UNIVERSIDADE FEDERAL DO CEARÁ

***CAMPUS* SOBRAL**

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

MESTRADO ACADÊMICO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

RHYAN XIMENES DE BRITO

SISTEMAS DE CLASSIFICAÇÃO E AUXÍLIO AO DIAGNÓSTICO DE TRANSTORNOS MENTAIS EM USUÁRIOS DE SUBSTÂNCIAS PSICOATIVAS COM BASE EM INTELIGÊNCIA COMPUTACIONAL

SOBRAL, CEARÁ

2021

RHYAN XIMENES DE BRITO

SISTEMAS DE CLASSIFICAÇÃO E AUXÍLIO AO DIAGNÓSTICO DE TRANSTORNOS MENTAIS EM USUÁRIOS DE SUBSTÂNCIAS PSICOATIVAS COM BASE EM INTELIGÊNCIA COMPUTACIONAL

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia Elétrica e de Computação do Programa de Pós-Graduação em Engenharia Elétrica e de Computação do *Campus* Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia Elétrica e de Computação. Área de Concentração: Sistemas de Informação.

Orientador: Prof. Dr. Carlos Alexandre Rolim Fernandes

Coorientadora: Profª. Dra. Eliany Nazaré Oliveira

SOBRAL, CEARÁ

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

B877s Brito, Rhyan Ximenes de.
Sistemas de Classificação e Auxílio ao Diagnóstico de Transtornos Mentais em Usuários de Substâncias Psicoativas com Base em Inteligência Computacional / Rhyan Ximenes de Brito. – 2021.
73 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Campus de Sobral, Programa de Pós-Graduação em Engenharia Elétrica e de Computação, Sobral, 2021.
Orientação: Prof. Dr. Carlos Alexandre Rolim Fernandes.
Coorientação: Profa. Dra. Eliany Nazaré Oliveira.

1. Transtornos Relacionados ao Uso de Substâncias. 2. Aprendizado de Máquina. 3. Mineração de Dados. I. Título.

CDD 621.3

RHYAN XIMENES DE BRITO

SISTEMAS DE CLASSIFICAÇÃO E AUXÍLIO AO DIAGNÓSTICO DE TRANSTORNOS MENTAIS EM USUÁRIOS DE SUBSTÂNCIAS PSICOATIVAS COM BASE EM INTELIGÊNCIA COMPUTACIONAL

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia Elétrica e de Computação do Programa de Pós-Graduação em Engenharia Elétrica e de Computação do *Campus* Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia Elétrica e de Computação. Área de Concentração: Sistemas de Informação.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Carlos Alexandre Rolim Fernandes (Orientador)
Universidade Federal do Ceará (UFC)

Profª. Dra. Eliany Nazaré Oliveira (Coorientadora)
Universidade Estadual do Ceará (UVA)

Prof. Dr. Rodrigo de Melo Souza Veras
Universidade Federal do Piauí (UFPI)

Profª. Dra. Francisca Raquel Silveira de Vasconcelos
Instituto Federal do Ceará (IFCE)

A Deus, eterna fonte de luz e amor. A meus filhos Ana Clara e Rafael, meu amor por você, me fez mais forte. Aos meus pais e a minha esposa, por acreditarem e me apoiarem. A meus avôs (Antonio e Manuel), minha tia Maria do Socorro e ao meu filho(a) todos *in memoriam*.

AGRADECIMENTOS

A Deus que permitiu que tudo isso acontecesse, estando presente em todos os momentos da minha vida.

Agradeço aos meus pais, Tarcísio e Liduina, que com muito sacrifício e renúncia, me trouxeram até aqui. Me proporcionaram educação e o conforto necessário para que pudesse seguir na minha caminhada acadêmica. Sem vocês essa vitória não teria sido possível!

À minha filha, Ana Clara, que mesmo tão pequena, por vezes abriu mão de momentos comigo, por tentar entender e aceitar minha ausência em alguns momentos importantes de sua vida, sempre com muito amor e um abraço bem apertado.

À minha esposa, Janaide, que sempre me apoiou, compreendendo minha ausência como esposo e como pai, me motivando e me encorajando a enfrentar os desafios e a superar os obstáculos que surgiram durante o percurso até aqui traçado. Agradeço pela companhia e amor doados em todo esse tempo. Agradeço por estar ao meu lado nos melhores e piores momentos de minha vida.

À Universidade Federal do Ceará e aos professores do Programa de Pós-Graduação em Engenharia Elétrica e de Computação, em especial aos profs. Drs. Iális Cavalcante, Jarbas Joaci, Rui Vigelis e Márcio Amora, por compartilharem seus conhecimentos durante a minha caminhada acadêmica.

A todos os colegas de mestrado, em especial a Joniel e Pablo, que me ajudaram em momentos de dificuldades. Foi um prazer enorme dividir momentos de companheirismo, alegrias e sabedorias compartilhadas com vocês.

Ao meu colega de trabalho e doutorando Adonias Caetano por compartilhar momentos de aprendizagem, conhecimento e sabedoria durante nossas conversas.

Agradeço especialmente ao meu orientador, prof. Dr. Carlos Alexandre Rolim Fernandes, pelo incentivo, confiança, acessibilidade, compreensão, apoio, conselhos e ajuda que foram essenciais para que, apesar de todos os obstáculos e dificuldades vivenciados, eu conseguisse produzir esse trabalho.

A minha coorientadora, profª Dra. Eliany Nazaré e a profª Ma. Roberta Magda por cederem gentilmente a base de dados e auxiliado na compreensão da mesma.

"O sonho é que leva a gente para frente. Se a gente for seguir a razão, fica aquietado, acomodado."

(Ariano Suassuna)

RESUMO

Os transtornos mentais estão entre as doenças mais prevalentes no mundo e muitos estudos observaram a relação entre o uso de substâncias psicoativas com o transtorno mental comum (TMC) ou mesmo com a depressão, caracterizados por sintomas depressivos, ansiosos e somáticos, como irritabilidade, fadiga, insônia e outros. Por outro lado, a aprendizagem de máquina (AM) tem sido amplamente utilizada para resolver muitos problemas em diferentes áreas. Nesse contexto, o presente trabalho tem como objetivo testar a eficácia da AM como ferramenta auxiliar no pré-diagnóstico do TMC e da depressão, por meio da classificação dos usuários de substâncias psicoativas quanto ao risco de depressão e ou mesmo do TMC. O objetivo principal é obter um modelo de previsão do risco de depressão e do TMC, bem como determinar quais os fatores que mais contribuem para a previsão do risco de depressão e do TMC. As bases de dados utilizadas neste trabalho foram compostas por 605 amostras de pessoas de oito cidades do estado do Ceará, no Brasil, coletadas de janeiro a julho de 2019. Os resultados mostraram a acurácia das técnicas de AM testadas na previsão do TMC e da depressão, atingindo uma acurácia de 82,81% e 81,98% respectivamente, com ênfase para o classificador *Support Vector Machine* (SVM) com a técnica de seleção de atributos *Sequential Backward Selection* (SBS) em ambas as bases de dados. Os resultados também mostraram que o uso de derivados do tabaco, álcool e cocaína/crack foram os fatores mais significativos na classificação nas bases de dados, apontando que o uso dessas substâncias psicoativas (SPA) causaram a recaída, contribuindo para o retorno do indivíduo ao uso de SPA, assim como quais SPA foram as mais utilizadas. Dessa forma, o estudo evidenciou que o uso de mineração de dados (MD) e técnica de AM podem contribuir de forma significativa no pré-diagnóstico de doenças como os transtornos mentais.

Palavras-chave: Transtornos Relacionados ao Uso de Substâncias. Aprendizado de Máquina. Mineração de Dados.

ABSTRACT

Mental disorders are among the most prevalent diseases in the world and many studies have observed the relationship between the use of psychoactive substances with CMD or even depression, characterized by depressive, anxious and somatic symptoms, such as irritability, fatigue, insomnia and others. On the other hand, ML has been widely used to solve many problems in different areas. In this context, this study aims to test the effectiveness of ML as an auxiliary tool in the pre-diagnosis of CMD and depression, through the classification of users of psychoactive substances regarding the risk of depression and/or even CMD. The main objective is to obtain a model to predict the risk of depression and CMD, as well as to determine which factors contribute most to the prediction of the risk of depression and CMD. The databases used in this work were composed of 605 samples from people from eight cities in the state of Ceará, Brazil, collected from January to July 2019. The results showed the accuracy of the ML techniques tested in the prediction of CMD and depression , reaching an accuracy of 82.81% and 81.98% respectively, with emphasis on the Support Vector Machine (SVM) classifier with the Sequential Backward Selection (SBS) attribute selection technique in both databases. The results also showed that the use of tobacco derivatives, alcohol and cocaine/crack were the most significant factors in the classification in the databases, pointing out that the use of these psychoactive substances (SPA) caused the relapse, contributing to the individual's return to the use of SPA, as well as which SPA were the most used. Thus, the study showed that the use of data mining (DM) and ML technique can significantly contribute to the pre-diagnosis of diseases such as mental disorders.

Keywords: Disorders Related to Substance Use. Machine Learning. Data Mining.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AG	algoritmo genético
AM	aprendizagem de máquina
AP	aprendizagem profunda
BDI	*Beck Depression Inventory*
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CRES	Coordenadoria Regional de Saúde
DPM	distúrbios psíquicos menores
DSM-IV	Manual Diagnóstico e Estatístico dos Transtornos Mentais-IV
EEG	eletroencefalograma
ELM	*Extreme Learning Machine*
FLD	*Fisher ´s Linear Discriminante*
FP	*false positive*
FUNCAP	Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico
GANs	*Generative Adversarial Network*
K-NN	*K-Nearest Neighbors*
KDD	*Knowledge Discovery in Databases*
LDA	*Linear Discrimant Analysis*
LR	*Logistic Regression*
MD	mineração de dados
MLP	*Multilayer Perceptron*
NB	*Naive Bayes*
OMS	Organização Mundial de Saúde
PCA	*Principal Component Analysis*
PHQ-9	*Patient Health Questionnaire-9*
PRIME-MD	*Primary Care Evaluation of Mental Disorders*
QDA	*Quadratic Discriminat Analysis*
RBF	*Radial Basis Function*
RBs	redes bayesianas
RF	*Random Forest*
RNA	rede neural artificial
RNAs	redes neurais artificiais

SBS	*Sequential Backward Selection*
SFS	*Sequential Forward Selection*
SPA	substâncias psicoativas
SRQ-20	*Self-Reporting Questionaire-20*
SVM	*Support Vector Machine*
TEA	transtorno do espectro do autista
TMC	transtorno mental comum
UFC	Universidade Federal do Ceará
UVA	Universidade Estadual Vale do Acaraú

SUMÁRIO

1 INTRODUÇÃO

Nos últimos 30 anos, muitos levantamentos epidemiológicos em todo o mundo mostraram que os transtornos mentais tornaram-se muito relevantes do ponto de vista da saúde pública devido à sua prevalência e persistência, respondendo por aproximadamente 12% dos diagnósticos globais de doenças (SKAPINAKIS *et al.*, 2013).

Sobre esse ponto de vista, o transtorno mental comum (TMC), responsável pela redução da capacidade de concentração e distúrbios de memória, é considerado o sofrimento mental mais predominante na população mundial e estará entre as maiores causas incapacitantes no ano de 2030, assim como a depressão (HARTMANN *et al.*, 2017; FENERICH *et al.*, 2020).

Da mesma forma, acredita-se que a depressão pode ser desencadeada por fatores biológicos, sendo a genética um fator significativo no desenvolvimento de um quadro depressivo (NÓBREGA *et al.*, 2015). Além disso, fatores psicológicos causam perda da autonomia e agravamento de quadros patológicos preexistentes, como os fatores sociais que interferem na capacidade funcional, no autocuidado e nas relações sociais (NÓBREGA *et al.*, 2015).

Nessa perspectiva, ressalta-se que o uso de substâncias psicoativas (SPA), por atuarem no sistema nervoso central provocando efeitos nas funções cognitivas, comportamentais e psicológicas, causando alterações de humor, comportamento e consciência, podem ter relação com o surgimento do TMC ou mesmo da depressão (MOREIRA *et al.*, 2020).

Diversos trabalhos observaram que vários problemas de saúde podem ter relação com o uso de SPA, como o TMC, que se caracteriza por sintomas, ansiosos e somáticos, como a irritabilidade, fadiga, insônia, preocupação excessiva, entre outros, ou mesmo com a depressão, que é caracterizada não apenas na forma de tristeza como também através da perda de interesse ou prazer em atividades cotidianas, perda de concentração ou memória, diminuição da autoestima, alteração de sono ou apetite, entre outros sintomas (MOREIRA *et al.*, 2020; LIMA *et al.*, 2019; LUCCHESE *et al.*, 2017; SAIDE, 2014; BARBOSA *et al.*, 2020). Para Hartmann *et al.* (2017) e Fenerich *et al.* (2020), o TMC e a depressão afligem milhões de pessoas em todo o mundo trazendo sofrimento para o paciente e familiares.

Por outro lado, a mineração de dados (MD) e a AM estão se tornando cada vez mais populares em muitas áreas do conhecimento como mecanismos auxiliares para a resolução de vários problemas. Atualmente a MD e a AM têm aplicações em uma grande variedade de campos do conhecimento e, em particular, elas se tornaram ferramentas poderosas nos campos da medicina, saúde e biologia (Ravì *et al.*, 2017; PAZMIÑO-MAJI *et al.*, 2017). Os sistemas

baseados em MD e/ou AM desenvolvidos nessas áreas visam identificar padrões em grandes quantidades de dados e auxiliar nas decisões clínicas, sendo uma poderosa ferramenta de auxílio no pré-diagnóstico e nos sistemas preditivos.

Mitchell (1997) define AM como a área de pesquisa que visa desenvolver programas computacionais capazes de automaticamente melhorar o desempenho de sistemas computacionais por meio da experiência. Assim, a aprendizagem de máquina tem como foco extrair informação a partir de dados de maneira automática (CASTRO; FERRARI, 2016). Sistemas desenvolvidos com base em AM buscam aprender a identificar padrões em grandes quantidades de dados, com o objetivo por exemplo de auxiliar na tomada de decisões clínicas em situações repetitivas (FENERICH *et al.*, 2020).

Para Amaral (2016), a MD é utilizada para explorar e analisar grandes volumes de dados em busca de padrões, previsões, erros, associações entre outros. Utilizando-se tanto de técnicas de AM como de redes neurais para extração de conhecimento, é parte integrante de um processo amplo conhecido como *Knowledge Discovery in Databases* (KDD) (CASTRO; FERRARI, 2016).

O presente estudo tem como objetivo testar a eficácia das técnicas de MD e AM como ferramentas auxiliares no pré-diagnóstico do TMC e da depressão, por meio da classificação de usuários de substâncias psicoativas de acordo com o risco. Em particular, o objetivo principal é obter um modelo de previsão do risco, com base em dados relativos ao uso de substâncias psicoativas e dados socioeconômicos. Outro objetivo é determinar quais fatores contribuem mais para a previsão do risco do TMC e da depressão. Esses dois objetivos combinados podem auxiliar os profissionais da área de saúde no rastreio de sinais e sintomas, assim como no desenvolvimento de políticas públicas de saúde que proporcionem qualidade de vida aos pacientes.

O modelo de previsão apresentado é baseado em sistema de classificação que segue as etapas do método KDD, ou seja na descoberta de conhecimento em bancos de dados para MD (PAZMIÑO-MAJI *et al.*, 2017). O KDD é um procedimento comumente usado para encontrar padrões explicáveis nos dados, que permitem a interpretação ou mesmo a previsão de eventos futuros, sendo a fase de mineração de dados apenas uma etapa desse processo (MOLINA-CORONADO *et al.*, 2020).

A abordagem KDD permite um melhor aproveitamento da base de dados, levando a uma utilização eficiente das técnicas de AM. Vários modelos de AM foram testados no sistema de classificação, a fim de encontrar aquele que tem a melhor capacidade de modelar a base

de dados considerada. Em particular, os nove seguintes classificadores foram testados por apresentarem resultados significativos na acurácia conforme a literatura: *K-Nearest Neighbors* (K-NN), *Multilayer Perceptron* (MLP), *Support Vector Machine* (SVM), *Linear Discrimant Analysis* (LDA), *Quadratic Discriminat Analysis* (QDA), *Naive LDA*, *Naive QDA*, *Extreme Learning Machine* (ELM) e *Random Forest* (RF).

Além disso, a fim de aumentar a precisão dos classificadores e reduzir o número de atributos, três técnicas de seleção/transformação de atributos foram testadas: *Principal Component Analysis* (PCA), *Sequential Forward Selection* (SFS) e *Sequential Backward Selection* (SBS). Objetivando determinar quais fatores contribuem mais para a predição do TMC e da depressão, foi realizada uma análise baseada na entropia (ganho de informação) e no índice de diversidade de Gini dos atributos.

As bases de dados utilizadas neste trabalho fazem parte de um estudo maior chamado "Saúde mental e o risco de suicídio em usuários de drogas", construídas pelas pesquisadoras da Universidade Estadual Vale do Acaraú (UVA) e da Universidade Federal do Ceará (UFC), professora Dra. Eliany Nazaré Oliveira e Roberta Magda Martins Moreira com dados coletados de 605 participantes, no período de janeiro a julho de 2019, em oito municípios do estado do Ceará/Brasil que possuem serviços de saúde mental e/ou comunidades terapêuticas que atendem usuários de substâncias psicoativas (MOREIRA, 2020). Os dados foram coletados em entrevistas apoiadas em três instrumentos: um formulário para perfil sociodemográfico, clínico e padrão de consumo, o *Self-Reporting Questionaire-20* (SRQ-20) e o *Patient Health Questionnaire-9* (PHQ-9).

A motivação para este trabalho está relacionada à importância da aplicação de técnicas de AM em situações que possam auxiliar profissionais da área de saúde no suporte ao diagnóstico de doenças como o TMC ou mesmo à depressão que aflige milhões de pessoas no mundo. Este trabalho mostra-se importante por trazer estudos que podem contribuir para a área de Psiquiatria, Psicologia, Enfermagem e Ciência da Computação, diferenciando-se principalmente com relação ao problema estudado para a tarefa de classificação, assim como na criação de diferentes modelos com abordagens distintas para verificar aquele(s) que tenham a maior capacidade de generalização para o problema aqui tratado.

1.1 Trabalhos Relacionados

Nesta seção, é apresentada uma breve discussão sobre algumas contribuições de outros trabalhos relacionados à aplicação de diversas técnicas de AM como mecanismo de apoio em problemas relacionados ao TMC, depressão e transtornos mentais semelhantes.

O trabalho Li e Fan (2006) apresenta o desenvolvimento de uma aplicação clínica mostrando que é possível diferenciar pacientes que sofrem de esquizofrenia, depressão e pessoas saudáveis, com base no ritmo do eletroencefalograma (EEG). Utilizaram duas redes neurais artificiais (RNAs), MLP e a *Self-organizing Competitive Network* para a classificação de três grupos de pacientes (10 normais, 10 esquizofrênicos e 10 depressivos), utilizando os ritmos EEG como vetores de atributos. Os resultados mostraram que as RNAs são uma abordagem eficaz para a classificação, com a MLP tendo o melhor desempenho frente a *Self-organizing Competitive Network*.

Em Fenerich *et al.* (2020), realizou-se um estudo para classificar os diversos tipos de cefaleias em pacientes, utilizando diferentes métodos de análise de dados, como redes bayesianas (RBs) e RNAs. Os dados foram coletados por meio de um processo de levantamento de dados aplicado a 2.177 pacientes com diagnóstico de cefaleia na Clínica Neurológica do município de Joinville-SC, Brasil, no período de janeiro de 2010 a novembro de 2014. Os resultados apresentaram uma boa acurácia em todos os testes realizados, mostrando que as RBs apresentaram melhor acurácia quando comparadas as RNAs.

Já em Hosseinifard *et al.* (2011), estudou-se o desempenho de diferentes técnicas de classificação como a *Logistic Regression* (LR) e SVM com intuito de discernir pacientes com depressão de indivíduos normais. Para selecionar as características mais importantes utilizaram algoritmo genético (AG). Para este propósito, foram registrados dados com EEG de 19 canais, com base em 30 pacientes depressivos e 30 normais. Assim, constataram que a SVM utilizando 15 atributos selecionados pelo AG atingiram uma taxa de precisão em torno de 88,60%.

Em Sau e Bhakta (2017), destaca-se o desenvolvimento de um modelo preditivo para diagnosticar ansiedade e depressão em pacientes idosos a partir de fatores sociodemográficos e relacionados à saúde, utilizando AM. Dez classificadores foram avaliados com um conjunto de dados de 510 pacientes geriátricos e testados com o método de validação cruzada *k-fold*. A maior precisão foi de 89% obtida com o classificador de RF. O modelo de RF foi testado com outro conjunto de dados de 110 pacientes idosos separados para sua validade externa. Sua precisão preditiva foi de 91% e a taxa de *false positive* (FP) foi de 10%.

No trabalho Khan e Wang (2017), desenvolveu-se uma ferramenta computacional com AM a partir de dados de sequenciamento do genoma completo em transtornos mentais. Os autores utilizaram um sistema de pontuação baseado em aprendizado profundo (*ncDeepBrain*) para analisar dados de sequenciamento de genomas pessoais por integração de contribuições de codificação, não-codificação, variantes estruturais, locus de traços quantitativos de expressão cerebral conhecidos (eQTLs) e picos de intensificador promotor de *PsychENCODE*. Observaram que, em estudos populacionais, o método pode ajudar a priorizar novas variantes que estão associadas a suscetibilidade a doença.

Em McGinnis *et al.* (2018), abordou-se o diagnóstico da ansiedade e depressão em crianças pequenas, através do uso da indução do medo de 90 segundos, durante a qual o movimento do participante foi monitorado usando um sensor vestível. O AM e os dados extraídos da fase de 20 segundos mais clinicamente viável da tarefa foram usados para prever o diagnóstico em uma amostra de crianças com e sem um diagnóstico de ansiedade e depressão. Os autores observaram que a regressão logística obteve o melhor desempenho com precisão de 80% no diagnóstico.

No artigo Moreira *et al.* (2020), investigou-se a presença do transtorno mental comum e sua associação com fatores relacionados ao perfil sociodemográfico de usuários de substâncias psicoativas. O estudo foi realizado com 497 usuários de substâncias psicoativas de oito municípios do interior do estado do Ceará. Os dados foram coletados com formulário para perfil sociodemográfico e o SRQ-20, analisados por estatística inferencial, com testes de associação, comparação e correlação. Por fim, constataram que 78,10% apresentaram rastreamento positivo para transtorno mental comum com maior índice para humor ansioso, depressivo e sintomas somáticos, como fatores de risco, o sexo feminino e ter menor idade, ao passo que, ter religião católica ou evangélica, ocupação, um companheiro fixo e filhos consistiram em fatores protetores.

Em Sapkal *et al.* (2021), um sistema neuro-fuzzy foi usado para reconhecer transtornos mentais como esquizofrenia, fobia, depressão, ansiedade e transtorno obsessivo-compulsivo, usando mineração de dados. Para a coleta de dados, foram utilizados questionários sobre sintomas e tipos de transtornos.

1.2 Produção Científica

Esta subseção aborda a produção científica com artigos publicados e submetidos em periódicos ou com participação em eventos:

- DE BRITO, Rhyan Ximenes; FERNANDES, Carlos Alexandre Rolim; MOREIRA, Roberta Magda Martins; OLIVEIRA, Eliany Nazaré. *Prediction Model for Common Mental Disorder in Users of Psychoactive Drugs Using Data Mining*, artigo a ser submetido à revista científica IEEE - *Latin America Transactions.*

O artigo acima mencionado foi elaborado utilizando os resultados obtidos a partir da base de dados sobre o TMC integrantes desta dissertação. Destaca-se ainda que os artigos mencionados logo abaixo foram desenvolvidos durante algumas disciplinas no decorrer do mestrado.

- DE BRITO, Rhyan Ximenes; FERNANDES, Carlos Alexandre Rolim; AMORA, Márcio André Baima. Análise de Desempenho com Redes Neurais Artificiais, Arquiteturas MLP e RBF para um Problema de Classificação de Crianças com Autismo. iSys-Revista Brasileira de Sistemas de Informação, v. 13, n. 1, p. 60-76, 2019.

Em Brito *et al.* (2019), realizou-se um estudo com a implementação e análise das redes neurais, MLP e *Radial Basis Function* (RBF), objetivando comparar os resultados baseados no treinamento, teste e classificação do diagnóstico de crianças com ou sem o transtorno do espectro do autista (TEA). A implementação foi realizada com base em 292 amostras de um banco de dados público, com validação cruzada *k-fold*, com k=10 *folds*. Nos resultados foi constatado que a rede neural MLP obteve a melhor média de acertos com 86,26% enquanto que a RBF foi de 83,12%.

- DE BRITO, Rhyan Ximenes; FERNANDES, Carlos Alexandre Rolim; XIMENES, Janaide Nogueira de Sousa. Avaliação de RNAs Durante Treinamento Supervisionado Para Classificação de Adolescentes com Autismo. In: Anais da VIII Escola Regional de Computação do Ceará, Maranhão e Piauí. SBC, 2020. p. 53-60.

Já em Brito *et al.* (2020) implementou-se as redes neurais ELM e MLP, comparando as acurácias resultantes de treinamentos com dados de adolescentes com ou sem o TEA. A metodologia foi baseada em um banco de dados público e na técnica de validação cruzada *k-fold* com e sem normalização *zscore*. Com relação aos resultados a rede MLP sem *zscore* obteve a melhor média atingindo 89,70% de acertos, por outro lados a ELM sem *zscore* obteve a pior média de acertos com 86,52%.

1.3 Divisão da Dissertação

Os capítulos desta dissertação estão organizados como apresentado abaixo:

Capítulo 2: apresenta os objetivos geral e específicos relacionados ao trabalho;

Capítulo 3: apresenta a fundamentação teórica com uma resenha técnica sobre os assuntos abordados durante esse trabalho;

Capítulo 4: descreve material e métodos propostos, apresentando uma descrição das bases de dados utilizadas, assim como os passos adotados na condução da pesquisa, com enfoque nas etapas do modelo de predição;

Capítulo 5: são discutidos os resultados encontrados com base nos treinamentos e testes dos classificadores e demais técnicas de AM utilizadas;

Capítulo 6: apresenta as conclusões obtidas no decorrer deste estudo, destacando os principais resultados obtidos, asssim como perspectivas para trabalhos futuros.

2 OBJETIVOS

2.1 Objetivo Geral

- Obter um modelo de previsão do risco de depressão e do TMC, bem como determinar quais os fatores que mais contribuem para a previsão do risco de depressão e do TMC.

2.2 Objetivos Específicos

- Desenvolver modelos de predição para o TMC e para a depressão com base no uso de substâncias psicoativas e em dados socioeconômicos. Os modelos de predição são baseados no método KDD e em técnicas de AM, através da classificação dos participantes do banco de dados considerado segundo o risco de TMC ou de depressão;
- Testar diversos classificadores e diversas técnicas de seleção de atributos para se determinar qual o modelo de AM que melhor se adequa à modelagem da base de dados considerada;
- Determinar quais atributos são mais relevantes para a predição do risco de TMC e depressão, usando entropia (ganho de informação) e índice de diversidade de Gini;
- Interpretar os resultados obtidos a partir dos treinamentos e testes com as bases de dados (TMC e depressão).

3 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta uma revisão técnica e teórica sobre os assuntos utilizados como base para a elaboração desse trabalho, estando dividido da seguinte forma. A Subseção 3.1 traz a definição da depressão e suas características. A Subseção 3.2 aborda o TMC. A Subseção 3.3 traz a definição de KDD. A Subseção 3.4 apresenta definições sobre AM, tais como classificadores, técnicas de seleção de atributos, entropia e coeficiente de Gini.

3.1 Depressão

De acordo com Association *et al.* (2014), a característica comum desse transtorno é a presença de humor triste ou melancólico, vazio ou irritável, acompanhado de alterações somáticas e cognitivas que afetam significativamente a capacidade de funcionamento do indivíduo, tendo como diferimento entre eles os aspectos de duração, momento ou etiologia presumida.

Para Beck e Alford (2016), a depressão ou melancolia é reconhecida como uma síndrome clínica há mais de 2 mil anos e ainda não foi encontrada uma explicação plenamente satisfatória para suas características intrigantes e paradoxais, existindo importantes questões não resolvidas sobre sua natureza, classificação e etiologia. Entre essas questões pode-se observar os seguintes pontos: (i) a depressão é o exagero de um estado de humor vivenciado por indivíduos normais, ou é qualitativa e quantitativamente diferente de um estado de humor normal?; (ii) quais são as causas, as características definidoras, os resultados e os tratamentos efetivos da depressão?; (iii) a depressão é um tipo de reação ou uma doença?; (iv) a depressão é causada principalmente por estresse psicológico e conflito, ou está basicamente relacionada a uma desordem biológica?

Não existem respostas para essas perguntas, existindo entretanto, uma nítida discordância entre clínicos e investigadores que escreveram sobre a depressão, com considerável controvérsia quanto à classificação da depressão. A natureza e a etiologia da depressão estão sujeitas a opiniões ainda divididas. Algumas autoridades afirmam que a depressão é sobretudo um transtorno psicogênico, outras afirmam que a causa está relacionada a fatores orgânicos. Um terceiro grupo defende o conceito de dois tipos diferentes de depressão: um psicogênico e outro orgânico (BECK; ALFORD, 2016).

Dentro dessa perspectiva, Association *et al.* (2014) afirma que o abuso de um grande número de substâncias como as SPA, alguns medicamentos e diversas condições médicas podem estar associadas a depressão, esse fato é reconhecido nos diagnósticos de transtorno depressivo.

O conhecimento das propriedades dos testes, quanto à capacidade de identificar corretamente indivíduos em risco de apresentar depressão é imprescindível em estudos epidemiológicos, uma vez que permite corrigir as estimativas de prevalência da doença em função dos erros de classificação, decorrentes da acurácia imperfeita (SANTOS *et al.*, 2013).

De acordo com Santos *et al.* (2013), entre os instrumentos usados para identificar indivíduos em risco de depressão, encontra-se o PHQ-9, o mesmo é derivado do *Primary Care Evaluation of Mental Disorders* (PRIME-MD), que foi originalmente desenvolvido para identificar cinco transtornos mentais comuns em atenção primária à saúde: depressão, ansiedade, abuso de álcool, transtornos somatoformes e transtornos da alimentação. O PHQ-9 caracteriza-se por ser um instrumento de aplicação relativamente rápida, contendo nove questões, o que seria uma vantagem em estudos epidemiológicos, em comparação a outros atualmente validados para o Brasil, como o *Beck Depression Inventory* (BDI).

3.2 Transtorno Mental Comum (TMC)

O TMC ou distúrbios psíquicos menores (DPM) são quadros de intenso sofrimento psíquico com importantes repercussões para a saúde do indivíduo e prejuízos em vários aspectos da vida em termos de desempenho de papéis, envolvendo trabalho, estudos e demais atividades do cotidiano. A expressão TMC foi cunhada por Goldeber e Huxley, cujo conceito desenvolveu-se na década de 1970, por meio de pesquisas sobre o adoecimento mental no âmbito da atenção primária em saúde, sendo caracterizado por uma sintomatologia não psicótica através de queixas de ansiedade, irritabilidade, somatização, diminuição da energia vital e humor depressivo (FALCO *et al.*, 2019).

No que concerne aos transtornos como enfermidade que acomete a população, o TMC é caracterizado pela presença de diferentes sintomas por pelo menos sete dias. A avaliação desses sintomas possibilita o diagnóstico precoce e o acompanhamento de transtornos depressivo, ansiedade, fobia, transtorno de pânico e transtorno obsessivo-compulsivo, caracterizados como alguns dos tipos de TMC (LUCCHESE *et al.*, 2017).

O TMC é caracterizado por sintomas depressivos, estados de ansiedade, irritabilidade, fadiga, insônia, dificuldade de memória e concentração e queixas somáticas e, manifesta-se como uma mistura de sintomas somáticos, ansiosos e depressivos. O diagnóstico precoce e correto desse transtorno é fundamental para evitar prejuízos físicos e psicológicos ao indivíduo e ônus ao sistema de saúde (PARREIRA *et al.*, 2017).

Deve-se salientar que a prevalência do TMC oscila mundialmente e é muito frequente na população. No Brasil, na região centro-oeste, a probabilidade de TMC esteve presente em um terço dos entrevistados (31,47%), com maior prevalência nas Regiões Sudeste (51,9% a 53,3%), Nordeste (64,3%) e Sul (57,7%) (LUCCHESE *et al.*, 2017).

A avaliação da saúde mental para o TMC é executada por meio da aplicação do *Self-Reporting Questionnaire* (SRQ-20), desenvolvido por Harring e Mcmullin. O questionário originalmente possuía 24 perguntas: vinte sobre distúrbios não psicóticos e quatro referentes a distúrbios psicóticos. A versão aplicada no Brasil foi validada por Mari e Williams, que observaram uma sensibilidade de 83%, especificidade de 80% e 19% de erros de classificação (MINAYO *et al.*, 2008).

Dessa maneira o SRQ-20 é um instrumento para rastreamento de transtornos mentais não psicóticos em que as respostas são categóricas do tipo sim/não. Cada resposta afirmativa pontua com o valor 1 para compor o escore final por meio do somatório desses valores. Os escores obtidos estão relacionados com a probabilidade de presença de transtorno não psicótico, variando de 0 (nenhuma probabilidade) a 20 (extrema probabilidade) (GONÇALVES *et al.*, 2008). Dessa forma, as respostas possibilitam o estabelecimento de um escore, em que acima de 7, o indivíduo apresenta rastreamento positivo para TMC (LIMA *et al.*, 2006).

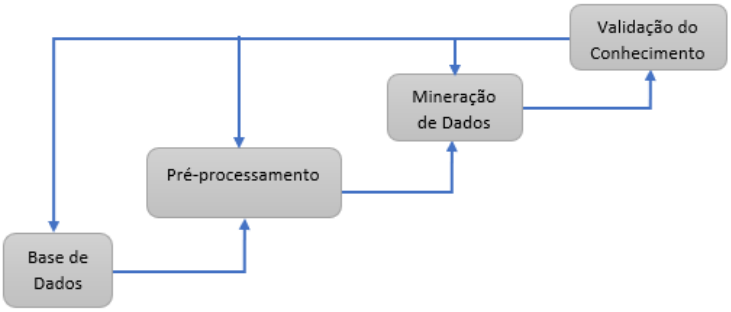
Além disso, o SRQ-20 é recomendado pela Organização Mundial de Saúde (OMS) para estudos comunitários e em atenção básica a saúde, principalmente nos países em desenvolvimento, por apresentar facilidade de uso e custo reduzido, sendo utilizado em vários países de culturas diferentes para rastreamento de transtornos não-psicóticos (GONÇALVES *et al.*, 2008).

3.3 *Knowledge Discovery in Databases* (KDD)

Para Castro e Ferrari (2016), a mineração de dados é parte integrante de um processo mais amplo, conhecido como descoberta de conhecimento em bases de dados. A KDD tem como objetivo encontrar padrões intrínsecos aos dados nela contidos, apresentando-os de forma a facilitar sua assimilação como conhecimento. Tal descoberta está associada a um processo analítico, sistemático e, até onde possível, automatizado (SILVA *et al.*, 2017).

A KDD geralmente segue os seguintes passos: a seleção e integração das bases de dados, a limpeza da base, a seleção e transformação dos dados, a mineração e a avaliação dos dados (CASTRO; FERRARI, 2016). Nessa perspectiva os autores mencionados sintetizam o processo da KDD em quatro partes principais, conforme observa-se na Figura 1.

Figura 1 – Processo de Descoberta de Conhecimento em Bases de Dados



Fonte: Castro e Ferrari (2016).

As etapas do KDD podem ser resumidas da seguinte forma:

- **Base de dados:** coleção organizada de dados, com valores quantitativos ou qualitativos que permitem uma recuperação eficiente dos dados.
- **Pré-processamento de dados:** são etapas anteriores à mineração que visam preparar os dados para uma análise eficiente e eficaz. Essa etapa inclui a limpeza (remoção de ruídos e dados inconsistentes), a integração (combinação de dados obtidos a partir de múltiplas fontes), a seleção ou redução (escolha dos dados relevantes à análise) e a transformação (transformação ou consolidação dos dados em formatos apropriados para a mineração;
- **Mineração de dados:** corresponde à aplicação de algoritmos capazes de extrair conhecimentos a partir dos dados pré-processados;
- **Validação do Conhecimento:** avaliação dos resultados da mineração objetivando identificar conhecimentos verdadeiramente úteis e não triviais.

Deve-se salientar que essas quatro etapas são correlacionadas e interdependentes de tal forma que a abordagem ideal para extrair informações relevantes consiste em considerar as inter-relações entre cada uma dessas etapas e sua influência no resultado final, permitindo que conhecimentos interessantes e úteis sejam extraídos da base de dados e validados sob diferentes perspectivas. Esses conhecimentos poderão ser usados para a tomada de decisões estratégicas, como por exemplo, controle de processos, gestão da informação e conhecimento, processamento de consultas entre outras aplicações (CASTRO; FERRARI, 2016).

3.4 *Aprendizagem de Máquina (AM)*

A AM é a área de pesquisa que tem como objetivo desenvolver programas computacionais com a capacidade automática de melhorar seu desempenho pela experiência. Assim, os algoritmos de AM são ferramentas poderosas para a descoberta de conhecimentos em bases de

dados (CASTRO; FERRARI, 2016).

Além do mais, a área de AM é fundamentada em conceitos de muitas outras áreas, como estatística, inteligência artificial, filosofia, teoria da informação, biologia, ciências cognitivas, complexidade computacional e teoria de controle (CASTRO; FERRARI, 2016).

3.4.1 Classificadores

Os classificadores são funções que utilizam como entrada padrões desconhecidos e como saída as classes que estes padrões provavelmente pertencem, para que seja possível a realização do reconhecimento em AM (CERQUEIRA, 2010).

3.4.1.1 Multilayer Perceptron (MLP)

Para Bassetto *et al.* (2020), as redes neurais artificiais MLP são modelos computacionais que apresentam uma estrutura formada por um conjunto de elementos chamados neurônios, similares aos existentes no cérebro humano, distribuídos paralelamente e compostos pelas camadas de entrada, camadas ocultas e de saída interligadas entre si por conexões. São do tipo *feedforward*, ou seja, cada camada se conecta à próxima camada, tal que cada neurônio fornece sua saída para cada unidade da camada seguinte (BONIFÁCIO, 2010).

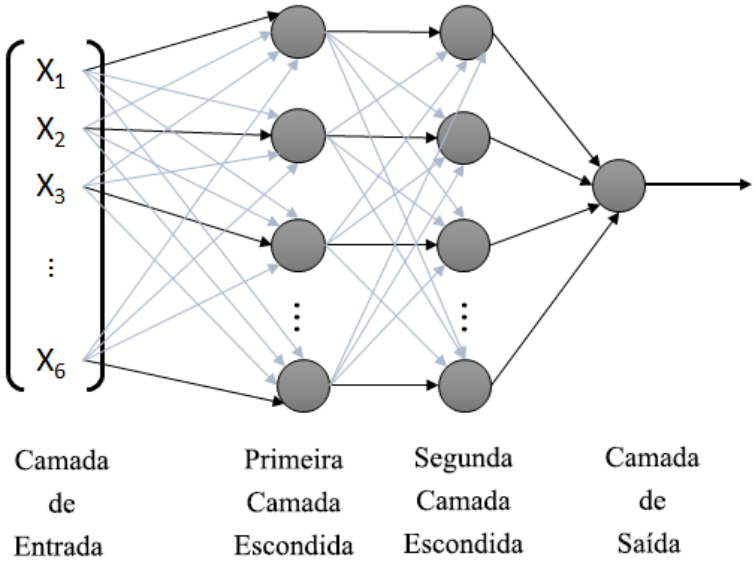
As funções de ativação dos neurônios devem ser não lineares e diferenciáveis, isto é, o gráfico da função não pode ser uma reta e deve ser possível calcular a derivada da função. A não linearidade serve para separar padrões que não são linearmente separáveis, a diferenciação permite o cálculo do gradiente da função, direcionando assim o ajuste dos pesos dos neurônios durante o treinamento (BONIFÁCIO, 2010).

Para Bocanegra (2002), as arquiteturas do tipo *perceptron* de múltiplas camadas constituem os modelos neurais artificiais mais utilizados e conhecidos atualmente. Os sinais de entrada são propagados pela rede em uma direção positiva, da entrada para a saída, representando uma generalização do *perceptron* simples.

Haykin (2007) afirma que as redes do tipo MLP têm sido utilizadas com eficiência para solucionarem vários problemas envolvendo altos graus de não linearidade, como por exemplo, reconhecimento, classificação de padrões, agrupamento, previsão, e nos últimos anos, na estimativa de variáveis astronômicas (BASSETTO *et al.*, 2020). Seu treinamento é do tipo supervisionado e utiliza um algoritmo muito popular chamado retro propagação do erro (*error backpropagation*), baseado em uma regra de aprendizagem que "corrige" o erro durante o

treinamento.

Figura 2 – Arquitetura de Rede Neural MLP



Fonte: Zaghetto *et al.* (2015).

Na Figura 2, é apresentada a estrutura básica de uma rede neural artificial (RNA) do tipo MLP normalmente utilizada em problemas de classificação e de aproximação (ou análise de regressão) o que inclui previsão e modelagem de séries temporais em áreas como: controle, diagnósticos e MD (FERREIRA *et al.*, 2016).

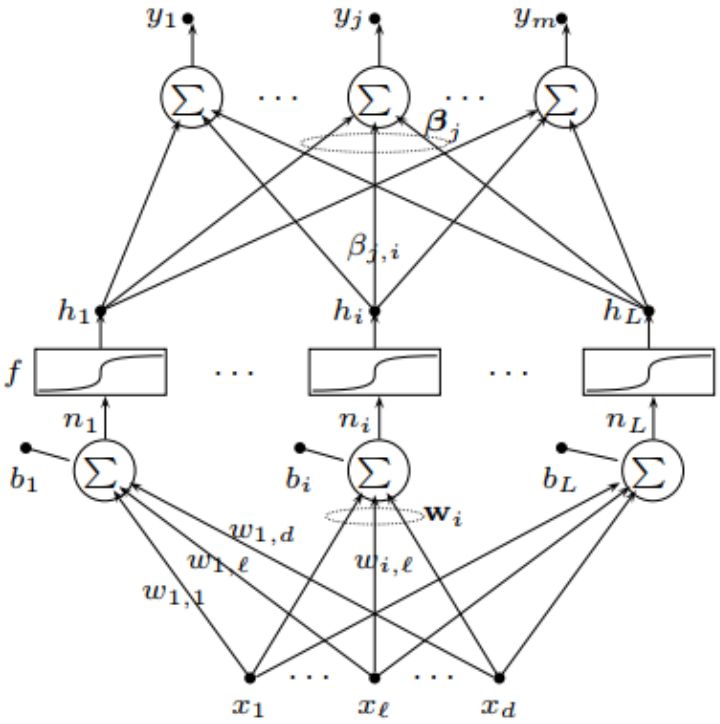
3.4.1.2 Extreme Learning Machine (ELM)

As ELM são redes neurais *feedforward* com estrutura muito semelhante às das redes neurais MLP (HAYKIN, 2007). A principal diferença entre essas estruturas está no processo de treinamento, em que as ELM não ajustam os pesos da camada intermediária, a qual possui neurônios gerados de forma aleatória e independente. O treinamento busca encontrar os melhores pesos da camada de saída, por meio de uma solução de um problema de otimização via mínimos quadrados, baseada no paradigma supervisionado, com a utilização de um sinal de referência. Esta característica faz com que os ajustes dos pesos da rede sejam rápidos e eficientes computacionalmente (HUANG *et al.*, 2006).

As ELM foram propostas inicialmente em Huang *et al.* (2006), em que os autores apresentaram, por meio de rigorosa demonstração matemática, que os pesos da camada intermediária podem ser escolhidos de forma arbitrária, com a condição de que a função de ativação dos pesos seja infinitamente diferenciável. Dessa forma, eles mostraram a capacidade

de aproximação universal da estrutura, onde as redes ELM podem aproximar, com erro arbitrário, qualquer mapeamento não linear. A Figura 3 mostra uma rede neural ELM com sua estrutura semelhante à da rede neural MLP.

Figura 3 – Arquitetura de Rede Neural ELM



Fonte: Boldt (2017).

3.4.1.3 Support Vector Machine (SVM)

A SVM constituem uma técnica de aprendizado que vem recebendo crescente atenção da comunidade de AM (MITCHELL, 1997). A SVM é uma técnica de aprendizado, onde os resultados da aplicação dessa técnica são comparáveis e muitas vezes superiores aos obtidos por outros algoritmos de aprendizado, como nas RNAs. Aplicações de sucesso podem ser encontradas em diversos domínios, como na categorização de textos, na análise de imagens, em bioinformática entre outras áreas (LORENA; CARVALHO, 2007).

A SVM é embasada na teoria de aprendizado estatístico desenvolvida por Vapnik. Esta teoria que estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores que possuem a capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado aconteceu (LORENA; CARVALHO, 2007).

Os algoritmos SVM têm como objetivo a determinação de limites de decisão que

produzam uma separação ótima entre classes por meio da minimização dos erros. A classificação é baseada no princípio de separação ótima entre classes, tal que se as classes são separáveis, a solução é escolhida de forma a maximizar a distância entre as classes (NASCIMENTO *et al.*, 2009).

As SVM implementam um mapeamento não-linear, executado por um produto interno *kernel*, em que um hiperplano ótimo é construído para separar os dados não lineares em duas classes. Assim, um *kernel* pode ser entendido como uma função que recebe dois vetores $\mathbf{x}_i$ e $\mathbf{x}_j$, retornando como saída um produto escalar a partir do mapeamento das entradas em uma dimensionalidade maior, possibilitando uma melhor distribuição e facilitando a classificação das amostras utilizadas (SOENTPIET *et al.*, 1999; SEMOLINI, 2002; CHANG; LIN, 2011).

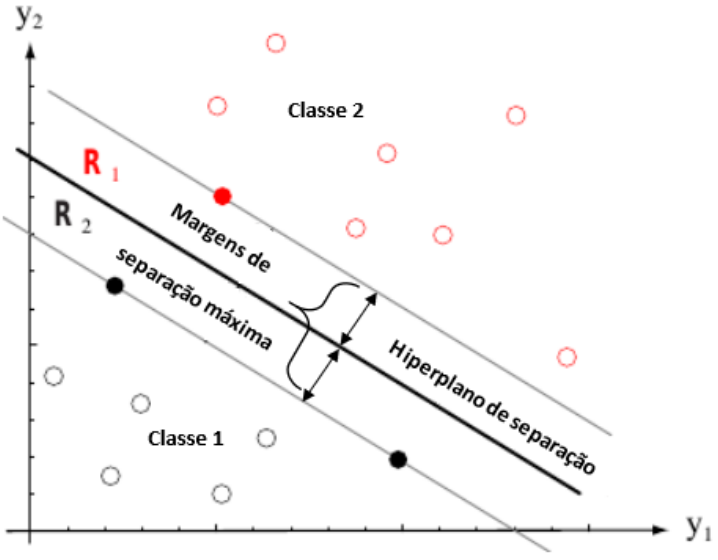
De acordo com Soentpiet *et al.* (1999), a SVM possui vários tipos de *kernels*, conforme pode ser visto na Tabela 1. A Figura 4 mostra um exemplo de SVM onde a mesma encontra o hiperplano com a distância máxima dos padrões de treinamento mais próximos (DUDA *et al.*, 2006).

Tabela 1 – Alguns *Kernels* Utilizados no SVM

Tipo *Kernel*	Função K(x_i, x_j)	Parâmetros
Gaussiano	$exp(-\sigma \mid\mid x_i - x_j \mid\mid^2)$	σ
Polinomial	$(\delta(x_i \cdot x_j) + \kappa)^d$	δ, κ e d
Sigmoidal	$tanh(\delta(x_i \cdot x_j)) + \kappa 1$	δ e κ

Fonte: Chang e Lin (2011).

Figura 4 – Exemplo de Classificação com SVM



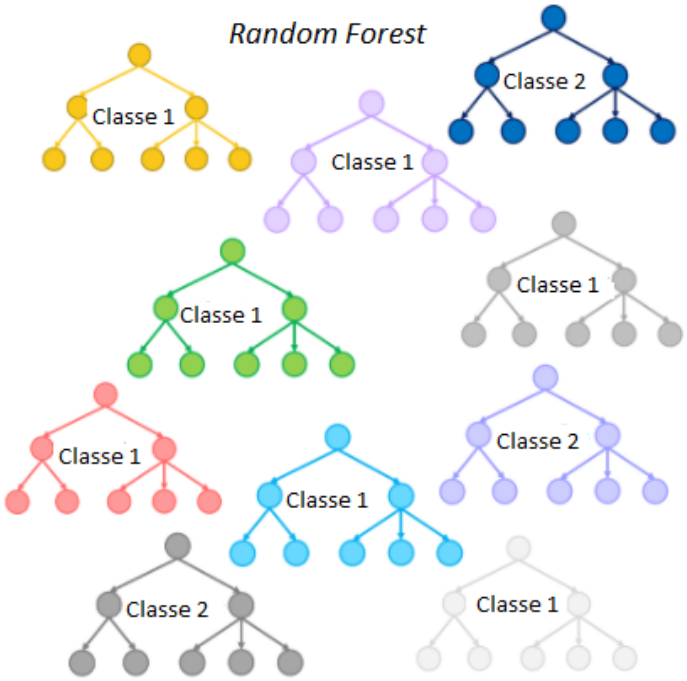
Fonte: Traduzido de Duda *et al.* (2006).

3.4.1.4 Random Forest (RF)

O RF é um método estatístico que não requer uma distribuição específica sobre a relação das covariáveis com a variável resposta. Assim, o RF é uma técnica não linear robusta que otimiza a acurácia de predição, realizando ajustes sobre o conjunto de árvores (BREIMAN, 2001). O RF é um algoritmo de classificação baseado na construção de árvores de decisão, em que árvore de decisão é uma técnica de MD utilizada para descobrir regras de classificação para um atributo a partir da subdivisão dos dados em um conjunto que está sendo analisado (APTÉ; WEISS, 1997).

Já Breiman (2001) salienta que o RF é um classificador composto por uma coleção de árvores de decisão com amostras aleatórias independentes e identicamente distribuídas, em que cada árvore vota na classe mais popular para uma entrada x. Cada árvore de decisão é gerada a partir de um novo conjunto de atributos selecionados aleatoriamente por uma técnica de amostragem com reposição chamada *Bootstrap*. A Figura 5 mostra uma combinação de árvores de decisão que são geradas para serem utilizadas na classificação de novas classes.

Figura 5 – Exemplo de Classificação com RF



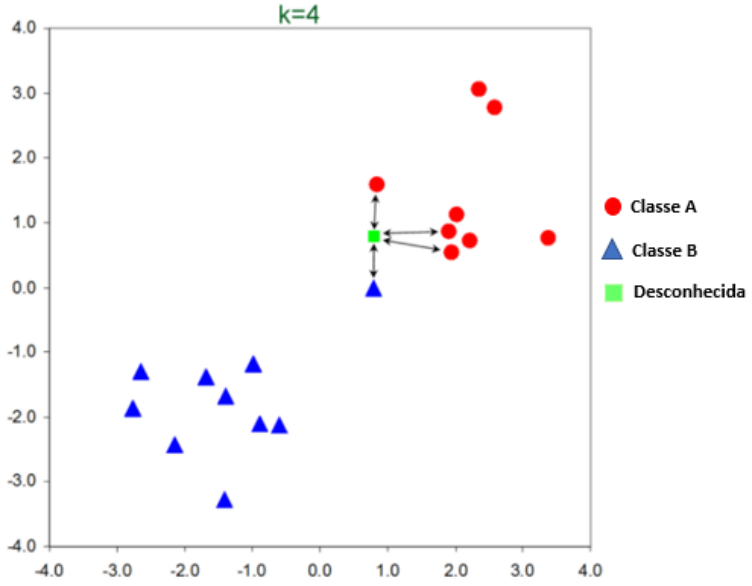
Fonte: Traduzido de Silipo e Melcher (2019).

3.4.1.5 K-Nearest Neighbors (K-NN)

Para Buani (2010), o algoritmo K-NN é utilizado de forma eficiente na resolução de problemas de classificação, dependentes da dimensionalidade dos dados. O K-NN é um classificador muito simples baseado na regra de atribuição de uma amostra desconhecida à classe das K amostras que estejam mais próximas, utilizando algumas métricas de distância como por exemplo, a euclidena, Mahalanobis, *cityblock* entre outras. Vale salientar que quando o valor de K>1, a classe dominante entre as K amostras será a classe do objeto desconhecido, contudo se houver empate, a decisão será feita pela amostra mais próxima do objeto desconhecido (BACKES; JUNIOR, 2019).

A seguir, têm-se algumas propriedades do algoritmo K-NN: (i) requer apenas um parâmetro, o número K de vizinhos mais próximos; (ii) não necessita de conhecimento prévio sobre a distribuição dos dados de treinamento; e (iii) tem provado convergir na abordagem ótima *Bayesiana* sob certas condições (BUANI, 2010).

Figura 6 – Exemplo de Uso do K-NN



Fonte: Traduzido de Peterson (2009).

A Figura 6 representa um problema com duas classes definidas como classe A representada pelo circulo vermelho, classe B representada pelo triângulo azul e a classe desconhecida representada pelo quadrado verde.

3.4.1.6 Linear Discriminant Analysis (LDA)

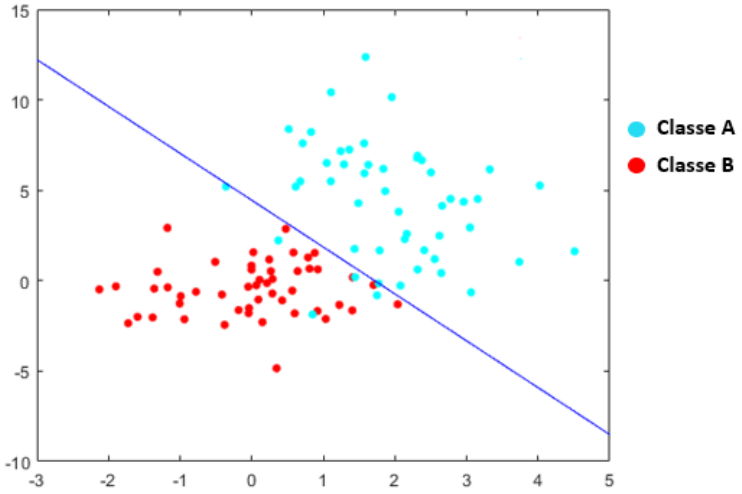
Para Belhumeur *et al.* (1997), Thomaz *et al.* (2006), Ye *et al.* (2006), o LDA é um método estatístico que busca maximizar a dispersão entre as classes enquanto minimiza a dispersão dos dados intra-classes. Por outro lado, para Santos (2005), o LDA é uma técnica clássica em reconhecimento de padrões desenvolvida primeiramente por Robert Fisher em 1936, também chamada de *Fisher´s Linear Discriminante* (FLD). Essa técnica é usualmente utilizada para classificação de dados e redução de dimensionalidade.

Na análise discriminante baseada em uma função discriminante linear, constrói-se uma função pela combinação das variáveis discriminantes. Nessa técnica, tenta-se construir a melhor função discriminante linear em termos de discriminação entre grupos (BARTH, 2004).

De acordo com Welling (2005), a LDA fornece um nível adequado de correção para classificação. Matematicamente, a LDA maximiza a razão de variâncias entre as classes e dentro da classe. Ela envolve menos poder computacional e resulta em separabilidade máxima.

Salienta-se que no LDA na versão *Naive Bayes* (NB) a matriz de covariâncias, assume que os atributos de entrada são descorrelacionados, ou seja, a matriz de covariância é considerada diagonal.

Figura 7 – Exemplo de Uso do LDA



Fonte: Próprio autor

A Figura 7 mostra um exemplo de um problema com duas classes, definidas como classe A, representada pelo circulo azul, e a classe B, representada pelo circulo vermelho, visualizadas em um espaço 2D, onde o LDA faz a separação no espaço abrangido.

Conforme observado na Figura 7 a LDA desenha uma região de decisão linear entre

uma série de classes dadas que representa um hiperplano no espaço de recursos para diferenciar as classes (ISHFAQUE *et al.*, 2013).

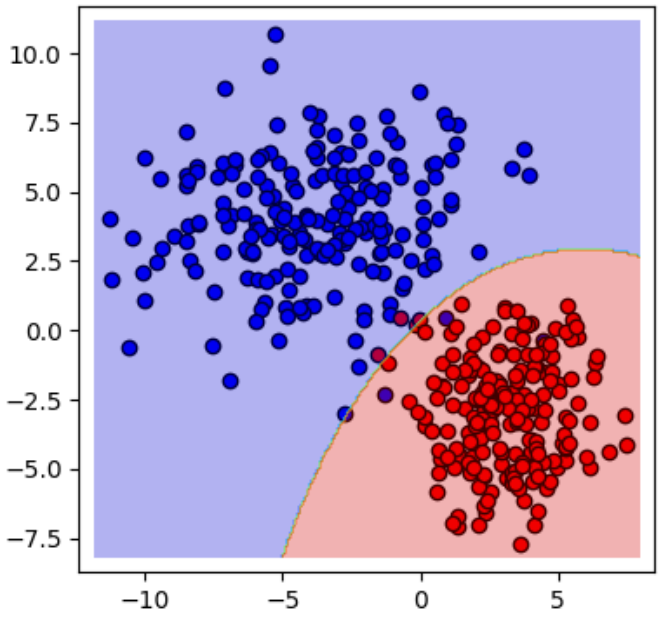
3.4.1.7 Quadratic Discriminant Analysis (QDA)

Para Muhammad *et al.* (2014), a QDA é uma abordagem padrão para problemas de classificação supervisionada. Na QDA, presume-se que as medições de cada classe são normalmente distribuídas. Os parâmetros para cada classe podem ser estimados a partir de pontos de treinamento com estimativa de máxima verossimilhança.

A QDA pressupõe que cada classe tem a sua própria matriz de covariância (BACKES; JUNIOR, 2019). Visto que a matriz de covariância é calculada por classe, o método QDA traça elípses, hipérboles etc. em espaços 2D; elipsoides, hiperboloides etc. em espaços 3D; e hiperelipsoides, hiper-hiperboloides etc. em espaços de maiores dimensões (BACKES; JUNIOR, 2019). Assim, Backes e Junior (2019) enfatizam que podem haver problemas no cálculo da matriz caso haja um número pequeno de amostras por classe. O ideal é que cada classe possua uma quantidade de amostras bem maior do que a quantidade de atributos dos vetores.

Destaca-se ainda que a QDA na versão NB é uma adaptação que assume que os atributos de entrada são descorrelacionados, ou seja, a matriz de covariância é considerada diagonal. A Figura 8 mostra o resultado de uma classificação com a utilização do QDA com base em duas classes sintéticas e uma região de separação não linear.

Figura 8 – Exemplo de Uso do QDA



Fonte: Ghojogh e Crowley (2019).

3.4.2 Técnicas de Seleção de Atributos

De acordo com Guyon e Elisseeff (2006), as técnicas de seleção de atributos são muito exploradas na área de MD, principalmente na tarefa de classicação, tendo como objetivo selecionar atributos relevantes de forma a obter os seguintes benefícios: (i) redução do tempo de execução do processo de classicação, com menos atributos avaliados, o processo de classicação tende a ser executado em menos tempo de processamento; (ii) aumento da capacidade preditiva do classicador, dessa maneira a seleção de atributos procura retirar atributos redundantes ou irrelevantes da base de dados, permitindo a geração de um classicador menos propenso a erros; e por fim (iii) obter uma representação mais compacta do conceito a ser aprendido visto que o conhecimento ficará concentrado somente nos atributos realmente importantes para a classicação (PAES *et al.*, 2013).

3.4.2.1 Principal Component Analysis (PCA)

PCA é uma técnica da estatística multivariada que consiste em transformar um conjunto de variáveis originais em outro conjunto de variáveis de mesma dimensão denominadas de componentes principais. Foi inicialmente descrita por Pearson (1901) e uma descrição de métodos computacionais práticos veio muito mais tarde com Hotelling, que utilizou com o propósito determinado de analisar as estruturas de correlação. Como técnica estatística de análise multivariada transforma linearmente um conjunto original de variáveis, inicialmente correlacionadas entre si, num conjunto substancialmente menor de variáveis não correlacionadas que contém a maior parte da informação do conjunto de dados original (HONGYU *et al.*, 2016). Dessa forma a PCA permite uma redução no número de atributos utilizados (JOLLIFFE; CADIMA, 2016).

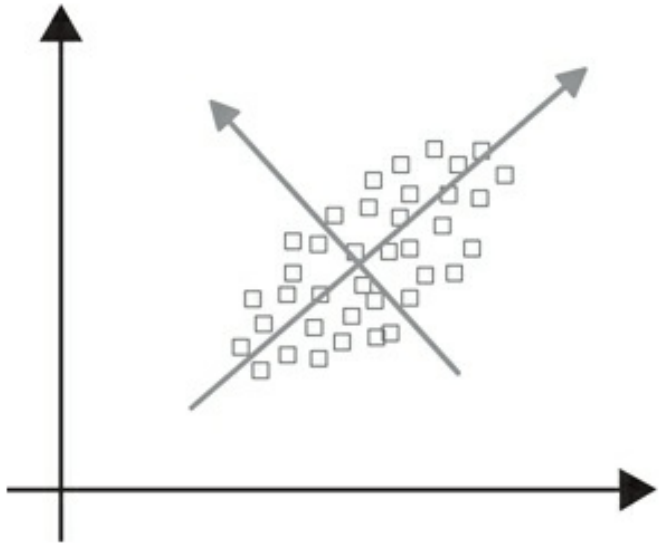
Vale destacar que os componentes principais apresentam propriedades importantes, onde cada componente principal é uma combinação linear de todas as variáveis originais, independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados (JOHNSON *et al.*, 2002); (HONGYU, 2015).

Para Hongyu *et al.* (2016), o objetivo principal da PCA é explicar a estrutura da variância e covariância de um vetor aleatório, composto de variáveis aleatórias, por meio de combinações lineares das variáveis originais. Essas combinações lineares são chamadas de

componentes principais e são não correlacionadas entre si.

Por outro lado, Jolliffe e Cadima (2016) salienta que a PCA reduz a dimensionalidade de um conjunto de dados, sendo obtido através da transformação dos dados originais em um novo conjunto de variáveis, chamadas componentes principais, correlacionadas e organizadas de forma que as primeiras componentes contêm a maior parte da variância contida no conjunto de dados original.

Figura 9 – Componentes Principais de Uma Base de Dados em 2D



Fonte: Castro e Ferrari (2016).

A Figura 9 apresenta um exemplo em que os dois componentes principais dos dados estão apresentados em um gráfico (CASTRO; FERRARI, 2016). Segundo Castro e Ferrari (2016), a PCA realiza um mapeamento linear dos dados em um espaço de dimensão menor, para que a variância dos dados nesse espaço seja maximizada. Na prática, a matriz de covariância dos dados é construída e seus autovetores são calculados, em que os autovetores que correspondem aos maiores autovalores podem ser usados para reconstruir uma grande fração da variância dos dados originais.

3.4.2.2 Sequential Forward Selection (SFS)

Para Marcano-Cedeño *et al.* (2010) o principal objetivo dos métodos de seleção de atributos é escolher um número de atributos do conjunto de atributos extraídos que produza um erro de classificação mínimo, podendo inclusive ser utilizado para seleção de características baseado na combinação de classificadores, como por exemplo MLP e o SFS. Assim o SFS começa

a partir de um conjunto de atributos vazio e adiciona gradualmente atributos selecionados por uma função de avaliação, que minimiza a taxa de erro de classificação. Deste modo, a cada iteração, o atributo a ser incluído no conjunto de atributos é selecionado entre os atributos disponíveis restantes do conjunto que não foram adicionados. Portanto, o novo conjunto de atributos estendidos deve produzir uma taxa de erro mínima de classificação em comparação com a adição de qualquer outro atributo. O SFS é amplamente utilizado por sua simplicidade e velocidade (MARCANO-CEDEÑO *et al.*, 2010). Dessa forma, o SFS, seleciona atributos que são adicionados sequencialmente a um conjunto candidato vazio até que a adição de outros atributos não diminua o critério (VISALAKSHI; RADHA, 2014).

Vale destacar que o SFS é um algoritmo de pesquisa ganancioso com uma carga computacional relativamente baixa, extraindo o subconjunto de atributos e maximizando a eficiência do subconjunto de atributos (LIOGIENĖ; TAMULEVIČIUS, 2015).

3.4.2.3 Sequential Backward Selection (SBS)

O SBS é um algoritmo de seleção de atributos que deduz o espaço do atributo em um subespaço com latência mínima no desempenho do classificador, reduzindo o tempo de execução do modelo, podendo assim melhorar a capacidade preditiva do classificador. Vale salientar que o SBS seleciona atributos que serão removidos sequencialmente de um conjunto completo de candidatos. Para calcular qual atributo será eliminado em cada etapa, é definida uma função do critério que é calculada através da diferença de desempenho do classificador antes e depois da eliminação de um determinado atributo. Assim, o atributo que é eliminado em cada etapa pode ser definido como o atribito que maximiza o critério (HAQ *et al.*, 2019). O SBS, assim como o SFS, podem ser utilizados baseados na combinação com um classificador como o MLP, ELM, SVM entre outros.

Deve-se destacar que o SFS e o SBS são algoritmos de busca com custo computacional relativamente baixo que melhoram a eficiência do classificador diminuindo o número de recursos usados. Além disso, podem melhorar a capacidade preditiva do classificador (HAQ *et al.*, 2019).

3.4.3 Importância dos Atributos

A seguir serão apresentadas os parâmetros entropia e índice de diversidade de Gini utilizados como medidas de importância dos atributos.

3.4.3.1 Entropia (Ganho de Informação)

A entropia tem origem na teoria da informação de Claude Shannon, campo dedicado ao estudo da quantificação da informação para comunicação (SHANNON, 1948). Dessa forma, para quantificar a quantidade de informação é necessário o uso de probabilidade. Assim, a teoria da informação afirma que: (i) a entropia está relacionada com a frequência dos símbolos transmitidos ou mesmo armazenados; (ii) a entropia será zero quando existir a certeza da transmissão de um único símbolo; (iii) a entropia máxima é obtida quando a frequência dos símbolos é equiprovável; (iv) a entropia não está relacionada ao significado ou a questões subjetivas (PINEDA *et al.*, 2006).

De acordo com Silipo e Melcher (2019), a entropia é um conceito utilizado para medir a pureza de um conjunto de dados. Assim a entropia pode ser uma medida de pureza, desordem ou informação.

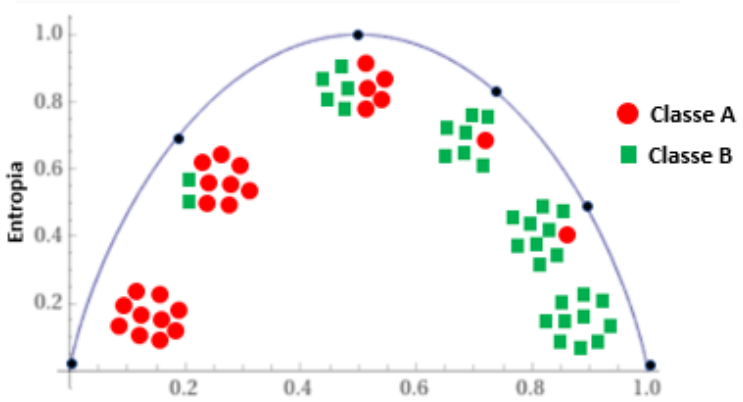
Considerando as classes alvo como possíveis status de um ponto em um conjunto de dados, a entropia de um conjunto de dados pode ser definida matematicamente como a soma de todas as probabilidades de cada classe multiplicada pelo logaritmo dela. Ressalta-se que para problemas de classificação binária, o intervalo da entropia fica entre 0 e 1 (SILIPO; MELCHER, 2019).

Segundo Kubat (2017), o processo começa pelo cálculo da entropia do sistema onde apenas os percentuais de classe são conhecidos. Em seguida, o algoritmo calcula o ganho de informação transmitido por cada atributo. O atributo que oferece o maior ganho de informação é considerado o melhor. Maximizando o ganho de informação, a entropia possibilita encontrar os atributos que fornecerão o maior ganho de informação. Nessa perspectiva a entropia é uma medida de incerteza ligada ao conteúdo da informação (WILMOTT, 2019).

Seguem abaixo as etapas para o cálculo da entropia segundo Kubat (2017), de forma que o algoritmo encontre os atributos com os maiores ganhos de informação:

1. Calcule a entropia do conjunto de treinamento T;
2. Para cada atributo, em que divide T em subconjuntos Ti, com tamanhos relativos, faça o seguinte:
   i calcular a entropia de cada subconjunto;
   ii calcular a entropia média;
   iii calcular o ganho de informação.
3. Escolha o atributo com o maior valor de ganho de informação.

Figura 10 – Exemplo de Entropia



Fonte: Adapatado de Wilmott (2019).

A Figura 10 mostra um exemplo de entropia com duas classes representadas pelos circulos vermelhos e quadrados verdes. Nela pode-se perceber que o eixo x representa a proporção de circulos vermelhos e quadrados verdes, já o eixo y representa a entropia. No gráfico é possível perceber que a entropia é maior no meio onde observa-se a mesma quantidade de circulos e quadrados no agrupamento. Por outro lado, a entropia nas extremidades do gráfico é igual a zero pois em ambos os agrupamento tem-se a mesma quantidade de circulos e quadrados.

3.4.3.2 Índice de Diversidade de Gini

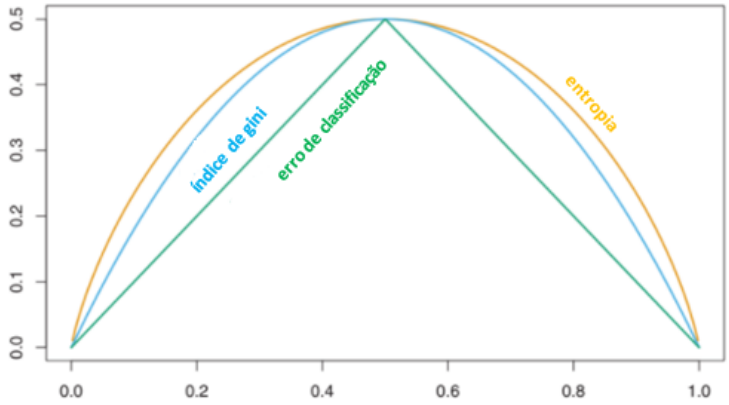
O índice de diversidade de Gini é uma medida de impureza usada em problemas de n classes. Foi desenvolvido por um matemático italiano chamado Conrado Gini como uma medida estatística no início da década de 1912. É responsável por medir o grau de heterogeneidade dos dados utilizados para medir a impureza, com valores variando no intervalo de 0 a 1. Vale salientar que quanto mais próximo de zero maior será a pureza (BASGALUPP, 2010; VILAS, 2020).

De acordo com Charu (2015), o índice de diversidade de Gini é comumente usado para medir o poder discriminativo de um determinado atributo. Normalmente é usado para variáveis categóricas, podendo ser generalizado para atributos numéricos pelo processo de discretização (SILIPO; MELCHER, 2019).

Para Xu (2003), o índice de Gini pode ser usado para medir a dispersão de uma distribuição de renda, consumo, riqueza, ou uma distribuição de qualquer outro tipo. Já para Santos (2016), o índice de Gini gera um índice de dispersão estatístico que mede a heterogeneidade dos dados que é calculado subtraindo a soma das probabilidades quadradas de cada classe por um

(JUNIOR, 2015). Salienta-se que quando, o critério de Gini é utilizado tende-se a isolar num ramo os dados que representam a classe mais frequente (SANTOS, 2016).

Figura 11 – Exemplo de Uso do Índice de Diversidade de Gini



Fonte: Adaptado e traduzido de Hastie *et al.* (2009), Wakefield (2013).

A Figura 11 faz uma comparação das medidas de impureza para classificação binária, com as classes A e B.

4 MATERIAL E MÉTODOS

Este capítulo objetiva descrever a metodologia utilizada nesse trabalho, sendo dividida em duas partes: descrição da base de dados e etapas do sistema de classificação. Para este trabalho, foram realizados testes baseados em 4 fases para cada base de dados utilizada: (i) todos os classificadores foram testados usando a normalização *z-score*; (ii) todos os classificadores foram testados com a técnica PCA; (iii) somente os dois classificadores que obtiveram os melhores resultados na fase (i) e (ii) foram testados (SVM e MLP) com as técnicas (SFS e SBS) e por fim; (iv) foram analisadas a Entropia (ganho de informação) e o Índice de Diversidade de Gini para a determinação dos atributos mais relevantes.

4.1 Descrição da Base de Dados

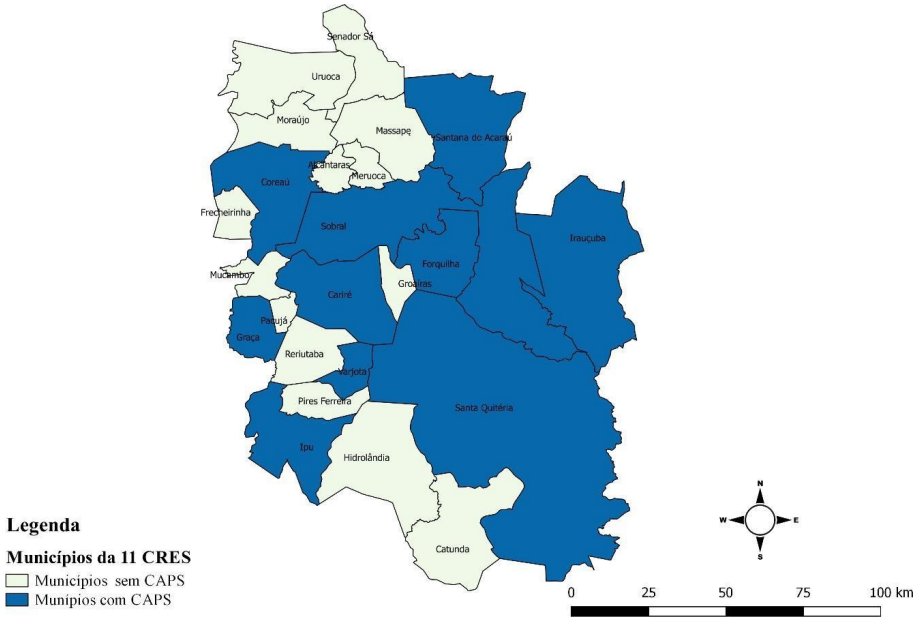
Os dados utilizados nesse trabalho fazem parte de uma pesquisa denominada "Saúde mental e o risco de suicídio em usuários de drogas", com parecer favorável do Comitê de Ética em Pesquisa no ano de 2018 e nº 2.739.560, coordenada pela professora Dra. Eliany Nazaré Oliveira e financiada pela Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP) através do Edital Programa de Bolsas de Produtividade em Pesquisa, Estímulo à Interiorização e Inovação Tecnológica (BPI). Vale ressaltar que o estudo contou com a participação de 07 bolsistas de Iniciação científica, 05 apoiados pela FUNCAP e 02 vinculados ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Importante enfatizar que do projeto em questão resultaram 03 dissertações de mestrado, 02 produzidas no Mestrado Acadêmico em Saúde da Família (UFC) e 01 no Mestrado Profissional em Saúde da Família (UVA/RENASF).

Os dados foram coletados junto a 605 participantes em oito municípios do estado do Ceará que possuem serviços de saúde mental para atendimento aos usuários de drogas psicoativas, como os Centros de Atenção Psicossocial Geral (CAPS Geral), Centros de Atenção Psicossocial Álcool e Drogas (CAPS AD) e em comunidades terapêuticas.

A escolha por esses serviços, conforme Moreira (2020), deu-se devido a população ser bem delimitada, contribuindo assim para que se pudesse ter uma amostra representativa, uma vez que o público pesquisado está presente em todos os segmentos da sociedade, minimizando o viés de amostragem ou de autosseleção, onde uns teriam mais oportunidade de participar que outros.

O estudo foi realizado na 11ª Coordenadoria Regional de Saúde (CRES), referente à macrorregião de Sobral, composta por 24 municípios conforme Figura 12. Os serviços da 11ª CRES são considerados referência para o atendimento ao público-alvo do estudo nas diferentes modalidades, fundamentais para a pesquisa (MOREIRA, 2020).

Figura 12 – Municípios Pertencentes a 11ª CRES



Fonte: Moreira (2020).

A pesquisa ocorreu no período de janeiro a julho de 2019, por meio de entrevistas com três instrumentos: um formulário de perfil sociodemográfico, clínico e padrão de consumo conforme (Anexo A), SRQ-20 (Anexo B) e o PHQ-9 (Anexo C), podendo ser encontrado também em (MOREIRA *et al.*, 2020; MOREIRA, 2020). As entradas do modelo de predição apresentado neste trabalho são os dados obtidos no formulário sociodemográfico, clínico e padrão de consumo, enquanto as saídas (a predição de TMC) são os resultados do questionário SRQ-20 e (a predição da depressão) os resultados do questionário PHQ-9.

O PHQ-9 é um questionário cuja descrição encontra-se no Manual Diagnóstico e Estatístico dos Transtornos Mentais-IV (DSM-IV), para os sintomas episódicos de depressão (ASSOCIATION *et al.*, 2002). Ele é composto por nove perguntas de aplicação rápida, validadas no Brasil por Osório *et al.* (2009), com a finalidade de avaliar a periodicidade de sinais e sintomas de transtono depressivo nas últimas duas semanas.

Dessa forma o PHQ-9 avalia os nove sintomas para o episódio depressivo caracterizados por problemas com o sono, cansaço ou falta de energia, distúrbios na concentração, humor deprimido, lentidão ou inquietação excessiva e pensamentos suicidas, anedonia, mudança

no apetite ou peso e sentimento de inutilidade (SANTOS *et al.*, 2013). Destaca-se ainda que o questionário possui uma décima pergunta para avaliar a influência desses sintomas no desempenho de atividades do cotidiano (SANTOS *et al.*, 2013). Para Bergerot *et al.* (2014) cada umas das perguntas podem ser pontuados em uma escala de *likert* em quatro pontos que vairam de 0 (nenhuma vez) a 3 (quase todos os dias), com a pontuação podendo ser de 0 a 30 com indicação positiva a partir do maior valor ou igual a 10.

De acordo com Gonçalves *et al.* (2008) o SRQ-20 foi validado em vários países além do Brasil, mostrando bom desempenho na identificação de casos positivos e negativos, assim como a efetividade para ser usado em larga escala. Ainda conforme os autores, o SRQ-20 é composto por 20 perguntas e utiliza-se como ponto de corte 7/8 para ambos os sexos para a existência do TMC, com destaque para o rastreamento e não diagnóstico, com sensibilidade de 86,33%, especificidade de 89,31%, com valores de 76,43% e 94,21% para valores preditivos positivos e negativos respectivamente.

O formulário sociodemográfico clínico e de padrão de consumo, busca caracterizar os participantes por meio de variáveis como sexo, idade, cor da pele/raça autorreferida, religião, escolaridade, ocupação, estado civil, número de filhos, renda familiar, número de residentes no situação familiar e habitacional. Quanto aos aspectos clínicos, a principal hipótese diagnóstica e a presença de comorbidades clínicas ou psiquiátricas também são investigadas na ficha, bem como sua relação com o uso de substâncias psicoativas. O SRQ-20 e o PHQ-9, por outro lado, são formulários bem estabelecidos contendo perguntas para rastreamento de transtornos não psicóticos (MOREIRA, 2020).

Uma parte desta base de dados foi analisada em Moreira *et al.* (2020) usando estatísticas inferenciais e testes de associação, comparação e correlação. Este estudo encontrou uma alta taxa de pessoas com humor deprimido, ansiedade e sintomas somáticos. Além disso, os principais fatores de risco social observados foram o sexo feminino e ser jovem, ser católico ou evangélico, ter companheiro fixo e filhos foram apontados como fatores de proteção. Em Moreira *et al.* (2020) e Moreira (2020), também foi verificado que o uso de substâncias psicoativas tem grande impacto no risco de TMC assim como na depressão.

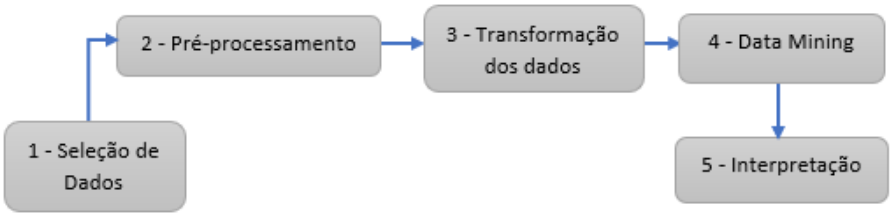
A coleta dos dados foi orientada conforme a resolução 466/2012 do Conselho Nacional de Saúde que aponta que as pesquisas com seres humanos devem ser desenvolvidas preferencialmente em indivíduos com autonomia plena, ou seja, pessoas independentes e cujos impulsos e desejos pessoais possam ser controlados, moderados e aprovados por ela mesma

(BRASIL, 2012).

4.2 Etapas do Modelo de Classificação

O modelo de predição apresentado neste trabalho é baseado em um sistema de classificação que segue as etapas do método KDD para DM, com as seguintes etapas: (1) seleção dos dados; (2) pré-processamento; (3) transformação de dados; (4) mineração de dados e (5) interpretação (FAYYAD *et al.*, 1996). Um esquema simplificado das etapas do modelo de predição é mostrado na Figura 13.

Figura 13 – Etapas do Método KDD



Fonte: Próprio autor.

Vale destacar que, para Castro e Ferrari (2016), conhecer e preparar de forma adequada os dados para análise pode tornar todo o processo de mineração muito mais eficiente e eficaz. Por outro lado, dados mal ou não pré-processados podem inviabilizar uma análise ou invalidar um resultado.

Na Etapa 1, iniciou-se o processo de pré-processamento onde algumas amostras com atributos ausentes foram descartadas, bem como algumas questões encontradas nos questionários que não estavam relacionadas à classificação como os atributos, naturalidade, tipo de serviço, município de residência. Além disso, algumas informações redundantes também foram descartadas. Exemplos de informações redundantes são data e idade de nascimento, renda familiar em reais e renda familiar em termos de salário mínimo. Além disso, são exemplos de dados irrelevantes para o problema investigado (TMC e depressão), face à classificação, o município de nascimento, o município de residência e o tipo de serviço público onde os dados foram recolhidos.

Na Etapa 2, as respostas do questionário foram codificadas e as inconsistências corrigidas. Algumas questões como, por exemplo, cor da pele/raça autorreferida, religião e tipo de moradia, foram codificadas em diversas variáveis binárias, uma para cada cor/raça da pele, cada religião e tipo de moradia, já outras como escolaridade foram codificadas de 0 a 7. Após o pré-processamento, os bancos de dados continham 84 atributos, 605 amostras e uma classe

binária de saída representando a previsão de TMC (sim ou não) e outra representando a previsão de saída da depressão (sim ou não), obtidos respectivamente através dos resultados da aplicação dos questionários SRQ-20 (Anexo B) e do PHQ-9 (Anexo C). Todos os dados contidos nas bases representam valores numéricos para as escolhas realizadas pelos entrevistados.

Posteriormente, na Etapa 3, algumas técnicas de seleção/ transformação de atributos foram aplicadas. Em particular, as seguintes técnicas foram testadas: *Principal Component Analysis* (PCA), *Sequential Forward Selecion* (SFS) e *Sequential Backward Selecion* (SBS). O PCA é uma técnica de transformação de atributos cuja ideia principal é reduzir a dimensionalidade e a correlação do conjunto de dados, que, no caso considerado, consiste em um grande número de variáveis correlacionadas. Como consequência, o PCA também pode diminuir consideravelmente o tempo de processamento. Neste trabalho, o número de componentes do PCA foi escolhido com base na acurácia, após testar alguns valores para este parâmetro.

Por outro lado, como mencionado anteriormente, no método de seleção de atributos SFS, os atributos são adicionados sequencialmente ao conjunto de atributos usados até que a adição de outro atributo não aumente a acurácia do classificador (VISALAKSHI; RADHA, 2014). Em contraste, no algoritmo de seleção de atributos SBS, os atributos são removidos sequencialmente de um conjunto completo de atributos, até que a remoção de outro atributo não diminua a acurácia (VISALAKSHI; RADHA, 2014).

Em seguida, na Etapa 4, os dados são enviados para um algoritmo de classificação, para a realização da previsão das classes dos participantes, relacionada à presença de TMC e da depressão. Nove classificadores foram testados neste trabalho: KNN, MLP, SVM, LDA, QDA, *Naive* LDA, *Naive* QDA, ELM e RF. Os classificadores *Naive* LDA e *Naive* QDA são versões do LDA e QDA padrão que assumem que os atributos estão correlacionados. Os hiperparâmetros de cada classificador foram obtidos por meio de testes, com base na acurácia obtida durante a fase de teste, por meio da técnica de validação cruzada *k-fold* com k = 10 *folds*. Os classificadores foram testados com ou sem a normalização do *z-escore*. Este método normaliza o desvio padrão e remove a média de cada atributo. Os melhores resultados foram obtidos com a normalização do *z-score*, portanto, os resultados apresentados correspondem aos casos com a normalização do *z-score*.

Como será visto nas seções de resultados para o TMC e depressão, as técnicas de MLP e SVM forneceram as melhores acurácias. Por esse motivo, a maior parte dos resultados foi obtida por meio desses classificadores.

No Passo 5 do método KDD, os resultados são visualizados e analisados. A figura de mérito utilizada neste trabalho é a taxa média da acurácia, obtida na classificação binária. Com relação à análise da importância das características, o ganho de informação, com base na entropia, e o índice de diversidade de Gini, também conhecido como índice de Gini-Simpson (KUBAT, 2017; CHARU, 2015; BREIMAN, 2001) foram utilizados para determinar os melhores atributos.

5 RESULTADOS

Neste capítulo, são apresentados os resultados e discussões sobre os resultados adquiridos durante o processo de treinamento e teste com as bases de dados relacionadas ao TMC e à depressão. Os resultados foram obtidos utilizando-se de nove classificadores e demais técnicas de AM. Os resultados a seguir representam as taxas médias de acertos com a técnica de validação cruzada *k-fold* com k=10. Como mencionado anteriormente, os resultados foram obtidos usando dois bancos de dados, um sobre TMC e outro sobre depressão, com normalização *z-score*, PCA, SBS, SFS e os parâmetros entropia e índice de diversidade de Gini.

5.1 Resultados do Experimento com a Base de Dados TMC

Os resultados a seguir apresentados são referentes ao TMC e estão divididos em três partes: (i) comparação entre classificadores; (ii) resultados com seleção/transformação de atributos; e (iii) importância dos atributos.

5.1.1 Comparação Entre Classificadores

O primeiro experimento teve como objetivo comparar o desempenho de vários algoritmos de classificação, sem utilizar as técnicas de seleção/transformação de características PCA, SFS e SBS. A Tabela 2 mostra as acurácias obtidas pelas nove técnicas de classificação testadas. Conforme anteriormente mencionado, muitas simulações foram realizadas para ajustar os hiperparâmetros desses classificadores. A Tabela 3 mostra os hiperparâmetros dos classificadores que forneceram os melhores resultados.

Tabela 2 – Comparação Entre Classificadores - sem Seleção/Transformação de Atributos

Classificador	Média de Acertos
MLP	77,21%
SVM	76,42%
ELM	74,75%
KNN	74,59%
Random Forest	73,05%
LDA	71,15%
QDA	70,98%
LDA (Versão *Naive Bayes*)	67,38%
QDA (Versão *Naive Bayes*)	67,21%

Fonte: elaborado pelo autor (2021).

Pode-se observar na Tabela 2 que as técnicas MLP e SVM obtiveram as maiores acurácias, com 77,21% e 76,42% respectivamente. Este resultado não é surpreendente, já que o MLP e o SVM são dois dos classificadores mais eficientes.

Por outro lado, os classificadores que obtiveram os piores resultados foram o LDA e o QDA na versão *Naive*, com acurácias de 67,38% e 67,21% respectivamente. Isso se deve ao fato de que esses classificadores assumem que os atributos são descorrelacionados, o que não é uma hipótese válida para o banco de dados em uso.

Tabela 3 – Hiperparâmetros Usados nos Classificadores

Classificador	Hiperparâmetro
MLP	2 camadas ocultas com 10 neurônios cada, função de ativação: *linear saturated*, *scaled conjugate gradient backpropagation*, *learning rate* $= 0,2$, *batch size* $= 1$, número de *epochs* $= 30$
SVM	*kernel* Polinomial (*non-homogeneous*) de grau 1, $C = 1$, *KernelScale*: 1/sqrt(2*0.01), *one-vs-one*
ELM	1 camada oculta com 25 neurônios
KNN	$K = 40$
RF	número de *seeds* $= 1$, número de *trees*$= 400$
LDA	-
QDA	-
Naive LDA	-
Naive QDA	-

Fonte: elaborado pelo autor (2021).

5.1.2 Resultados com Seleção/Transformação de Atributos

Os resultados a seguir têm como objetivo avaliar o desempenho dos algoritmos de classificação utilizando o método de transformação de características PCA e as técnicas de seleção de características SFS e SBS. A Tabela 4 mostra as acurácias dos nove classificadores testados usando o PCA, com a segunda coluna mostrando o número de componentes usados. Para este resultado, para cada classificador, foram testados vários valores para o número de componentes do PCA como por exemplo: 1, 3, 5, 10, 15, ..., 84, sendo escolhido o valor que proporcionou a maior acurácia.

Assim como na Tabela 2, os classificadores MLP e SVM proporcionaram os melhores

Tabela 4 – Comparação Entre Classificadores - com PCA

Classificador	Nº Comp. PCA	Média de Acertos
MLP	81	77,70%
SVM	80	77,54%
ELM	35	75,90%
KNN	55	75,90%
QDA	60	73,61%
Random Forest	55	73,05%
LDA (Versão *Naive Bayes*)	75	72,46%
LDA	75	71,64%
QDA (Versão *Naive Bayes*)	25	71,64%

Fonte: elaborado pelo autor (2021).

resultados, atingindo 77,70% e 77,54% de acurácia, respectivamente, enquanto o LDA e o QDA na versão *Naive* obtiveram os piores resultados, com 71,64% de acurácia para ambas as técnicas.

Além disso, ao comparar os resultados das Tabelas 2 e 4 pode-se concluir que todos os classificadores alcançaram as melhores acurácias com PCA, quando comparados ao caso sem PCA. Contudo, para a maior parte dos classificadores, o ganho na acurácia fornecido pelo uso do PCA pode ser considerado pequeno. Em particular, para o método RF, o PCA não forneceu um ganho de desempenho. Por outro lado, para o QDA na versão *Naive*, o uso de PCA melhorou a acurácia em 4,4%. Deve ser destacado que, mesmo quando o PCA não oferece desempenho ganho, ele permite uma redução no número de atributos, diminuindo o custo computacional.

Por conta dos resultados obtidos, decidiu-se resumir os testes aos classificadores MLP e SVM, por terem proporcionado as melhores acurácia. Os próximos resultados foram obtidos usando as técnicas de seleção de atributos SFS e SBS. As Tabelas 5 e 6 mostram as matrizes de confusão com as classes verdadeiras e preditas, obtidas pelo SVM, juntamente com as técnicas SFS e SBS, respectivamente, enquanto as Tabelas 7 e 8 mostram as matrizes de confusão do MLP com o SFS e SBS, respectivamente.

Pode-se ver nas Tabelas 5 e 6 que ambas as técnicas de seleção de atributos foram capazes de aumentar a acurácia do SVM. Em particular, o SVM com o SBS atingiu uma acurácia de 82,81%, que é 6,39% maior do que a taxa obtida apenas pelo SVM e 5,27% maior que com PCA. Esse ganho de desempenho se deve ao fato de que as técnicas SFS e SBS selecionam subconjuntos de atributos que são mais adequados ao problema. Uma conclusão semelhante pode ser tirada das Tabelas 7 e 8, no entanto, o ganho de acurácia obtido com o MLP é menos significativo, com o SBS atingindo uma taxa de 78,68% com ganho de 1,47% maior que a taxa obtida apenas pelo MLP e 0,98% maior que com PCA.

Pode-se notar também que o SBS proporcionou melhores resultados que o SFS,

Tabela 5 – Matriz de Confusão com Classes Verdadeiras e Preditas (SVM - SFS)

		Classes Verdadeiras		
		Com TMC	Sem TMC	Média
Classes	Com TMC	**387**	67	
Preditas	Sem TMC	58	**93**	
Acurácia (%)		86,97	58,13	79,34

Fonte: elaborado pelo autor (2021).

Tabela 6 – Matriz de Confusão com Classes Verdadeiras e Preditas (SVM - SBS)

		Classes Verdadeiras		
		Com TMC	Sem TMC	Média
Classes	Com TMC	**395**	54	
Preditas	Sem TMC	50	**106**	
Acurácia (%)		88,76	66,25	82,81

Fonte: elaborado pelo autor (2021).

para ambos os classificadores. No entanto, o SBS usou muito mais atributos do que o SFS. Na verdade, o SFS selecionou apenas 4 atributos, tanto para o SVM (sexo, idade, CID10-F19 - transtorno mental devido ao uso de múltiplas drogas e sem ocupação) quanto para o MLP (estado civil/amasiado, renda familiar, tipo moradia/ocupação/invasão e CID10-F17- transtorno mental devido ao uso do fumo), enquanto o SBS excluiu apenas 2 atributos para o SVM (religião e tipo moradia) e 1 atributo para o MLP (religião/nenhuma). Também deve-se notar que o SVM com SFS atingiu uma acurácia de 79,34% com apenas 4 atributos.

Outra observação que pode ser feita a partir dessas matrizes de confusão é que a classe que representa o diagnóstico positivo para TMC obteve melhores acurácias do que a classe que representa a ausência de TMC, ou seja, a sensibilidade é maior que a especificidade nas Tabelas 5, 6, 7 e 8.

Isso significa que a taxa de falsos negativos é menor do que a taxa de falsos positivos. Do ponto de vista das políticas públicas de saúde, isso pode ser visto como uma característica desejada para um sistema de previsão, pois tem menor probabilidade de não detectar TMC em pessoas que estão em um grupo de risco. No melhor caso evidenciado na Tabela 6, a taxa de falsos negativos é de 11,24%.

5.1.3 Resultados com Importância dos Atributos

Nesta subseção, são apresentados alguns resultados que analisam a importância dos atributos no processo de classificação das amostras da base de dados considerada, a fim de determinar quais fatores são mais relevantes para a predição de TMC. Dois dos parâmetros mais

Tabela 7 – Matriz de Confusão com Classes Verdadeiras e Preditas (MLP - SFS)

		Classes Verdadeiras		
		Com TMC	Sem TMC	Média
Classes	Com TMC	**383**	68	
Preditas	Sem TMC	62	**92**	
Acurácia (%)		86,07	57,60	78,51

Fonte: elaborado pelo autor (2021).

Tabela 8 – Matriz de Confusão com Classes Verdadeiras e Preditas (MLP - SBS)

		Classes Verdadeiras		
		Com TMC	Sem TMC	Média
Classes	Com TMC	**380**	64	
Preditas	Sem TMC	65	**96**	
Acurácia (%)		85,39	60,00	78,68

Fonte: elaborado pelo autor (2021).

comuns usados para medir a relevância das características são o ganho de informação, com base na entropia, e o índice de diversidade de Gini (KUBAT, 2017; CHARU, 2015; BREIMAN, 2001). Esses parâmetros são comumente usados no projeto de classificadores de árvore de decisão, como o RF, mas também indicam a relevância dos atributos para uma tarefa de classificação.

As Tabelas 9 e 10 apresentam, respectivamente, os 10 maiores ganhos de informação e índices de diversidade de Gini dos atributos, com as respectivas descrições dos atributos. Nessas tabelas, os atributos “CID10-F19” e “CID10-F17” representam, respectivamente, diagnósticos de transtorno mental por uso de múltiplas drogas e tabagismo, e os atributos “problema com drogas psicoativas: tabaco” e “problema com drogas psicoativas: cocaína/crack” onde os participantes tiveram que informar se tiveram recaídas devido a derivados do tabaco e/ou cocaína/crack, respectivamente. Ressalta-se que os atributos do tipo “SPA problema” respresentam o fator que o participante da pesquisa identificou como aquele que causou maior problema ou está relacionado ao ápice da recaída, ou seja, determinando o retorno ao uso de SPA. Por outro lado, que os atributos do tipo “SPA mais utilizada” estão relacionados à SPA de maior consumo.

A primeira conclusão que pode ser tirada desses resultados é que as Tabelas 9 e 10 apresentam uma grande semelhança entre as melhores características em termos de ganho de informação e o índice de diversidade de Gini. Além disso, os resultados mostram uma alta influência do uso de substâncias psicoativas, sendo o uso de cocaína/crack o fator mais relevante para o TMC. O uso de bebidas alcoólicas e derivados do tabaco também é um fator importante para predizer o risco de TMC. Com efeito, o consumo de múltiplas substâncias psicoativas pode interferir diretamente na saúde mental dos usuários, aumentando a probabilidade

Tabela 9 – Os 10 Maiores Ganhos de Informação dos *Feactures*

Ganho de Informação	Descrição do Atributo
0.0165	depressão
0.0175	CID10-F19 (transtorno mental devido ao uso de múltiplas drogas)
0.0187	transtornos gastrointestinais
0.0191	SPA problema: derivados do tabaco
0.0193	idade
0.0209	SPA mais utilizada: álcool
0.0222	SPA problema: cocaína/crack
0.0223	sem ocupação
0.0241	CID10-F17 (transtorno mental devido ao uso do fumo)
0.0324	SPA mais utilizada: cocaína/crack

Fonte: elaborado pelo autor (2021).

Tabela 10 – Os 10 Maiores Índices de Diversidade Gini dos *Feactures*

Índice de Gini	Descrição do Atributo
0.3724	depressão
0.3745	renda familiar
0.3768	CID10-F19 (transtorno mental devido ao uso de múltiplas drogas)
0.3772	SPA problema: derivados do tabaco
0.3778	SPA problema: cocaína/crack
0.3778	idade
0.3780	sem ocupação
0.3793	SPA mais utilizada: álcool
0.3800	CID10-F17 (transtorno mental devido ao uso do fumo)
0.3812	SPA mais utilizada: cocaína/crack

Fonte: elaborado pelo autor (2021).

de rompimento ou fragilização das relações sociais, ocasionando redução da autoestima e, consequentemente, sentimento de solidão (DALGALARRONDO, 2018). Além disso, o uso de substâncias psicoativas pode prejudicar o tratamento clínico do indivíduo, causando um maior risco para o agravamento do TMC (DALGALARRONDO, 2018).

Além disso, as Tabelas 9 e 10 mostram que não ter ocupação (sem emprego) é uma importante fator de risco, assim como a idade dos participantes, onde ser mais jovem constitui fator de risco. Esses resultados estão de acordo com a conclusão de (MOREIRA *et al.*, 2020). Ambas as tabelas também indicam que a depressão é um fator de risco para TMC. Em (ADAN *et al.*, 2017), observou-se que existe uma relação direta entre depressão e TMC em usuários de substâncias psicoativas. Na verdade, a depressão e o uso de substâncias psicoativas interferem diretamente na qualidade de vida desses indivíduos, principalmente na saúde física, social e mental, contribuindo para a presença de TMC (ADAN *et al.*, 2017). Além disso, o transtorno relacionado ao uso de múltiplas substâncias e os distúrbios gastrointestinais também

são importantes fatores preditivos para TMC. Portanto, deve-se destacar o elevado número de fatores relevantes e importantes para o TMC, o que requer intervenção precoce para melhor prognóstico. Por fim, vale ressaltar que, enquanto a entropia do banco de dados é igual a 0,833, a soma dos ganhos de informação de todos os 84 atributos é igual a 0,440, o que pode ser considerado um ganho de informação significativo.

5.2 Resultados do Experimento com a Base de Dados Depressão

Esta subseção apresenta os resultados obtidos através dos testes realizados com a base de dados depressão, divididos em três etapas, a saber: (i) comparação entre classificadores; (ii) resultados com seleção/transformação de atributos; e (iii) importância de atributos.

5.2.1 Comparação Entre Classificadores

O primeiro experimento realizado com a base de dados foi a comparação entre os classificadores com o objetivo de verificar os desempenhos alcançados pelos nove classificadores, sem a utilização de qualquer tipo de técnica de seleção ou transformação de características como o PCA, SBS ou SFS. A Tabela 11 mostra a acurácia obtida pelas técnicas de classificação testadas.

Tabela 11 – Comparação Entre Classificadores - sem Seleção/Transformação de Atributos

Classificador	Média de Acertos
MLP	72,46%
SVM	72,30%
ELM	70,82%
Random Forest	70,57%
KNN	69,67%
QDA	67,70%
LDA	65,90%
QDA (Versão *Naive Bayes*)	64,59%
LDA (Versão *Naive Bayes*)	63,93%

Fonte: elaborado pelo autor (2021).

Para tal finalidade, foram realizadas várias simulações com os ajustes de vários hiperparâmetros dos classificadores. A Tabela 12 mostra todos os hiperparâmetros testados para os classificadores que fornceram as melhores acurácias para a base de dados sobre a depressão. Pode-se observar na Tabela 11 que as técnicas SVM e MLP, obtiveram os melhores resultados com as acurácias de 72,46% e 72,30%, tal como observado nos experimentos realizados com a base de dados sobre TMC. Esse resultado corrobora o fato de que esses dois classificadores estão entre os mais

eficientes dentro da área da AM.

Em contrapartida, percebeu-se que os classificadores QDA e LDA nas versões *Naive* obtiveram os piores resultados com acurácias de 64,59% e 63,93%, pois assumem que os atributos são descorrelacionados, o que não se configura uma hipótese válida, tal como para o banco de dados anterior (TMC). Ademais, comparando os resultados das Tabela 4 e 11, pode-se observar que as acurácias obtidas com a base de dados depressão são significativamente menores que aquelas obtidas com a base de dados TMC, indicando que a depressão é mais difícil de ser modelada por técnicas de AM, quando se usa os atributos considerados sem a utilização de técnicas de seleção de atributos, por apresentar sintomas com aspectos subjetivos para suas características intrigantes e paradoxais, com questões não resolvidas sobre sua natureza e classificação (BECK; ALFORD, 2016).

Tabela 12 – Hiperparâmetros Usados nos Classificadores

Classificador	Hiperparâmetro
MLP	2 camadas ocultas com 10 neurônios cada, função de ativação: *tangente hiperbólica)*, *resilient backpropagation*, *learning rate* $= 0,1$, *batch size* $= 1$, número de *epochs* $= 20$
SVM	*kernel* Gaussiano, $C = 0,89$, *KernelScale*: 1/sqrt(2*0.02), *one-vs-one*
ELM	1 camada oculta com 40 neurônios
KNN	$K = 70$
RF	número de *seeds* $= 1$, número de *trees*$= 100$
LDA	-
QDA	-
Naive LDA	-
Naive QDA	-

Fonte: elaborado pelo autor (2021).

5.2.2 Resultados com Seleção/Transformação de Atributos

Os resultados a seguir mostram o desempenho dos classificadores utilizando o método PCA, assim como com as técnicas de seleção de características SFS e SBS.

A Tabela 13 mostra as acurácias adquiridas com o classificadores utilizando a técnica do PCA, em que a segunda coluna mostra o número de componentes usados para cada um dos classificadores e a última coluna a média de acertos. Salienta-se que para estes resultados

Tabela 13 – Comparação Entre Classificadores - com PCA

Classificador	Nº Comp. PCA	Média de Acertos
MLP	80	71,48%
SVM	79	71,31%
ELM	40	70,82%
Random Forest	55	70,74%
KNN	30	70,49%
QDA	40	69,02%
QDA (Versão *Naive Bayes*)	40	68,52%
LDA (Versão *Naive Bayes*)	55	67,70%
LDA	35	67,05%

Fonte: elaborado pelo autor (2021).

foram testados vários valores para chegar à escolha do número de componentes, utilizando como critério de escolha aqueles que forneceram as maiores acurácias.

Conforme pode-se perceber na Tabela 13, os classificadores MLP e SVM tiveram os melhores resultados, algo evidenciado também na Tabela 11, atingindo, neste caso, acurácias de 71,48% e 71,31%. Por outro lado, os classificadores LDA Versão *Naive* e QDA obtiveram as piores acurácias, atingindo 67,70% e 67,05% respectivamente.

Ao se comparar os resultados da Tabela 11 com os da Tabela 13, percebe-se que os classificadores MLP e SVM não melhoraram suas taxas de acerto com o PCA, obtendo perdas de 0,98% e 0,99% respectivamente. Por outro lado o classificador ELM manteve a mesma taxa de acerto. Contudo, com o uso do PCA, os outros classificadores obtiveram taxas de acerto superiores àquelas adquiridas sem o PCA. Por exemplo, as versões *Naive* do QDA e LDA conseguiram melhorias de 3,93% e 3,77% respectivamente.

A seguir, são apresentados os resultados alcançados com os classificadores MLP e SVM, que proporcionaram as melhores taxas de acerto, com as técnicas de seleção de atributos SFS e SBS. As Tabelas 14 e 15 mostram as matrizes de confusão exibindo as classes verdadeiras e preditas, obtidas com o classificador SVM e as técnicas SFS e SBS, respectivamente. Na sequência tem-se as Tabelas 16 e 17 que exibem as matrizes de confusão do classificador MLP com o SFS e SBS, respectivamente.

Tabela 14 – Matriz de Confusão com Classes Verdadeiras e Preditas (SVM - SFS)

		Classes Verdadeiras		
		Com Depressão	Sem Depressão	Média
Classes	Com Depressão	**332**	41	
Preditas	Sem Depressão	78	**154**	
Acurácia (%)		80,98	78,97	80,33

Fonte: elaborado pelo autor (2021).

Tabela 15 – Matriz de Confusão com Classes Verdadeiras e Preditas (SVM - SBS)

		Classes Verdadeiras		
		Com Depressão	Sem Depressão	Média
Classes	Com Depressão	**338**	37	
Preditas	Sem Depressão	72	**158**	
Acurácia (%)		82,44	81,03	81,98

Fonte: elaborado pelo autor (2021).

Conforme observado nas Tabelas 14 e 15, o classificador SVM, com ambas as técnicas de seleção de atributos, foi capaz de aumentar a acurácia em comparação com os resultados obtidos apenas com o SVM ou com PCA, como pode ser observado nas Tabelas 11 e 13.

De forma particular, o SVM com SBS atingiu uma taxa de 81,98% com um ganho de 9,68% maior que a taxa obtida apenas pelo SVM e um ganho de 10,67% se comparado com a PCA. Sobre essa perspectiva pode-se relacionar esse resultado ao fato de que as técnicas SFS e SBS selecionam atributos que são os mais adequados ao problema em questão. Algo semelhante pode ser observado nas Tabelas 17 e 16 para os resultados da MLP com relação aos ganhos de acurácia obtidos.

Tabela 16 – Matriz de Confusão com Classes Verdadeiras e Preditas (MLP - SFS)

		Classes Verdadeiras		
		Com Depressão	Sem Depressão	Média
Classes	Com Depressão	**337**	54	
Preditas	Sem Depressão	73	**141**	
Acurácia (%)		82,20	72,31	79,01

Fonte: elaborado pelo autor (2021).

Tabela 17 – Matriz de Confusão com Classes Verdadeiras e Preditas (MLP - SBS)

		Classes Verdadeiras		
		Com Depressão	Sem Depressão	Média
Classes	Com Depressão	**334**	67	
Preditas	Sem Depressão	76	**128**	
Acurácia (%)		81,46	65,64	76,36

Fonte: elaborado pelo autor (2021).

Pode-se observar que o SBS proporcionou ganhos significados para ambos classificadores (SVM e MLP). Porém, o classificador MLP com SFS atingiu o menor ganho com 6,55% se comparado com apenas o MLP e de 7,53% se comparado com PCA, conforme Tabelas 11 e 13. Se comparado os resultados do SBS com apenas o MLP, o ganho foi de 3,90% e de 4,88% se

comparado com PCA. Mediante o exposto, o classificador SVM com a técnica de seleção de atributos SBS obteve a melhor taxa de acerto com 81,98% próximo aos 82,21% observados no melhor resultado adquirido também com o SVM e o SBS na base de dados do TMC.

5.2.3 Resultados com Importância dos Atributos

A seguir são apresentados resultados sobre a importância dos atributos durante o processo de classificação, usando a base de dados sobre a depressão, com a finalidade de determinar quais atributos são mais relevantes para a predição.

Tabela 18 – Os 10 maiores ganhos de informação dos *Feactures*

Ganho de Informação	Descrição do Atributo
0.0191	idade
0.0163	religião católica
0.0217	sem ocupação
0.0158	número de moradores domicílio
0.0250	CID10-F17 (transtorno mental devido ao uso do fumo)
0.0272	CID10-F19 (transtorno mental devido ao uso de múltiplas drogas)
0.0185	Primeiro uso: tabaco
0.0352	SPA mais utizada: cocaína/crack
0.0177	SPA problema: tabaco
0.0315	SPA problema: cocaína/crack

Fonte: elaborado pelo autor (2021).

Tabela 19 – Os 10 Maiores Índices de Diversidade Gini dos *Feactures*

Índice de Gini	Descrição do Atributo
0.4249	idade
0.4270	religião católica
0.4238	sem ocupação
0.4206	CID10-F17 (transtorno mental devido ao uso do fumo)
0.4199	CID10-F19 (transtorno mental devido ao uso de múltiplas drogas)
0.4255	primeiro uso: derivados do tabaco
0.4274	SPA mais utilizada: bebidas alcoólicas
0.4164	SPA mais utilizada: cocaina/crack
0.4256	SPA problema: derivados do tabaco
0.4188	SPA problema: cocaina/crack

Fonte: elaborado pelo autor (2021).

As Tabelas 18 e 19 apresentam os 10 maiores ganhos de informação e índice de diversidade de Gini dos atributos, seguidos das descrições dos mesmos. Pode-se perceber que ambas apresentam grande semelhança entre as características em termos de ganho de informação

e o índice de diversidade de Gini, exibindo uma grande influência de substâncias psicoativas como fatores relavantes como substâncias depressoras. Nessa perspectiva, salienta-se que os atributos do tipo “SPA problemas” observados nas Tabelas 18 e 19 representam a SPA que o participante da pesquisa identificou como problema, ou seja, aquela que causou maior problema ou foi o ápice da recaída, marcando o retorno do indivíduo ao uso de SPA. Vale ressaltar que era solicitado aos particiantes a SPA mais utilizada, o que gerava os atributos do tipo “SPA mais utilizada”. Contudo, visto que muitos participantes não conseguiam identificar somente uma SPA, eles podiam apontar duas SPAs mais utilizadas.

Nas Tabelas 18 e 19, observa-se os atributos que obtiveram melhores ganhos de informação e índice de diversidade de Gini, como, exemplo, cocaína/crack, uso dos derivados do tabaco e uso de álcool. Estes resultados apontam para o uso de bebidas alcoólicas assim como o uso de cocaína/crack como fatores relevantes para a predição da depressão. No trabalho de Schenker e Minayo (2005), os autores chegaram à conclusão que corrobora com os resultados mostrando que o uso de drogas lícitas ou ilícitas interfere de forma direta na saúde mental dos usuários, configurando-se como fatores de riscos associados a ocorrência de resultados negativos para a saúde, o bem-estar e o desempenho social.

Ademais, deve-se ressaltar que a entropia do banco de dados depressão é igual a 0,906, já a soma dos ganhos de informação dos 84 atributos é igual 0,470, considerado um ganho de informação expressivo.

Conforme mencionado anteriormente, este estudo pode subsidiar não só a discussão sobre políticas públicas voltadas para prevenção do consumo de SPA, mas também confirma a importância do apoio psicológico, com vista à elaboração ou orientação de uma política institucional de prevenção ao uso de SPA (ARADILLA-HERRERO *et al.*, 2014).

6 CONCLUSÕES E TRABALHOS FUTUROS

Este estudo é baseado no TMC e na depressão, pois sabe-se que esses transtornos estão presentes em usuários de SPA. Assim, um sistema de classificação mostra-se importante no contexto social. Nesses termos, o trabalho apresentou um sistema de classificação baseado em MD e AM para classificação de pessoas quanto ao risco de TMC e de depressão, com base no uso de substâncias psicoativas e dados socioeconômicos, com o objetivo de auxiliar no desenvolvimento de políticas públicas de saúde. O banco de dados utilizado é composto por 605 pessoas de oito municípios do estado do Ceará, no Brasil, com coleta realizada no período de janeiro a julho de 2019. Os resultados mostraram a eficácia do sistema de previsão como ferramenta auxiliar no pré-diagnóstico do TMC e da depressão. Na verdade, o classificador SVM, junto com a técnica SBS, atingiu um índice de sucesso de 82,81% para o TMC enquanto que 81,98% para a depressão, o que pode ser considerado um índice razoavelmente bom, dada a complexidade dos problemas propostos. O estudo também analisou quais características são mais importantes no processo de classificação, a fim de identificar os fatores mais significativos na previsão do risco de TMC e da depressão. Os resultados mostraram que o uso de cocaína/crack é o fator mais relevante para o TMC. Além disso, o uso de álcool e derivados do tabaco, e não ter ocupação também constituem fatores importantes para a previsão do TMC. Por outro lado os resultados para a depressão mostraram que o uso de cocaína/crack é o fator mais relevante. Além disso, o uso de álcool, derivados do tabaco e o uso de múltiplas drogas constituem-se fatores relevantes para a previsão da depressão. Por fim, quando comparado o banco de dados sobre a depressão com o TMC observou-se que o banco sobre a depressão apresentou melhor resultado com relação à entropia e à soma dos ganhos de informação, com uma melhora de 0,073 para entropia e 0,03 para a soma dos ganhos de informação.

6.1 Perspectivas Para Trabalhos Futuros

Como trabalhos futuros sugere-se a aplicação da aprendizagem profunda (AP) com a arquitetura *Generative Adversarial Network* (GANs), nas bases de dados utilizadas, a fim de investigar uma possível melhora do desempenho em comparação com as demais técnicas aplicadas neste trabalho. Além disso, pretende-se estudar o TEA a partir da montagem de um banco de dados relacionados a crianças com ou sem TEA, objetivando a aplicação de técnicas de AM com a finalidade de classificá-los em níveis de gravidade.

REFERÊNCIAS

ADAN, A.; E, J. M.-A.; GILCHRIST, G. Comparison of health-related quality of life among men with different co-existing severe mental disorders in treatment for substance use. **Health and quality of life outcomes**, BioMed Central, v. 15, n. 1, p. 1–12, 2017.

AMARAL, F. **Aprenda mineração de dados: teoria e prática**. [S.l.]: Alta Books Editora, 2016. v. 1.

APTÉ, C.; WEISS, S. Data mining with decision trees and decision rules. **Future generation computer systems**, Elsevier, v. 13, n. 2-3, p. 197–210, 1997.

ARADILLA-HERRERO, A.; TOMÁS-SÁBADO, J.; GÓMEZ-BENITO, J. Associations between emotional intelligence, depression and suicide risk in nursing students. **Nurse education today**, Elsevier, v. 34, n. 4, p. 520–525, 2014.

ASSOCIATION, A. P. *et al.* **Manual diagnóstico e estatístico de transtornos mentais: texto revisado (DSM-IV-TR)**. [S.l.]: Artmed, 2002.

ASSOCIATION, A. P. *et al.* **DSM-5: Manual diagnóstico e estatístico de transtornos mentais**. [S.l.]: Artmed Editora, 2014.

BACKES, A. R.; JUNIOR, J. J. d. M. S. **Introdução à visão computacional usando Matlab**. [S.l.]: Alta Books Editora, 2019.

BARBOSA, L. N. F.; ASFORA, G. C. A.; MOURA, M. C. de. Ansiedade e depressão e uso de substâncias psicoativas em jovens universitários. **SMAD Revista Eletrônica Saúde Mental Álcool e Drogas (Edição em Português)**, v. 16, n. 1, p. 1–8, 2020.

BARTH, N. L. **Inadimplência**. [S.l.]: NBL Editora, 2004.

BASGALUPP, M. P. **LEGAL-Tree: um algoritmo genético multi-objetivo para indução de árvores de decisão**. Tese (Doutorado) — Universidade de São Paulo, 2010.

BASSETTO, E. L.; DESTRO, J. F. Z.; FINOCCHIO, M. A. F.; MODESTO, R. A.; MARQUES, A. de S. Rede percepton multicamadas (mlp) na estimativa da fração difusa da radiação global. In: **VII Congresso Brasileiro de Energia Solar-CBENS 2018**. [S.l.: s.n.], 2020.

BECK, A. T.; ALFORD, B. A. **Depressão: causas e tratamento**. [S.l.]: Artmed Editora, 2016.

BELHUMEUR, P. N.; HESPANHA, J. P.; KRIEGMAN, D. J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, v. 19, n. 7, p. 711–720, 1997.

BERGEROT, C. D.; LAROS, J. A.; ARAUJO, T. C. C. F. d. Avaliação de ansiedade e depressão em pacientes oncológicos: comparação psicométrica. **Psico-USF**, SciELO Brasil, v. 19, n. 2, p. 187–197, 2014.

BOCANEGRA, C. W. **Procedimentos para tornar mais efetivo o uso das redes neurais artificiais em planejamento de transportes**. Dissertação (Mestrado) — Engenharia Civil - Escola de Engenharia de São Carlos, Universidade de São Paulo, 2002.

BOLDT, F. **Classifier ensemble feature selection for automatic fault diagnosis**. Tese (Doutorado) — Programa de Pós-graduação em Informática da Universidade Federal do Espírito Santo, 2017.

BONIFÁCIO, F. N. Comparação entre as redes neurais artificiais MLP, RBF e LVQ na classificação de dados. **Paraná: Universidade Estadual do Oeste do Paraná**, 2010.

BRASIL. Ministério da Saúde. Conselho Nacional de Saúde. **Resolução nº 466, de 12 de dezembro de 2012.**, Brasília: Ministério da Saúde, 2012.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BRITO, R. de; FERNANDES, C. A.; XIMENES, J. de S. Avaliação de rnas durante treinamento supervisionado para classificação de adolescentes com autismo. In: SBC. **Anais da VIII Escola Regional de Computação do Ceará, Maranhão e Piauí**. [S.l.], 2020. p. 53–60.

BRITO, R. X. de; FERNANDES, C. A. R.; AMORA, M. A. B. Análise de desempenho com redes neurais artificiais, arquiteturas mlp e rbf para um problema de classificação de crianças com autismo. **iSys-Revista Brasileira de Sistemas de Informação**, v. 13, n. 1, p. 60–76, 2019.

BUANI, B. E. Z. **Aplicação da Lógica Fuzzy kNN e análises estatísticas para seleção de características e classificação de abelhas.** Tese (Doutorado) — Universidade de São Paulo, 2010.

CASTRO, L. N. d.; FERRARI, D. G. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações**. [S.l.]: Saraiva Educação SA, 2016.

CERQUEIRA, P. H. R. **Um estudo sobre reconhecimento de padrões: um aprendizado supervisionado com classificador bayesiano**. Tese (Doutorado) — Universidade de São Paulo, 2010.

CHANG, C.-C.; LIN, C.-J. Libsvm: a library for support vector machines. **ACM transactions on intelligent systems and technology (TIST)**, Acm New York, NY, USA, v. 2, n. 3, p. 1–27, 2011.

CHARU, C. A. Data mining: The textbook. **Switzerland: Springer**, 2015.

DALGALARRONDO, P. **Psychopathology and semiology of mental disorders. (in Portuguese)**. [S.l.]: Artmed Editora, 2018.

DUDA, R. O.; HART, P. E. *et al.* **Pattern classification**. [S.l.]: John Wiley & Sons, 2006.

FALCO, C. B.; FABRI, J. M. G.; OLIVEIRA, E. B.; SILVA, A. V.; FARIA, M. G. de A.; KESTENBERG, C. C. F. Transtornos mentais comuns em residentes de enfermagem: uma análise a partir do self reporting questionnaire [mental disorders common among nursing residents: an analysis based on the self-reporting questionnaire][trastornos mentales comunes en residentes de enfermería: un análisis a partir del self reporting questionnaire]. **Revista Enfermagem UERJ**, v. 27, p. 39165, 2019.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *et al.* Knowledge discovery and data mining: Towards a unifying framework. In: **KDD**. [S.l.: s.n.], 1996. v. 96, p. 82–88.

FENERICH, A. T.; STEINER, M. T. A.; NIEVOLA, J. C.; MENDES, K. B.; TSUTSUMI, D. P.; SANTOS, B. S. dos. Diagnosis of headaches types using artificial neural networks and bayesian networks. **IEEE Latin America Transactions**, IEEE, v. 18, n. 01, p. 59–66, 2020.

FERREIRA, A.; FERREIRA, R. P.; SILVA, A. M. da; FERREIRA, A.; SASSI, R. J. Um estudo sobre previsão da demanda de encomendas utilizando uma rede neural artificial. **Blucher Marine Engineering Proceedings**, v. 2, n. 1, p. 353–364, 2016.

GHOJOGH, B.; CROWLEY, M. Linear and quadratic discriminant analysis: Tutorial. **arXiv preprint arXiv:1906.02590**, 2019.

GONÇALVES, D. M.; STEIN, A. T.; KAPCZINSKI, F. Avaliação de desempenho do self-reporting questionnaire como instrumento de rastreamento psiquiátrico: um estudo comparativo com o structured clinical interview for dsm-iv-tr. **Cadernos de Saúde Pública**, SciELO Public Health, v. 24, p. 380–390, 2008.

GUYON, I.; ELISSEEFF, A. An introduction to feature extraction. In: **Feature extraction**. [S.l.]: Springer, 2006. p. 1–25.

HAQ, A. U.; LI, J.; MEMON, M. H.; MEMON, M. H.; KHAN, J.; MARIUM, S. M. Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection. In: IEEE. **2019 IEEE 5th International Conference for Convergence in Technology (I2CT)**. [S.l.], 2019. p. 1–4.

HARTMANN, J. M.; MENDOZA-SASSI, R. A.; CESAR, J. A. Depressão entre puérperas: prevalência e fatores associados. **Cadernos de Saúde Pública**, SciELO Public Health, v. 33, p. e00094016, 2017.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer Science & Business Media, 2009.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2007.

HONGYU, K. **Comparação do GGE biplot-ponderado e AMMI-ponderado com outros modelos de interação genótipo x ambiente**. Tese (Doutorado) — Universidade de São Paulo, 2015.

HONGYU, K.; SANDANIELO, V. L. M.; JUNIOR, G. J. de O. Análise de componentes principais: resumo teórico, aplicação e interpretação. **E&S Engineering and Science**, v. 5, n. 1, p. 83–90, 2016.

HOSSEINIFARD, B.; MORADI, M. H.; ROSTAMI, R. Classifying depression patients and normal subjects using machine learning techniques. In: IEEE. **2011 19th Iranian Conference on Electrical Engineering**. [S.l.], 2011. p. 1–4.

HUANG, G.-B.; ZHU, Q.-Y.; SIEW, C.-K. Extreme learning machine: theory and applications. **Neurocomputing**, Elsevier, v. 70, n. 1-3, p. 489–501, 2006.

ISHFAQUE, A.; AWAN, A. J.; RASHID, N.; IQBAL, J. Evaluation of ann, lda and decision trees for eeg based brain computer interface. In: IEEE. **2013 IEEE 9th International Conference on Emerging Technologies (ICET)**. [S.l.], 2013. p. 1–6.

JOHNSON, R. A.; WICHERN, D. W. *et al.* **Applied multivariate statistical analysis**. [S.l.]: Prentice hall Upper Saddle River, NJ, 2002. v. 5.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, The Royal Society Publishing, v. 374, n. 2065, p. 20150202, 2016.

JUNIOR, R. C. **Uso da mineração de dados na identificação de alunos com perfil de evasão do ensino superior.** 2015. Monografia (Bacharel em Ciência da Computação), UNISC (Universidade de Santa Cruz do Sul), Santa Cruz do Sul, Brasil.

KHAN, A.; WANG, K. A deep learning based scoring system for prioritizing susceptibility variants for mental disorders. In: IEEE. **2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)**. [S.l.], 2017. p. 1698–1705.

KUBAT, M. **An introduction to machine learning**. [S.l.]: Springer, 2017.

LI, Y.-j.; FAN, F.-y. Classification of schizophrenia and depression by EEG with ANNs. In: IEEE. **2005 IEEE Engineering in Medicine and Biology 27th Annual Conference**. [S.l.], 2006. p. 2679–2682.

LIMA, A. I. O.; DIMENSTEIN, M.; FIGUEIRÓ, R.; LEITE, J.; DANTAS, C. Prevalência de transtornos mentais comuns e uso de álcool e drogas entre agentes penitenciários. **Psicologia: Teoria e Pesquisa**, SciELO Brasil, v. 35, 2019.

LIMA, M. C. P.; DOMINGUES, M. d. S.; CERQUEIRA, A. T. d. A. R. Prevalência e fatores de risco para transtornos mentais comuns entre estudantes de medicina. **Revista de Saúde Pública**, SciELO Brasil, v. 40, n. 6, p. 1035–1041, 2006.

LIOGIENĖ, T.; TAMULEVIČIUS, G. Sfs feature selection technique for multistage emotion recognition. In: IEEE. **2015 IEEE 3rd Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)**. [S.l.], 2015. p. 1–4.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007.

LUCCHESE, R.; SILVA, P. C. D.; DENARDI, T. C.; FELIPE, R. L. de; VERA, I.; CASTRO, P. A. de; BUENO, A. de A.; FERNANDES, I. L. Transtorno mental comum entre indivíduos que abusam de álcool e drogas: estudo transversal. **Texto & Contexto Enfermagem**, Universidade Federal de Santa Catarina, v. 26, n. 1, p. 1–7, 2017.

MARCANO-CEDEÑO, A.; QUINTANILLA-DOMÍNGUEZ, J.; CORTINA-JANUCHS, M.; ANDINA, D. Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In: IEEE. **IECON 2010-36th annual conference on IEEE industrial electronics society**. [S.l.], 2010. p. 2845–2850.

MCGINNIS, R. S.; MCGINNIS, E. W.; HRUSCHAK, J.; LOPEZ-DURAN, N. L.; FITZGERALD, K.; ROSENBLUM, K. L.; MUZIK, M. Rapid anxiety and depression diagnosis in young children enabled by wearable sensors and machine learning. In: IEEE. **2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**. [S.l.], 2018. p. 3983–3986.

MINAYO, M. C. d. S.; SOUZA, E. R. d.; CONSTANTINO, P. **Missão prevenir e proteger: condições de vida, trabalho e saúde dos policiais militares do Rio de Janeiro**. [S.l.]: Editora Fiocruz, 2008.

MITCHELL, T. Machine learning, mcgraw-hill higher education. **New York**, 1997.

MOLINA-CORONADO, B.; MORI, U.; MENDIBURU, A.; MIGUEL-ALONSO, J. Survey of network intrusion detection methods from the perspective of the knowledge discovery in databases process. **IEEE Transactions on Network and Service Management**, IEEE, v. 17, n. 4, p. 2451–2479, 2020.

MOREIRA, R. M. M. **Transtorno Mental e o Risco de Suicídio em Usuários de Substâncias Psicoativas**. Dissertação (Dissertação de Mestrado) — Universidade Federal do Ceará, 2020.

MOREIRA, R. M. M.; OLIVEIRA, E. N.; LOPES, R. E.; LOPES, M. V. de O.; ALMEIDA, P. C. de; ARAGÃO, H. L. Transtorno mental comum em usuários de substâncias psicoativas. **Enfermagem em Foco**, v. 11, n. 1, 2020.

Muhammad, F.; Rashid, N.; Akhtar, H.; Muhammad, Z.; Gilani, S. O.; Ansari, U. Evaluation of lda, qda and decision trees for multifunctional controlled below elbow prosthetic limb using emg signals. In: **2014 International Conference on Robotics and Emerging Allied Technologies in Engineering (iCREATE)**. [S.l.: s.n.], 2014. p. 115–117.

NASCIMENTO, R. F. F.; ALCÂNTARA, E.; KAMPEL, M.; STECH, J. L.; NOVO, E.; FONSECA, L. M. G. O algoritmo support vector machines (svm): avaliação da separação ótima de classes em imagens ccd-cbers-2. **Simpósio Brasileiro de Sensoriamento Remoto**, v. 14, p. 2079–2086, 2009.

NÓBREGA, I. R. A. P. d.; LEAL, M. C. C.; MARQUES, A. P. d. O.; VIEIRA, J. d. C. M. Fatores associados à depressão em idosos institucionalizados: revisão integrativa. **Saúde em Debate**, SciELO Public Health, v. 39, p. 536–550, 2015.

OSÓRIO, F. de L.; MENDES, A. V.; CRIPPA, J. A.; LOUREIRO, S. R. Study of the discriminative validity of the phq-9 and phq-2 in a sample of brazilian women in the context of primary health care. **Perspectives in psychiatric care**, Wiley Online Library, v. 45, n. 3, p. 216–227, 2009.

PAES, B. C.; PLASTINO, A.; FREITAS, A. A. Seleção de atributos aplicada à classificação hierárquica. In: **Symposium on Knowledge Discovery, Mining and Learning-KDMiLe**. [S.l.: s.n.], 2013.

PARREIRA, B. D. M.; GOULART, B. F.; HAAS, V. J.; SILVA, S. R. da; MONTEIRO, J. C. dos S.; GOMES-SPONHOLZ, F. A. Transtorno mental comum e fatores associados: estudo com mulheres de uma área rural. **Revista da Escola de Enfermagem da USP**, v. 51, p. e03225, 2017.

PAZMIÑO-MAJI, R. A.; GARCÍA-PEÑALVO, F. J.; CONDE-GONZÁLEZ, M. Statistical implicative analysis approximation to KDD and data mining: A systematic and mapping review in knowledge discovery database framework. IARIA XPS Press, 2017.

PETERSON, L. E. K-nearest neighbor. **Scholarpedia**, v. 4, n. 2, p. 1883, 2009.

PINEDA, J. O. d. C. *et al.* **A entropia segundo Claude Shannon: o desenvolvimento do conceito fundamental da teoria da informação**. Dissertação (Dissertação de Mestrado) — Pontifícia Universidade Católica. São Paulo, 2006.

Ravì, D.; Wong, C.; Deligianni, F.; Berthelot, M.; Andreu-Perez, J.; Lo, B.; Yang, G. Deep learning for health informatics. **IEEE Journal of Biomedical and Health Informatics**, v. 21, n. 1, p. 4–21, 2017.

SAIDE, O. Depressão e uso de drogas. **Revista Hospital Universitário Pedro Ernesto (TÍTULO NÃO-CORRENTE)**, v. 10, n. 2, 2014. ISSN 1983-2567. Disponível em: <https://www.e-publicacoes.uerj.br/index.php/revistahupe/article/view/8852>.

SANTOS, A. R. d. **Identificação de faces humanas através de PCA-LDA e redes neurais SOM**. Tese (Doutorado) — Universidade de São Paulo, 2005.

SANTOS, D. D. N. d. Extração de características de rnas não-codificadores longos utilizando o algoritmo random forest. 2016. Monografia (Bacharel em Engenharia da Computação), UNB (Universidade de Brasília), Brasília, Brasil.

SANTOS, I. S.; TAVARES, B. F.; MUNHOZ, T. N.; ALMEIDA, L. S. P. d.; SILVA, N. T. B. d.; TAMS, B. D.; PATELLA, A. M.; MATIJASEVICH, A. Sensibilidade e especificidade do patient health questionnaire-9 (phq-9) entre adultos da população geral. **Cadernos de Saúde Pública**, SciELO Public Health, v. 29, p. 1533–1543, 2013.

SAPKAL, D.; MEHTA, C.; NIMGAONKAR, M.; DEVASTHALE, R.; PHANSALKAR, S. Prediction of mental disorder using artificial neural network and psychometric analysis. In: **Data Management, Analytics and Innovation**. [S.l.]: Springer, 2021. p. 369–377.

SAU, A.; BHAKTA, I. Predicting anxiety and depression in elderly patients using machine learning technology. **Healthcare Technology Letters**, IET, v. 4, n. 6, p. 238–243, 2017.

SCHENKER, M.; MINAYO, M. C. d. S. Fatores de risco e de proteção para o uso de drogas na adolescência. **Ciência & Saúde Coletiva**, SciELO Public Health, v. 10, p. 707–717, 2005.

SEMOLINI, R. **Support vector machines, inferência transdutiva e o problema de classificação**. Dissertação (Mestrado) — Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, 2002.

SHANNON, C. E. A mathematical theory of communication. **The Bell system technical journal**, Nokia Bell Labs, v. 27, n. 3, p. 379–423, 1948.

SILIPO, R.; MELCHER, K. **From a Single Decision Tree to a Random Forest**. 2019. [Online; Acessed 28-dezembro- 2020]. Disponível em: <https://www.dataversity.net/from-a-single-decision-tree-to-a-random-forest/>.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados: com aplicações em R**. [S.l.]: Elsevier Brasil, 2017.

SKAPINAKIS, P.; BELLOS, S.; KOUPIDIS, S.; GRAMMATIKOPOULOS, I.; THEODORAKIS, P. N.; MAVREAS, V. Prevalence and sociodemographic associations of common mental disorders in a nationally representative sample of the general population of greece. **BMC psychiatry**, Springer, v. 13, n. 1, p. 163, 2013.

SOENTPIET, R. *et al.* **Advances in kernel methods: support vector learning**. [S.l.]: MIT press, 1999.

THOMAZ, C. E.; KITANI, E. C.; GILLIES, D. F. A maximum uncertainty lda-based approach for limited sample size problems—with application to face recognition. **Journal of the Brazilian Computer Society**, Springer, v. 12, n. 2, p. 7–18, 2006.

VILAS, L. B. **Capital humano e desigualdade: uma análise do caso europeu**. Dissertação (Dissertação de Mestrado) — Instituto Universitário de Lisboa. Portugal, 2020.

VISALAKSHI, S.; RADHA, V. A literature review of feature selection techniques and applications: Review of feature selection in data mining. In: **2014 IEEE International Conference on Computational Intelligence and Computing Research**. [S.l.: s.n.], 2014. p. 1–6.

WAKEFIELD, J. **Bayesian and frequentist regression methods**. [S.l.]: Springer Science & Business Media, 2013.

WELLING, M. **Fisher linear discriminant analysis. department of computer science, university of toronto**. [S.l.], 2005.

WILMOTT, P. **Machine Learning: An Applied Mathematics Introduction**. Panda Ohana Publishing, 2019. ISBN 9781916081604. Disponível em: <https://books.google.com.br/books?id=f_WaxQEACAAJ>.

XU, K. How has the literature on gini's index evolved in the past 80 years? **Dalhousie University, Economics Working Paper**, 2003.

YE, J.; XIONG, T.; LI, Q.; JANARDAN, R.; BI, J.; CHERKASSKY, V.; KAMBHAMETTU, C. Efficient model selection for regularized linear discriminant analysis. In: **Proceedings of the 15th ACM international conference on Information and knowledge management**. [S.l.: s.n.], 2006. p. 532–539.

ZAGHETTO, C.; AGUIAR, L.; ZAGHETTO, A.; VIDAL, F. Projeto e implementação de uma rede neural artificial para detecção do mal-posicionamento rotacional de dedos em dispositivos de captura de impressões digitais multivista sem toque. In: SBC. **Anais Principais do XI Simpósio Brasileiro de Sistemas de Informação**. [S.l.], 2015. p. 211–218.

ANEXO A – FORMULÁRIO SOCIODEMOGRÁFICO, CLÍNICO E PADRÃO DE CONSUMO

Município: _____ Data: __/__/__ Serviço de saúde mental: _____

Nome: _____ Registro Nº: __

Pesquisador responsável: _____

ASPECTOS SOCIODEMOGRÁFICOS	
1. **Sexo**	( ) Masculino ( ) Feminino
2. **Data de nascimento**	__/_____/__
3. **Idade**	_____ anos
4. **Naturalidade**	_____
5. **Município de residência**	_____
6. **Cor de pele/raça**	( ) Branca ( ) Preta ( ) Parda ( ) Amarela ( ) Indígena
7. **Religião**	( ) Católica ( ) Evangêlico ( ) Outras: _____
8. **Escolaridade**	( ) Sem escolaridade ( ) Ensino fundamental incompleto ( ) Ensino fundamental completo ( ) Ensino médio incompleto ( ) Ensino médio completo ( ) Ensino superior incompleto ( ) Ensino superior completo ( ) Pós-graduação
9. **Ocupação**	( ) Sem ocupação ( ) Informal ( ) Aposentado(a) ( ) Estudante ( ) Formal ( ) Outra:_____
10. **Estado civil**	( ) Solteiro(a) ( ) Casado(a) ( ) Amasiado(a) ( ) Viúvo(a) ( ) Separado(a)/Divorciado(a)
11. **Nº filhos**	( ) Sim: _____ filhos ( ) Não
12. **Renda familiar**	( ) Menos de 1 salário mínimo ( ) 1 salário mínimo ( ) 1 a 3 salários mínimos ( ) 4 a 6 salários mínimos ( ) Mais de 6 salários mínimos ( ) Não quer declarar Valor exato: R$ _____
13. **Situação moradia**	( ) Casa própria ( ) Casa financiada ( ) Casa cedida ( ) Casa alugada ( ) Em situação de rua ( ) Ocupação/invasão ( ) Institucionalizado ( ) Outra:_____
14. **Nº moradores no domicílio**	_____ pessoas ( ) Não se aplica

ASPECTOS CLÍNICOS	
15. **Hipótese diagnóstica principal(CID-10)**	Especificar: _____ ( ) Não se aplica/Não sabe
16. **Comorbidades clínica**	( ) Sim: _____ ( ) Não ( ) Não se aplica/Não sabe
17. **Comorbidades psiquiátrica**	( ) Sim: _____ ( ) Não ( ) Não se aplica/Não sabe **Se sim**, antes do uso do SPA ( ) depois do uso de SPA ( )
18. **Histórico familiar de uso de SPA**	( ) Sim ( ) Não ( ) Não sabe
ASPECTOS RELACIONADOS AO CONSUMO DE SPA	
19. **Idade do primeiro uso de SPA**	_____ anos ( ) Não se aplica/Não sabe
20. **SPA de primeiro uso**	( ) derivados do tabaco ( ) bebidas alcoólicas ( ) maconha ( ) cocaína/crack ( ) inalantes ( ) hipnóticos/sedativos ( ) alucinógenos ( ) opiódes ( ) estimulantes/anfetaminas ou êxtase ( ) Outras: _____
21. **SPA mais utilizadas atualmente (de escolha)**	( ) derivados do tabaco ( ) bebidas alcoólicas ( ) maconha ( ) cocaína/crack ( ) inalantes ( ) hipnóticos/sedativos ( ) alucinógenos ( ) opiódes ( ) estimulantes/anfetaminas ou êxtase ( ) Outras: _____
22. **SPA problema**	( ) derivados do tabaco ( ) bebidas alcoólicas ( ) maconha ( ) cocaína/crack ( ) inalantes ( ) hipnóticos/sedativos ( ) alucinógenos ( ) opiódes ( ) estimulantes/anfetaminas ou êxtase ( ) Outras: _____
23. **Há quanto tempo está sem utilizar SPA)**	_____( ) horas ( ) dias ( ) meses ( ) anos

ANEXO B – *SELF-REPORTING QUESTIONNAIRE* (SRQ-20)

Serviço de saúde mental: _____ Registro Nº: ___

Por favor, leia estas instruções antes de preencher as questões abaixo. É muito importante que todos que estão preenchendo o questionário sigam as mesmas instruções.

INSTRUÇÕES

Estas questões são relacionadas a certas dores e problemas que podem ter lhe incomodado nos últimos 30 dias. Se você acha que a questão se aplica a você e você teve o problema descrito nos últimos 30 dias responda **SIM**. Por outro lado, se a questão não se aplica a você e você não teve o problema nos últimos 30 dias, responda **NÃO**.

Perguntas	**Respostas**	
1. Você tem dores de cabeça com frequência?	( ) Sim	( ) Não
2. Você tem falta de apetite?	( ) Sim	( ) Não
3. Você dorme mal?	( ) Sim	( ) Não
4. Assusta-se ou fica com medo com facilidade?	( ) Sim	( ) Não
5. Suas mãos tremem?	( ) Sim	( ) Não
6. Você se sente nervoso(a), tenso(a) ou preocupado(a)?	( ) Sim	( ) Não
7. Tem má digestão ou sofre de perturbação digestiva?	( ) Sim	( ) Não
8. Não consegue ou tem dificuldades para pensar com clareza?	( ) Sim	( ) Não
9. Sente-se infeliz ou triste ultimamente?	( ) Sim	( ) Não
10. Você tem chorado mais do que o comum?	( ) Sim	( ) Não
11. Tem dificuldades para gostar ou realizar com satisfação suas atividades diárias?	( ) Sim	( ) Não
12. Tem dificuldades para tomar decisões?	( ) Sim	( ) Não
13. Seu trabalho causa sofrimento ou tormento (Tem dificuldade de realizá-lo)?	( ) Sim	( ) Não
14. Sente-se incapaz de ter um papel útil na vida?	( ) Sim	( ) Não
15. Você perdeu ou tem perdido o interesse nas coisas?	( ) Sim	( ) Não
16. Acha que é uma pessoa inútil ou que não vale nada?	( ) Sim	( ) Não
17. Já pensou, alguma vez, em acabar com a sua vida / tentar suicídio?	( ) Sim	( ) Não
18 Você tem dores de cabeça com frequência?	( ) Sim	( ) Não
19. Tem sensações desagradáveis no estômago?	( ) Sim	( ) Não
20. Fica cansado(a) com facilidade?	( ) Sim	( ) Não

ANEXO C – QUESTIONÁRIO SOBRE A SAÚDE DO PACIENTE (PHQ-9)

Serviço de saúde mental: _____ Registro Nº: ___

AGORA VAMOS FALAR SOBRE COMO O (A) SR.(A) TEM SE SENTIDO NAS DUAS ÚLTIMAS SEMANAS.

1) Nas duas últimas semanas, quantos dias você teve pouco interesse ou pouco prazer em fazer as coisas?

(0) Nenhum dia.

(1) Menos de uma semana.

(2) Uma semana ou mais.

(3) Quase todos os dias.

2) Nas duas últimas semanas, quantos dias você se sentiu para baixo, deprimido (a) ou sem perspectiva?

(0) Nenhum dia.

(1) Menos de uma semana.

(2) Uma semana ou mais.

(3) Quase todos os dias.

3) Nas duas últimas semanas, quantos dias você teve dificuldade para pegar no sono ou permanecer dormindo ou dormiu mais do que de costume?

(0) Nenhum dia.

(1) Menos de uma semana.

(2) Uma semana ou mais.

(3) Quase todos os dias.

4) Nas duas últimas semanas, quantos dias você se sentiu cansado (a) ou com pouca energia?

(0) Nenhum dia.

(1) Menos de uma semana.

(2) Uma semana ou mais.

(3) Quase todos os dias.

5)Nas duas últimas semanas, quantos dias você teve falta de apetite ou comeu demais?

(0) Nenhum dia.

(1) Menos de uma semana.

(2) Uma semana ou mais.

(3) Quase todos os dias.

6) Nas duas últimas semanas, quantos dias você se sentiu mal consigo mesmo (a) ou achou que é um fracasso ou que decepcionou sua família ou a você mesmo(a)?

(0) Nenhum dia.

(1) Menos de uma semana.

(2) Uma semana ou mais.

(3) Quase todos os dias.

7) Nas duas últimas semanas, quantos dias você teve dificuldade para se concentrar nas coisas (como ler o jornal ou ver televisão)?

(0) Nenhum dia.

(1) Menos de uma semana.

(2) Uma semana ou mais.

(3) Quase todos os dias.

8) Nas duas últimas semanas, quantos dias você teve lentidão para se movimentar ou falar (a ponto das outras pessoas perceberem) ou, ao contrário, esteve tão agitado(a) que você ficava andando de um lado para o outro mais do que de costume?

(0) Nenhum dia.

(1) Menos de uma semana.

(2) Uma semana ou mais.

(3) Quase todos os dias.

9) Nas duas últimas semanas, quantos dias você pensou em se ferir de alguma maneira ou que seria melhor estar morto (a)?

(0) Nenhum dia.

(1) Menos de uma semana.

(2) Uma semana ou mais.

(3) Quase todos os dias.

10) Considerando as últimas duas semanas, os sintomas anteriores lhe causaram algum tipo de dificuldade para trabalhar ou estudar ou tomar conta das coisas em casa ou para se relacionar com as pessoas?

(0) Nenhum dia.

(1) Menos de uma semana.

(2) Uma semana ou mais.

(3) Quase todos os dias.