Predição da Depressão

Diego Souza, Hugo Soares, Kamilah Santos

Dom Helder Escola Superior
Ciência da Computação – Ciências de dados
{E01131, E01381, E01499}@academico.domhelder.edu.br

Resumo: O trabalho consiste em uma fase de etapas, desde a colheita e análise da base, passando pela análise exploratória até chegar a execução do algoritmo Florest Random. Com o objetivo de predizer a depressão, tendo como base de dados, os dados retirados do IBGE, onde uma gigantesca quantidade de informação está alocada e categorizada, fizemos uma análise superficial e crescendo e incrementando nossa análise pouco a pouco, conhecendo mais sobre o tema, eliminando dados ausentes, outliers, nulos, informações irrelevantes, reduzido a quantidade de variáveis, verificando a possibilidade de enviesar o código.

Palavras-Chave: Ciência de dados, analise exploratória, machine learn, regressão linear, logística, depressão, Random Florest, inteligência artificial, mineração de dados

1 Introdução

Para entender toda a complexidade de uma predição, primeiro precisamos saber o que é ciência de dados, esse conceito pode ser entendido como a união de diversas disciplinas, sendo elas teóricas, matemáticas ou prática, sendo algumas delas Estatística, Inteligência Artificial, Mineração de dados entre outras, a união de toda essas características formam um profissional de ciência de dados (COMARELA et al., 2019).

A ciência de dados nos da uma base muito grande de análise, e inferência no mundo real, possibilitando a automatização, predição, suporte e tomada de decisão de milhares de pessoas diariamente, projetos e algoritmos reduzem o tempo gasto, tornam a vida das pessoas mais dinâmicas e lúdicas, as aplicações são ilimitadas. Por esse motivo, e não poderia ser diferente a ciência de dados, mas especificamente Machine Learn e mineração de dados são a base pratica e teórica respectivamente para nosso trabalho de predizer a depressão.

O algoritmo de Random Florest e Support Vector Machine caem perfeitamente nesse aspecto por possibilitar caminhos de decisões múltiplas, cruzando diversas variáveis (informações colhidas e tratadas exaustivamente na etapa de análise exploratória) entre si, criando um modelo preditivo, normalmente mais preciso, o que e exatamente buscado nessa situação.

1.1 Problema

Segundo a OMS a depressão é a segunda maior causa de morte entre jovens de 15 a 29 anos, ficando conhecida como o “Mal do Século”. Por isso a sua prevenção é de

extrema importância, marcando o mês de setembro com campanhas e mobilizações com o objetivo de conscientizar a população (EDUCACAO, 2019).

Segundo a Amazon (2003), Machine Learning (ML) pode ser usado na área da saúde, auxiliando médicos no diagnóstico e no tratamento, já que as analises de um algoritmo são bem mais velozes, o que auxilia a identificar um quadro de depressão em seus estágios iniciais.

1.2 Justificativa

Foi uma grande reviravolta para escolhemos o tema depressão, antes dele passamos por autismo e TDHA e por falta de atributos compatíveis, resolvemos mudar o assusto, chegamos a fazer os mapas mentais e analisar alguns atributos, mas nos deparamos com esse empecilho.

Agora que fico explicado essas mudanças, fica claro que o tema nos escolheu e não nós escolhemos o tema. Apesar disso, o assunto chama atenção pelo sua complexidade e tamanho de presença na sociedade brasileira, e com isso compreendemos que além de muito interessante e de grande impacto social e privado abordamos esse tema tão sensível com tamanha delicadeza e sensibilidade que merece.

1.3 Objetivos

1.3.1 Objetivo Geral

Desenvolver um modelo preditivo a fim de realizar o diagnóstico da doença depressão, por meio da análise de tal transtorno e de seus aspectos correlacionados, baseado na investigação dos dados extraídos pela Pesquisa Nacional de Saúde de (PNS) 2019 realizada pelo IBGE (Instituto Brasileiro de Geografia e Estatística).

1.3.2 Objetivos Específicos

- Desenvolver um modelo preditivo por intermédio de algoritmos de “Machine Learning” classificatórios objetivando predizer a depressão, a fim de auxiliar os profissionais de saúde a diagnosticar pacientes com sintomas depressivos;
- Identificar os principais atributos presentes na base de dados de estudo que possuem alta correlação com a doença depressão;
- Investigar possíveis hábitos, sintomas e características comuns entre brasileiros que foram diagnosticados com depressão presentes na base de dados PNS 2019;
- Avaliar algoritmos de “Machine Learning” a serem utilizados na elaboração do modelo de predição da doença depressão.

1.4 Organização

O artigo em questão encontra-se organizado em cinco capítulos, sendo eles: Introdução, Referencial Teórico, Fluxo Metodológico, Materiais e Métodos e, por fim, Conclusão, os quais abordarão as etapas e processos realizados a fim de atingir o

modelo preditivo. Dessa forma, no primeiro capítulo é apresentado a contextualização da temática a ser trabalhada, assim como a problemática e motivo da escolha do tema depressão, tal qual os objetivos a serem alcançados.

No segundo capítulo, Referencial Teórico, serão abordados estudos relacionados à temática da doença depressão, algoritmos de “Machine Learning” e trabalhos que uniram ambos. Na sequência, no terceiro capítulo, será explanado o fluxo metodológico desenvolvido e seguido no processo de pesquisa e desenvolvimento do modelo preditivo da depressão.

Sendo assim, no capítulo Materiais e Métodos, serão expostos os materiais e métodos utilizados e que serviram de base para o desenvolvimento do modelo de “Machine Learning”. Da mesma forma, serão discorridos todos os processos realizados ao decorrer do trabalho de pesquisa, os procedimentos e informações que foram levantadas, tal qual a base de dados utilizada e as etapas realizadas desde a limpeza até a discussão dos resultados obtidos. Por fim, retratar-se-à a conclusão atingida pela equipe de pesquisa.

2 Referencial Teórico

2.1 Machine Learning (ML)

De acordo com Escovedo e Koshiyama (2020) Machine Learning (ML) é um subconjunto de técnicas utilizadas na área de Inteligencia artificial que se concentra na descoberta de padrões que relacionem os dados. Como o modelo de ML não conhece essa relação previamente deve-se treina-lo, fornecendo a ele as seguintes combinações de entrada/saída e em seguida o algoritmos calcula a relação entre as entradas e saídas fornecidas (AMAZON, 2003).

MLs podem ser divididos em diversas partes, mas as mais comuns são, supervisionados e não supervisionados. Os Machine Learning supervisionados são algoritmos que trabalham com dados de treinamento rotulados, utilizados para predizer um conjunto limitado de dados ou dividi-los em categorias. Os Machine Learning não supervisionados, são algoritmos treinados com dados não rotulados e tentam estabelecer relações com os dados, podendo categoriza-los e identificar padrões (AMAZON, 2003).

2.2 Depressão

Segundo Rufino et al. (2018), a depressão é caracterizada como um transtorno de humor que pode afetar as pessoas em qualquer fase da vida, apesar de ser mais comum nas idade médias.

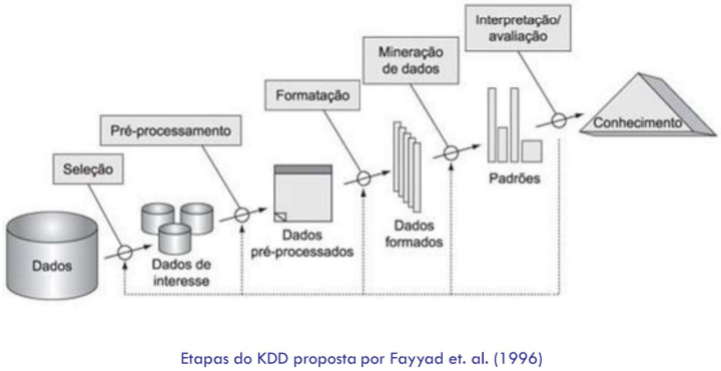
De acordo com Varella (2020), as causas da depressão são acontecimentos traumáticos na infância, estresse físico e psicológico, consumo de drogas lícitas e ilícitas e certos tipos de medicamentos. Varella (2020) também cita que pode existir predisposição genética em algumas pessoas.

O diagnóstico da depressão é feito a partir da percepção de determinados sintomas que se manifestam numa certa duração e intensidade que são analisados de acordo com o histórico de vida do paciente (GRUBITS; GUIMARÃES, 2007).

3 Fluxo Metodológico

Sabemos que uma análise/projeto de ciência de dados possuem grandes etapas, dividido em diversos tópicos e sub-tópicos muitas vezes complexos e com uma grande extensão de conteúdo e etapas a serem validadas.

Figura 1: Etapas do KDD

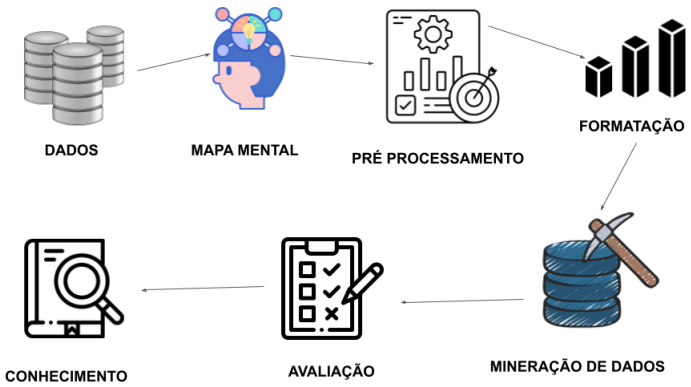


Fonte: Retirado do Slide CD 02 - Marcos Wander

A imagem acima demostra uma trajetória que os dados faz até pode ser dito como conhecimento, ou seja, algo que possa ser usado ou demostrado como verdadeiro ou falso sem grandes preocupação com a qualidade da informação, devido já ter sido feito uma análise encima disso nos dados. Nossa pesquisa passou de modo geral pelos seguintes caminho.

Dados (IBGE) → Seleção (Tema) → Dados de Interesses (Mapa mental, escolha das variáveis) → Pré-processamento (limpeza, remoção, tratamento e preenchimento dos dados) → Formatação (padronização dos dados) → Mineração de dados (aplicação do algoritmo) → Interpretação/ avaliação (análise dos resultados finais).

Figura 2: Fluxo seguido no desenvolvimento do trabalho



Fonte: Produzido pelos autores

4 Materiais e Métodos

Segundo Gil (2008, p. 26), a pesquisa é um “processo formal e sistemático de desenvolvimento do método científico”, e complementa indicando que o objetivo é apresentar soluções e respostas para problemas por intermédio de procedimentos científicos. Dessa maneira, neste estudo, a problemática trabalhada é o diagnóstico de uma pessoa que possa estar com depressão, para tanto, foi-se empregado procedimentos de exploração do espaço problema, análise, limpeza e pré-processamento dos dados, assim como técnicas de modelagem e algoritmos para o desenvolvimento de uma predição.

Nesse sentido, essa seção possui como objetivo descrever a metodologia, métodos e materiais adotados pela equipe de pesquisa para a efetivação do estudo em questão.

Em relação ao aspecto da natureza da pesquisa, trata-se de uma pesquisa aplicada, visto que possui o intuito de gerar conhecimento direcionado para uma aplicação prática (MORESI, 2003), voltado para auxiliar o âmbito da saúde a fim de diagnosticar pacientes depressivos.

Trata-se de um estudo quantitativo, em que obteve-se valores numéricos nos resultados, visto que ao realizar a construção de dois modelos preditivos, cada qual utilizando um algoritmo classificatório, e, executar o treinamento desses, gerou-se dados estatísticos referentes às métricas relacionadas à eficiência dos modelos desenvolvidos. Segundo Moresi (2003), pesquisas quantitativas realizam a tradução de informações em números, fazendo uso de técnicas estatísticas.

A análise dos dados quantitativos obtidos foi executada por meio da análise de informações referentes à depressão e técnicas estatísticas. Nesse sentido, a discussão dos resultados alcançados foi executada à luz da revisão bibliográfica, juntamente com a investigação das métricas encontradas em relação ao desempenho dos modelos criados

Além disso, foram empregadas ao decorrer do estudo metodologias utilizadas na área de Descoberta de Conhecimento em Base de Dados (“Knowledge Discovery in Databases” - KDD), que tem como objetivo principal estabelecer uma ordenação de regras e tarefas a serem seguidas para se obter resultados satisfatórios (BRACHMAN; ANAND, 1996 apud BOENTE et al., 2008).

4.1 Materiais - Base de Dados

O desenvolvimento do trabalho em questão aderiu, como material principal para a pesquisa, a base de dados da Pesquisa Nacional de Saúde (PNS) realizada no ano de 2019. Segundo a própria organização (PNS, 2021), tal pesquisa trata-se de um levantamento epidemiológico de saúde, de base domiciliar e de âmbito nacional, executado pelo Ministério da Saúde juntamente com o IBGE.

Tal base de dados tem sido utilizada para obtenção de informações referentes à morbidade e estilos de vida saudáveis. Devido a sua realização periódica, torna-se possível consolidar informações colhidas tomando como referência populacional em diferentes aspectos, tendo destaque para doenças crônicas e seus aspectos determinantes (PNS, 2021).

Nesse sentido, os dados coletados representam uma amostra da população brasileira, os quais foram tratados pela equipe de pesquisa, tendo como enfoque a doença depressão, para que se tornassem entradas para o treinamento dos modelos predi-

tivos criados. Sendo assim, foram-se correlacionando as informações adquiridas via pesquisas bibliográficas com os módulos existentes na base de estudo.

Diante do fato da pesquisa coletar dados referentes à saúde dos brasileiros e, principalmente, conter um módulo denominado doenças crônicas, o módulo Q, o qual contém informações específicas à respeito da depressão, foi-se de extrema relevância para o desenvolvimento do trabalho.Assim como fonte de alimentação dos modelos preditivos, visto que tal doença não possui fronteiras, dessa forma, proporciona um aumento da capacidade dos modelos de predizer a probabilidade de qualquer brasileiro estar ou não acometido pela doença depressão.

De forma semelhante, foram utilizados referenciais literários e exemplos da empregabilidade de algoritmos de “machine learning”. Tendo destaque para os algoritmos “Random-Forest” e “Support Vector Machine” (SVM), ambos classificatórios, os quais foram empregados para a criação dos modelos preditivos, em que retornam as classificações: 1 = tem depressão e 2 = não tem depressão.

4.2 Métodos

4.2.1 Seleção de atributos

A base de dados disponibilizada pelo IBGE, em sua forma bruta, “raw base”, constituise de 1116 atributos, a qual contém todas as perguntas do questionário realizado pelo PNS de 2019. Tais atributos, referem-se a distintos aspectos, tais como o desempenho nacional de saúde, as condições de saúde, assim como a vigilância das doenças e agravos de saúde e fatores de risco associados.

Dessa maneira, a base original continha alta quantidade de atributos, sendo que muitos deles não eram relevantes para temática da pesquisa. Sendo assim, foi realizada a primeira seleção de atributos, em que foram analisados manualmente todos os atributos que continham na base, realizando sua correlação com a doença depressão.

Nesse aspecto, foi efetuada a primeira redução por intermédio da associação das informações adquiridas pela pesquisa do referencial teórico juntamente com os atributos constantes na base de dados, e assim, foram reduzidos para 108 atributos. Tendo destaque para o atributo Q092, o qual foi utilizado como rótulo pelos modelos desenvolvidos, em que informa se a pessoa entrevistada já foi diagnosticada com depressão.

Posteriormente, foi realizada uma segunda etapa de seleção de atributos, pautada pela verificação da presença de dados ausentes dentre os 108 atributos. Dessa forma, foi criado um método para percorrer todos os atributos presentes na última base montada e verificar se havia dados ausentes em cada atributo, assim como a porcentagem desses dados.

Diante desta análise, foram eliminados todos os atributos com mais de 70% dos dados ausentes, visto que pela falta de dados, poderia enviesar nossos modelos, tal como realizar imputações poderiam não ser tão representativas quanto fosse necessário. Outros atributos, após identificação de quantidade pequena de dados ausentes, foram analisados e removidos os que realmente não eram relevantes em comparação aos demais.

Assim como foram avaliados os atributos que não existiam nenhum dado ausente, tal como sua relevância para o estudo em questão. Dessa forma, foram selecionados 26 atributos, ambos com nenhum dado ausente, visto que os demais atributos foram eliminados, seja pelo motivo que continham muitos dados ausentes, ou que requeririam

imputação de dados, o que poderia ocasionar um modelo tendencioso.

Por fim, foi realizada uma última redução de atributos por intermédio da análise das correlações, e foram notados 3 atributos com baixa correlação com o atributo rótulo, os quais eram mais relacionados após a pessoa ter ciência que estava com depressão. Dessa forma, identificou-se que esses atributos não eram tão relevantes para a predição da depressão, mas possivelmente para tratamento pós diagnóstico, para tanto, esses foram eliminados.

Portanto, a base final contém 23 atributos, os quais estão 100% preenchidos, alguns com presença de outliers, no entanto que são relevantes de permanecerem, visto que parte da pesquisa pauta-se na análise desses outliers. Assim como a base final foi construída com uma quantidade de atributos consideráveis para que os modelos preditivos fossem criados e obtivessem um bom desempenho, não sendo enviesados.

Segue tabela com os códigos dos atributos da base final, tal como o que significam:

V0026	Tipo de situação censitária
C006	Sexo
C009	Cor ou raça
J00101	Considerando saúde como um estado de bem-estar físico e mental, e não somente a ausência de doenças, como é o estado de saúde
J002	Nas duas últimas semanas, ___ deixou de realizar quaisquer de suas atividades habituais (trabalhar, ir à escola, brincar, afazeres domésticos etc.) por motivo da própria saúde
J007	Algum médico já deu o diagnóstico de alguma doença crônica, física ou mental, ou doença de longa duração (de mais de 6 meses de duração)

M01401	Com quantos familiares ou parentes ___ pode contar em momentos bons ou ruins
M01501	Com quantos amigos próximos ___ pode contar em momentos bons ou ruins (Sem considerar os familiares ou parentes)
M01601	Nos últimos doze meses, com que frequência o(a) Sr(a) se reuniu com outras pessoas para prática de atividades esportivas, exercícios físicos, recreativos ou artísticos
M01701	Nos últimos doze meses, com que frequência o(a) Sr(a) participou de reuniões de grupos como associações de moradores ou funcionários, movimentos sociais/comunitários, centros acadêmicos ou similares
N010	Nas duas últimas semanas, com que frequência o(a) Sr(a) teve problemas no sono, como dificuldade para adormecer, acordar frequentemente à noite ou dormir mais do que de costume?
N011	Nas duas últimas semanas, com que frequência o(a) Sr(a) teve problemas por não se sentir descansado(a) e disposto(a) durante o dia, sentindo-se cansado(a), sem ter energia?
N012	Nas duas últimas semanas, com que frequência o(a) Sr(a) teve pouco interesse ou não sentiu prazer em fazer as coisas?
N014	Nas duas últimas semanas, com que frequência o(a) Sr(a) teve problemas na alimentação, como ter falta de apetite ou comer muito mais do que de costume?
N016	Nas duas últimas semanas, com que frequência o(a) Sr(a) se sentiu deprimido(a), “pra baixo” ou sem perspectiva?
N017	Nas duas últimas semanas, com que frequência o(a) Sr(a) se sentiu mal consigo mesmo, se achando um fracasso ou achando que decepcionou sua família?
N018	Nas duas últimas semanas, com que frequência o(a) Sr(a) pensou em se ferir de alguma maneira ou achou que seria melhor estar morto?
P027	Com que frequência o(a) Sr(a) costuma consumir alguma bebida alcoólica?
P034	Nos últimos três meses, o(a) Sr(a) praticou algum tipo de exercício físico ou esporte?
P04502	Em um dia, quantas horas do seu tempo livre (excluindo o trabalho), o(a) Sr(a) costuma usar computador, tablet ou celular para lazer, tais como: utilizar redes sociais, para ver notícias, vídeos, jogar etc?

4.2.2 Pré-processamento

Na etapa de pré-processamento foi realizada a importação da base de dados já com os atributos selecionados, como descrito no tópico 4.2.1.

Nesta etapa também realizamos a análise exploratória dos dados, que segundo Escovedo e Koshiyama (2020) é muito importante compreender bem os dados com o objetivo de possivelmente obter resultados melhores nos algoritmos de Machine Learning.

Ao criar vários gráficos foi possível visualizar a disposição dos dados, identificando valores faltantes, valores discrepantes (outliers) e se seria necessário realizar a transformação dos dados. Também podemos analisar a relação dos atributos e identificar quais eram redundantes ou que não agregam muita informação. Após o entendimento da base foi possível remover os atributos e dados da qual não seriam necessários.

4.2.3 Modelo de treino-teste

Nessa etapa, foi construído alguns testes e métodos para tentar se aproximar de uma melhor eficácia do algoritmo escolhido. Aqui optamos por testar a base com 30% para treino e 70% teste, também tentamos com 40% e 60% sendo eles de teste e treino respectivamente, e com uma metodologia diferente validamos a opção de separar a base em pedaços iguais e testar eles entre elas, chamado de validação cruzada ou K-fold.

Teste de dados passando 40% da base

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.40,
    random_state=101)
# define os nomes da variaveis de teste e treino, a quantidade de dados a
    ser usado e a semente para gerar
```

Teste de dados passando 30% da base

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30,
    random_state=101)
# define os nomes da variaveis de teste e treino, a quantidade de dados a
    ser usado e a semente para gerar
```

Foi testado os K-fols's porém por alteração significativa no código foi descartado na versão final do projeto.

4.2.4 Modelo de aprendizado de maquina

4.2.4.1 Random Forest

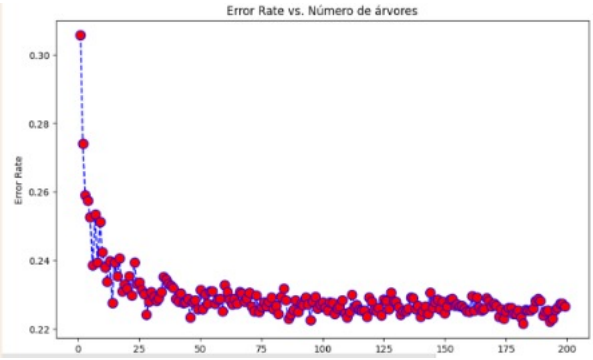
O modelo Random forest, é constituído por diversas árvores de decisão que são geradas de forma aleatórias, inicialmente é realizada a seleção de maneira aleatória de algumas amostras dos dados de treino. Assim, usa o método reamostragem, em que as amostras podem ser repetidas em uma mesma seleção e assim, essa primeira seleção forma a primeira árvore de decisão.

Esse processo de escolha se repete em todos os nós e a quantidade de variáveis pode ser definida no algoritmo. Como a seleção ocorre de forma aleatória e serão construídas diversas árvores, torna-se uma forma eficiente e geralmente evita o overfitting. Quanto mais árvores forem criadas, melhor serão os resultados do modelo, até

chegar em uma quantidade em que mesmo aumentando a quantidade de árvores não melhoram os resultados.

Nesse sentido, cada árvore apresentará seu resultado e, o resultado que mais vezes for apresentado pelas diversas árvores será escolhido.

Figura 3: Taxa de erros vs Número de arvores



Fonte: Produzido pelos autores

4.2.4.2 Support Vector Machine

O SVM (Support Vector Machine) é um algoritmo de aprendizado de máquina usado principalmente para classificação e regressão. Ele é um método supervisionado, o que significa que requer dados rotulados para treinamento.

O funcionamento básico do SVM envolve a criação de um hiperplano que separa os dados em diferentes classes. Um hiperplano é uma superfície de decisão que divide o espaço de características em duas ou mais partes, para cada classe existente. O objetivo do SVM é encontrar o hiperplano ótimo que maximize a margem entre as classes, ou seja, a maior distância possível entre os pontos de dados mais próximos de classes diferentes. O Hiperplano traçado pode ser um linha ou com curvas a fim de separar os dados com uma melhor precisão.

Para que o algoritmo funcione corretamente é necessário que a padronização e normalização dos dados seja realizada, colocando eles em uma mesma escala. Além disso, dois principais parâmetros influenciam o resultado final do algoritmo, sendo eles o Gamma e o C.

O parâmetro gamma, controla a distância que as variáveis serão consideradas para a definição do limite entre as fronteiras, quanto maior o valor do gamma mas variáveis serão desconsideradas na bordas, causando um overfitting , esse parâmetro e como seu fosse um definição de pesos para as variáveis, no qual através deles e demonstradas a importância de cada variável no treinamento do algoritmo.

O parâmetro C, ao contrário de seu colega, está presente tanto no modelo linear como não linear do modelo SVM, ele é responsável por cuidar da tolerância de erros de classificação, ou seja, é como se fosse uma penalidade aplicada a cada amostra tida como errada. Valores de C muito altos tendem a buscar a separação completa, mas normalmente tende a fazer o algoritmo demorar a executar e gerar overfitting, já quando o valor de C é baixo, a etapa do treinamento é mais simples, permitindo geração de fronteiras com erros, mas podem causar underfitting.

4.2.5 Avaliação do modelo

Conhecendo a base de dados e com os modelos treinados, foi possível construir alguns dados sintéticos para testar os modelos e realizar comparações com os testes anteriores. Porém para avaliar de fato os modelos geramos um relatório de classificação utilizando o método `classification_report()` que nos entrega a *precisão*, a *acurácia*, o *recall*, e o *f1-score*.

As analises foram feitas em peso sobre a *acurácia*, que indica a performance geral do modelo, o quanto que ele acertou em relação às previsões realizadas. Também foi analisado o resultado do *f1-score*, que realiza a média entre a *precisão* e o *recall*, pois o resultado da *acurácia* dos dois modelos testados foi muito próxima.

	precision	recall	f1-score	support
1.0	0.74	0.82	0.78	2487
2.0	0.80	0.72	0.76	2513
accuracy			0.77	5000
macro avg	0.77	0.77	0.77	5000
weighted avg	0.77	0.77	0.77	5000

Figura 4: Relatório de classificação do algoritmo Random Forest

	precision	recall	f1-score	support
1.0	0.75	0.85	0.79	2487
2.0	0.82	0.72	0.77	2513
accuracy			0.78	5000
macro avg	0.79	0.78	0.78	5000
weighted avg	0.79	0.78	0.78	5000

Figura 5: Relatório de classificação do algoritmo Support Vector Machine

4.2.6 Discussão dos resultados

Ambos os modelos desenvolvidos apresentaram bons índices referentes às métricas estatísticas de análise dos resultados apresentados pelos modelos. Apesar de serem modelos desenvolvidos com algoritmos distintos, sendo o Random Forest, um algoritmo ensemble e o Support Vector Machine, classificatório, apresentaram valores bem semelhantes referentes à capacidade preditiva.

Ao que se refere ao modelo desenvolvido utilizando Random Forest, dentre os 16.663 registros, nesse cenário de análise seriam os pacientes, foi gerada uma matriz de confusão com os seguintes indicativos de predição: 2048 correspondente a true negative, ou seja, pessoas que não possuíam depressão e a predição foi realizada corretamente. Tal como, o valor 1816 representa true positive, demonstra numericamente a quantidade prevista corretamente da classe em que estávamos buscando, sendo assim, as pessoas diagnosticadas com depressão em que foram previstas corretamente.

Em contrapartida, o valor 439 foi a quantidade de pessoas que foram preditas com depressão, no entanto, no conjunto real, não possuem depressão, são os falsos positivos. Em relação ao valor de 697, refere-se aos diagnósticos falsos negativos, ou seja, as pessoas que realmente têm depressão, porém o modelo predisse um diagnóstico negativo.

Nesse aspecto, realizando a análise do relatório de classificação, o qual apresenta as métricas estatísticas, tem-se que a precisão atingida foi de 75% para pessoas com depressão e 81% para pessoas sem depressão. Sendo que a precisão representa a proporção de identificações positivas (tem depressão) que estavam corretas dentre todas as classificações de positivo que o modelo fez, ou seja, a proporção de dados que foram classificados com o respectivo rótulo real dentre o total que foi predito como positivo.

Apresentou um recall, proporção do que se quer predizer acertado corretamente, dentre todos os valores que de fato eram correspondentes no conjunto real, sendo 82% para a predição de pessoas com depressão e 72% para pessoas sem depressão. Enquanto que referente ao f1-score, relativo a quem tem depressão apresentou 78% e em relação a quem não foi diagnosticado, 76% , sendo esse valor a média entre a precision e recall.

De maneira semelhante, o modelo desenvolvido com SVM, apresentou a matriz de confusão com os seguintes indicadores: 2103 referente ao true negative, pessoas que foram preditas sem depressão e que realmente não tinham depressão; 1809 como true positive, ou seja pessoas preditas como depressivas e que foram diagnosticadas com tal doença. Tal como 384 que foram preditos como falsos positivos e 704 como falsos negativos.

No que se refere às métricas, apresentou precisão de 75% para diagnosticados e 82% para não diagnosticados; 85% de recall para diagnosticados e 72% para não diagnosticados. Assim como o f1-score de 79% para pessoas com depressão e 77% para pessoas que não têm depressão.

Dessa maneira, nota-se que ambos modelos apresentaram valores bem próximos nas diferentes métricas avaliadas. Tal como o primeiro modelo apresentou acurácia de 77%, enquanto que o segundo modelo apresentou 78%.

Portanto, é possível perceber que os diferentes modelos obtiveram performance geral positiva, sendo que ambos ultrapassaram 60% de acurácia. Outrossim, os resultados foram satisfatórios dentro os recursos e tempo disponibilizados para o seu desenvolvimento, visto que atenderam ao objetivo principal do trabalho, predizer pessoas com depressão a fim de auxiliar os profissionais de saúde nos diagnósticos desses.

5 Conclusão

A aplicação dos algoritmos Random Forest e Support Vector Machine (SVM) para a previsão da depressão, utilizando a base de dados da Pesquisa Nacional de Saúde (PNS) de 2019, fornecida pelo Instituto Brasileiro de Geografia e Estatística (IBGE), demostrou resultados satisfatórios dentro dos parâmetros passados e da dificuldade apresentada na efetivação da construção desse projeto.

O uso de algoritmos de aprendizado de máquina tem crescido em grande escala e não foi diferente no setor da saúde, mas como deve ser lembrado nosso trabalho não deve ser usado como base verídica para afirmar ser alguém tem depressão ou

não. Mas de uma forma complementar ela interage com as avaliações clínicas dando uma segunda visão de um diagnóstico pré-estabelecido, e na adoção de medidas de cuidados dos pacientes.

Em suma, este estudo demonstrou que os algoritmos de Random Forest e Support Vector Machine são ferramentas poderosas para prever a depressão com base nos dados da PNS 2019. Essas técnicas oferecem uma abordagem inovadora e complementar para a compreensão e prevenção da depressão, confiante para o avanço da saúde mental e bem-estar da população brasileira.

Sem sobras de dúvidas a capacidade de gerar resultados através de variáveis muitas vezes desconexa para a maioria das pessoas e demostrar uma situação de causa e efeito entre nossos hábitos e não saúde física e mental tem seu lugar garantido no hall de criações tecnológicas incríveis do ultimo século, com isso conseguimos alcançar não somente a predição da causa, mas demostrar e entender que cada habito, relação social, trauma e vivências moldam agente, e entender isso, trouxe um panorama amplo para o grupo, que mas uma vez desde o ingresso na faculdade pôde ter a sua visão de mudo e da sociedade reformulada e acrescida mesmo que minimamente para melhor.

5.1 Pontos Fortes e fracos

Como pontos fracos podemos definir:

- Acurácia longe da desejada (devido ser tratar de uma doença)
- Dificuldade de correlação das variáveis
- Falta de uso de outros tipos de algoritmos (Redes neurais, regressores e etc)
- Co-dependência entre variáveis (variáveis que depende de um outra para existir ou ter o dado)

Os pontos fortes são:

- Variável especifica
- Uso de 2 Classificadores
- Poucos Outliners e ruídos
- Sem necessidade de imputação

5.2 Trabalhos Futuros

Aqui se encontra as etapas que poderiam ter sido testadas e implementadas no nosso projeto, mas por falta de tempo, verba, ou até mesmo capacidade de implementação não foram possíveis inserir no trabalho como: Análise de variáveis possíveis para atingir diferentes resultados, usos de diferentes tipos de algoritmos para realizar o treinamento da base de dados, exploração de diferentes métodos de seleção de dados para treino e teste, definição de um escopo de pessoas que terá os dados tratados, por exemplo faixa etária, classe social etc, aplicação em algoritmos mais robustas como redes neurais.

Referências

AMAZON, W. S. **O que é Machine Learning?** Amazon Web Services, 2003. Disponível em: <https://aws.amazon.com/pt/what-is/machine-learning/>. Acesso em: 07 jun. 2023.

BOENTE, A. N. P. et al. Uma metodologia de suporte ao processo de descoberta de conhecimento em bases de dados. In: **Anais do V Simposio de Excelencia em Gestão e Tecnologia-SEGeT**. [S.l.: s.n.], 2008. v. 1, p. 4–5.

BRACHMAN, R. J.; ANAND, T. The process of knowledge discovery in databases. In: **The KDD Process for Extracting Useful Knowledge from Volumes of Data**. [S.l.: s.n.], 1996. p. 37–57.

COMARELA, G. et al. Introdução à ciência de dados: Uma visão pragm ática utilizando python, aplicações e oportunidades em redes de computadores. **Sociedade Brasileira de Computação**, 2019.

EDUCACAO, M. d. **Depressão É uma das principais causas de Suicídio, Aponta Oms**. Ministério da Educação, 2019. Disponível em: <https://www.gov.br/ebserh/pt-br/hospitais-universitarios/regiao-nordeste/hul-ufs/comunicacao/noticias/depressao-e-uma-das-principais-causas-de-suicidio-aponta-oms>.

ESCOVEDO, T.; KOSHIYAMA, A. **Introdução a Data Science - Algoritmos de Machine Learning e métodos de análise**. São Paulo: Casa do Codigo, 2020.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Ediitora Atlas SA, 2008. Disponível em: <https://ayanrafael.files.wordpress.com/2011/08/gil-a-c-mc3a9todos-e-tc3a9cnicas-de-pesquisa-social.pdf>. Acesso em: 26 abril. 2023.

GRUBITS, S.; GUIMARÃES, M. A. L. Psicologia da saúde. especificidades e diálogo interdisciplinar. Artmed, Porto Alegre, p. 145–146, 2007.

MORESI, E. Metodologia da pesquisa. **Brasília: Universidade Católica de Brasília**, v. 108, n. 24, p. 5, 2003. Disponível em: <http://www.inf.ufes.br/~pdcosta/ensino/2010-2-metodologia-de-pesquisa/MetodologiaPesquisa-Moresi2003.pdf>. Acesso em: 08 jun. 2023.

PNS, P. N. d. S. **O que é PNS?** IBGE, 2021. Disponível em: <https://www.pns.icict.fiocruz.br/>. Acesso em: 08 jun. 2023.

RUFINO, S. et al. Aspectos gerais, sintomas e diagnóstico da depressão. **Revista Saúde em Foco**, v. 10, n. 1, p. 837–843, 2018.

VARELLA, D. D. **Depressão**. UOL, 2020. Disponível em: <https://drauziovarella.uol.com.br/doencas-e-sintomas/depressao/>.

A Apêndice

Figura 6: Mapa Mental sobre depressão