Navegando o cenário da Saúde Mental na Indústria Tech

Uma análise baseada em aprendizado de máquina

Augusto Scardua Oliveira
Ciência da Computação
PUC Minas
Belo Horizonte, MG, Brasil
augusto.oliveira.1388188@sga.pucminas.br

Bruno Santiago de Oliveira
Ciência da Computação
PUC Minas
Belo Horizonte, MG, Brasil
bruno.santi.oli@email.com

Carolina Morais Nigri
Ciência da Computação
PUC Minas
Belo Horizonte, MG, Brasil
carolina.nigri@sga.pucminas.br

Fábio Freire Kochem
Ciência da Computação
PUC Minas
Belo Horizonte, MG, Brasil
fabio.freire@sga.pucminas.br

Pedro Heinrich Sales Pena
Ciência da Computação
PUC Minas
Belo Horizonte, MG, Brasil
pedroheinri@gmail.com

Pedro Miranda Rodrigues
Ciência da Computação
PUC Minas
Belo Horizonte, MG, Brasil
pedro.rodrigues.1373336@sga.pucminas.br

RESUMO

Este artigo traz uma análise dos resultados de uma pesquisa realizada pela *Open Sourcing Mental Health* (OSMI), em 2014, sobre a saúde mental dos trabalhadores da indústria de tecnologia. Utilizando técnicas de aprendizado de máquina, este estudo busca identificar padrões e tendências nos dados coletados pela pesquisa descrita. Após realizar um pré-processamento da base de dados, removendo e alterando atributos conforme necessário, foi aplicado o algoritmo de *Random Forest* para prever se os funcionários procurariam ou não tratamento. Os resultados do modelo criado e sua respectiva análise são apresentados neste documento. Com isso, busca-se, como objetivo principal, contribuir para a promoção de um ambiente de trabalho mais saudável na indústria tech, visando auxiliar na formulação de estratégias que tragam melhorias para a saúde mental desses trabalhadores.

KEYWORDS

Saúde Mental no Trabalho, Indústria da Tecnologia, Aprendizado de Máquina, Classificador Random Forest, Modelagem Preditiva, Cultura no Ambiente de Trabalho Tech, Análise de Dados, Pesquisa com empregados, Suporte de Decisão, Algoritmo de Classificação, Iniciativas de Apoio no Local de Trabalho.

ACM Reference format:

Augusto S. Oliveira, Bruno S. de Oliveira, Carolina M. Nigri, Fábio F. Kochem, Pedro H. Sales Pena, Pedro M. Rodrigues. 2023. Navegando o cenário da Saúde Mental na Indústria Tech: Uma análise baseada em aprendizado de máquina. PUC Minas, Belo Horizonte, Brasil.

1. Introdução

A *Open Sourcing Mental Health* (OSMI) é uma organização sem fins lucrativos dedicada a aumentar a conscientização, educar e fornecer recursos para apoiar o bem-estar mental nas comunidades de tecnologia e código aberto.[1] Em busca dessa missão, a OSMI conduziu uma pesquisa abrangente em 2014, com o objetivo de avaliar as atitudes em relação à saúde mental e a prevalência de transtornos mentais no ambiente de trabalho de tecnologia.[2]

De maneira notável, em toda a Europa, tem havido um preocupante aumento nas faltas relacionadas à doença e aposentadorias precoces atribuídas a problemas de saúde mental.[3] Essa tendência preocupante não afeta apenas o bem-estar das pessoas, mas também representa um crescente fardo para a economia devido aos custos crescentes associados à saúde e à produtividade perdida.

Neste artigo, aprofundaremos as informações obtidas na pesquisa da OSMI em 2014 e aproveitaremos o poder da aprendizagem de máquina para analisar esses dados. Usando o modelo do *Random Forest*, nosso objetivo é descobrir padrões e tendências valiosas na interseção entre saúde mental e a indústria de tecnologia. Em última análise, nosso objetivo é contribuir para uma abordagem mais informada, baseada em dados e voltada para o futuro

para promover uma força de trabalho de tecnologia mais saudável e resiliente.

2. Descrição da base de dados

O conjunto de dados usado neste estudo abrange uma ampla gama de informações, incluindo detalhes demográficos, como idade, gênero, país e estado (para residentes dos EUA). Também se aprofunda em aspectos relacionados ao emprego, com perguntas relacionadas ao status de autoemprego, histórico familiar de doenças mentais e se os indivíduos buscaram tratamento para condições de saúde mental.

Além disso, o conjunto de dados explora o impacto da saúde mental no trabalho, observando se os indivíduos sentem que sua condição interfere em seu emprego. Fatores relacionados ao empregador, como a oferta de benefícios de saúde mental, opções de cuidados, programas de bem-estar e recursos para buscar ajuda são examinados. São feitas perguntas sobre anonimato, licença médica e percepções sobre discutir questões de saúde mental com empregadores e colegas lançam luz sobre a atitude do ambiente de trabalho em relação à saúde mental.

Por fim, o conjunto de dados inclui um campo aberto para comentários adicionais, dando aos participantes a oportunidade de compartilhar insights e experiências pessoais relacionadas à saúde mental.[2] Totalizando 26 atributos, tendo como maior parte atributos categóricos com respostas limitadas, a exceção de alguns com alta variedade de respostas, visto que as perguntas eram abertas. Idade sendo o único atributo numérico presente na base.

A seguir, é apresentado na Figura 1 todos os atributos citados acima de forma visual, organizando-os por tipo de informação fornecida de forma a facilitar o entendimento dos dados.

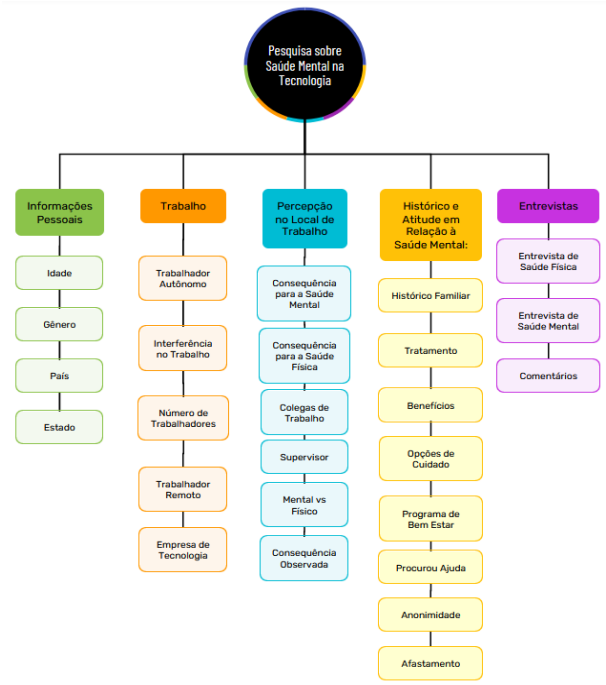


Figura 1 – Descrição visual dos atributos.

3. Etapas de pré-processamento

O pré-processamento da base iniciou-se pela remoção dos campos de *timestamp*, que no contexto de aprendizado de máquina era irrelevante, *comments*, que são campos em sua maioria vazios e sem padrão, e *state*, que era válido somente para os Estados Unidos. Em seguida, foi tratado o atributo idade, que possuía alguns dados incorretos, como idades negativas, e outros que foram considerados incoerentes e podem ser *outliers* que prejudicam a análise, como idades menores que 15, por se tratar de uma pesquisa em um local de trabalho, ou maiores que 100. Assim, filtrou-se as idades entre 15 e 100 anos.

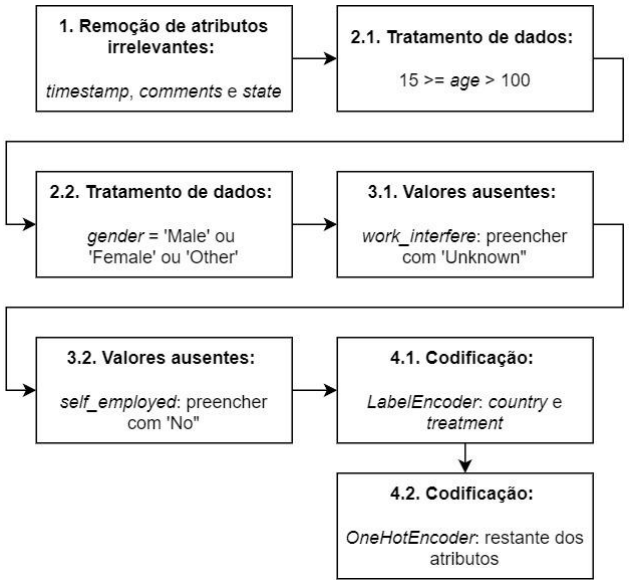
O próximo atributo tratado foi o de gênero, que possuía respostas muito variadas e que, em alguns casos, indicavam o mesmo dado, como respostas “M” equivalentes a *“Male”*. Esses dados foram tratados de forma que houvesse apenas três tipos de resposta: *“Male”*, *“Female”* e *“Other”*. Por fim, foram identificados valores ausentes apenas nos atributos *work_interfere* e *self_employed,* os quais foram preenchidos com *Unknown* e *No* respectivamente. A base escolhida já estava balanceada, logo essa etapa não foi necessária.

A codificação dos atributos categóricos foi feita, em sua maioria, utilizando-se da técnica do

OneHotEncoder, tendo em vista que não havia uma ordenação entre as respostas e os atributos em sua maioria não eram binários. Apenas os atributos *country* e *treatment* foram codificados pelo *LabelEncoder*, o primeiro por possuir muitos tipos de resposta (logo o *OneHot* geraria muitas colunas, dificultando o processo) e o segundo por ser binário, indicando as classes “Sim” e “Não”.

A Figura 2 mostra as etapas descritas, na ordem em que foram realizadas.

Figura 2 – Ordem das etapas de pré-processamento.



Após o processamento da base, separou-se a coluna referente ao *target* da classificação, o atributo *treatment*, do restante das colunas para realizar o treinamento do modelo. Para o conjunto de teste, foram separados 20% da amostragem da pesquisa. A Tabela 1 traz a quantidade de instâncias, de cada classe, divididas para teste e treinamento além do seu total.

Tabela 1 – Divisão dos dados em teste e treinamento.

Classe	Base original	Teste	Treinamento
Sim	632	505	127
Não	619	495	124
Total	1251	1000	251

4. Descrição dos métodos utilizados

No contexto desta análise, utilizamos um método de aprendizado de máquina: o *Random Forest*. Optamos por este como a principal abordagem devido à sua capacidade de aprimorar a precisão e a robustez do modelo. Este algoritmo é uma técnica de *ensemble* que combina múltiplas Árvores de Decisão para fornecer previsões mais confiáveis e reduzir o risco de *overfitting*. Desta forma, o uso do *Random Forest* nos permite explorar relações complexas nos dados, ajudando-nos a identificar qual método se adapta melhor ao nosso conjunto de dados e aos objetivos da análise.[4]

Para selecionar os hiperparâmetros que gerassem modelos de árvores melhores no *Random Forest*, utilizou-se o *GridSearch*, que testa diferentes combinações de hiperparâmetros em busca da melhor. A combinação de hiperparâmetros retornada pelo *GridSearch* e utilizada no modelo treinado foi a seguinte: critério de entropia, profundidade máxima da árvore de 6 níveis e escolha da raiz quadrada para calcular a quantidade de atributos usados.

O código foi feito em um *notebook* do *Google Colab*, que utiliza o Python 3.10.12. Para execução do código, é necessário fazer o upload da base de dados em formato csv na pasta ‘*/content/sample_data/*’.

5. Resultados e discussões

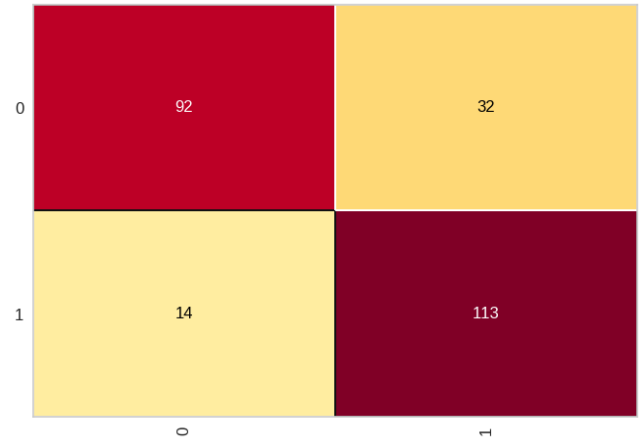
Após o treinamento do modelo do *Random Forest*, as instâncias de teste foram classificadas e obtiveram-se os resultados apresentados a seguir. A Tabela 2 mostra as métricas de avaliação do modelo para a classe “Sim” e a classe “Não”. Em seguida, a Figura 3 apresenta a matriz de confusão gerada pelos testes classificados.

Tabela 2 – Métricas de avaliação do modelo.

Classe	Precisão	Recall	F1-Score
Sim	78%	89%	83%
Não	87%	74%	80%

Acurácia	82%

Figura 3 – Matriz de confusão.



Como observado na Tabela 1, a acurácia do modelo foi de 82%, o que, no caso da base de dados estudada, é um indicativo bom sobre o modelo como um todo, visto que a base está balanceada. Esse fato diminui o risco da acurácia indicar um acerto alto do modelo não pelo fato dele estar acertando bem ambas as classes, mas pelo fato dele acertar mais a previsão de uma classe majoritária do que da minoritária. Portanto, percebe-se que o modelo treinado conseguiu classificar corretamente boa parte das instâncias, tanto da classe “Sim” quanto da classe “Não”, o que fica claro na Figura 3, onde se observa um elevado número de verdadeiros positivos e verdadeiros negativos na matriz de confusão.

Ao realizar uma análise mais particular de cada classe, observa-se que, para a classe “Sim”, o *recall* é mais alto que a precisão, o que indica que há uma taxa mais baixa de falsos negativos. Já para a classe “Não”, ocorreu o contrário, a precisão foi maior que o *recall*, mostrando que há uma taxa menor de falsos positivos. Por fim, os *F1-Score* de ambas as classes, que foram calculados pela média harmônica padrão, sem colocar mais peso em precisão ou *recall*, foi similar, de 80% e 83%.

6. Código desenvolvido

Todo o código desenvolvido está disponível no *COLAB*.

7. Referências Bibliográficas

1. Open Sourcing Mental Health (OSMI). (https://osmihelp.org)
2. Kaggle Dataset: Mental Health in Tech Survey. (https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey)
3. European Union - Compass for Action on Mental Health and Well-being. (https://policycommons.net/artifacts/1943653/eu-compass-for-action-on-mental-health-and-well-being/2695422/)
4. Devetyarov, D., Nouretdinov, I. (2010). Prediction with Confidence Based on a Random Forest Classifier. In: Papadopoulos, H., Andreou, A.S., Bramer, M. (eds) Artificial Intelligence Applications and Innovations. AIAI 2010. IFIP Advances in Information and Communication Technology, vol 339. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-16239-8_8