APLICAÇÃO DE TÉCNICAS DE CLASSIFICAÇÃO EM DADOS DE ACIDENTES DE TRABALHO DO SETOR DE SERVIÇOS

Lucimara Ferreira da Silva (PIBIC/FA/CNPq/Uem), Gislaine Camila Lapasini Leal (Orientadora), Edwin Vladmir Cardoza Galdamez (Co-Orientador), e-mail: ra108396@uem.br

Universidade Estadual de Maringá, Departamento de Engenharia de Produção/Maringá, PR.

Área: Engenharia de Produção
Subárea: Higiene e Segurança do Trabalho

Palavras-chave: Serviços; Doenças ocupacionais; Acidentes ocupacionais.

Resumo:

No Brasil os dados referentes à acidentes de trabalho e doenças ocupacionais são alarmantes. Esses dados são registrados por meio da Comunicação de Acidentes de Trabalho (CAT), com abrangência nacional e acesso público. Empregando estes dados, o objetivo deste estudo é fornecer uma análise dos dados de acidentes, doenças e óbitos ocupacionais no setor de serviços, utilizando o conjunto de dados entre julho de 2018 a março de 2021, tendo sido analisados 14.098 registros. Com a finalidade de fornecer uma análise da aplicação de técnicas de classificação em dados de acidente de trabalho do setor de serviços, esta pesquisa compara a execução de doze técnicas de mineração de dados, mediante os resultados de cinco métricas de desempenho, quanto à eficácia preditiva da ocorrência de óbitos por acidentes de trabalho, tendo como base os registros disponíveis na CAT referentes ao setor de serviços. Os algoritmos *Naïve Bayes* (NB) e *Random Forest* (RF) apresentaram a maior capacidade preditiva, onde em ambos destacou se a métrica curva ROC/AUC.

Introdução

No Brasil, os acidentes de trabalho se constituem em problema de saúde pública, pois além de causar danos aos trabalhadores, acarretam grandes consequências sociais e econômicas para o país (SANTANA *et al*., 2003).

De acordo com estimativas globais da Organização Internacional do Trabalho (OIT), anualmente a economia perde cerca de 4% do Produto Interno Bruto (PIB) em razão de doenças e acidentes de trabalho, ocasionando perdas humanas e perdas de produtividade provocada por ambientes de trabalho inadequados.

O setor de serviços assumiu um papel de destaque na economia, em 2013 a participação do setor de serviços no PIB representou quase 70%. A predominância do setor de serviços se estendeu para além do PIB, sendo primordial no que se

refere à criação de firmas e de empregos no Brasil, o setor respondia a 72,3% do total de empregos em 2012 (ARBACHE, 2006).

A visão das consequências financeiras de lesões e doenças ocupacionais fornece ao governo e organizações dados relevantes que podem contribuir para definir prioridades de intervenção e auxiliar no desenvolvimento de políticas de SST. Além disso, insights sobre essas despesas ajudarão a aumentar a conscientização sobre a magnitude do problema (TOMPA *et al*., 2019).

Diante dessa conjuntura, este estudo tem como objetivo fornecer uma análise da aplicação de técnicas de classificação em dados de acidente de trabalho do setor de serviços. Para a pesquisa é utilizado os dados da CAT entre o período de julho de 2018 a março de 2021, buscando apresentar um estudo geral dos dados referentes ao setor de serviço.

Na análise são utilizadas 12 técnicas de mineração de dados, aplicadas ao conjunto de dados do setor de serviço disponibilizados pela CAT. Os resultados da mineração de dados são comparados por meio de métricas, que demostram quais foram os algoritmos com melhor desempenho.

Materiais e métodos

O presente estudo realizou uma pesquisa exploratória, com fonte de pesquisa secundária utilizando como base os dados da CAT. Para realizar a análise optou-se em utilizar uma abordagem quantitativa. O estudo foi realizado em três etapas, sendo estas: planejamento; coleta e pré-processamento; aplicação de técnicas de mineração e por fim análise dos resultados.

A etapa de planejamento baseou-se no estudo do contexto do tema, composto por dados e informações encontrados na literatura, referente ao setor de serviços e acidente ocupacional. Em seguida sucedeu-se a etapa de coleta e pré-processamento, onde foi utilizado o banco de dados da CAT, em razão do seu alcance, credibilidade e disponibilidade. Esta etapa foi realizada utilizando o software Microsoft Excel e o Power BI.

Entre o período disposto de julho de 2018 a março de 2021 foi contabilizado 1.213.057 ocorrências, esses dados foram agrupados em uma planilha que continha inicialmente 25 colunas. A partir de uma análise verificou-se a presença de informações repetidas ou sem grande importância para o estudo, possibilitando assim a redução das variáveis a serem consideradas.

Tendo em consideração o enfoque do estudo, aplicou-se um filtro para obtenção apenas dos dados referentes ao setor de serviços, reduzindo os dados a serem analisados para 14.098 ocorrências, onde apenas 33 registros indicaram óbito do trabalhador.

Para aplicação da mineração de dados foi utilizado o ambiente do *Jupyter Notebook*, com modelos desenvolvidos na linguagem *Python*. O conjunto de dados balanceado foi processado em 12 técnicas de mineração de dados, buscando prever a ocorrência de óbitos a partir de variáveis registrados com a abertura da CAT. Na etapa de aplicação de técnicas de mineração foi realizado um balanceamento entre os dados de óbito e não óbito, deixando a quantidade de ocorrências equilibrada, resultando ao final 66 registros, composto por 50% de ocorrências com óbitos e 50% sem óbito, sendo aleatória a seleção dos dados sem óbito.

Posteriormente foram executados os algoritmos selecionados: *Bagging*, *Extra Trees*, *Random Forest*, *Stacking*, *Voting*, *XGBoost*, *Decision Trees*, *K-Nearest Neighbors*, *Logistic Regression*, *Naive Bayes*, *Neural Networks* e *Support Vector Machine.*

Os resultados das técnicas foram analisados a partir de cinco métricas: acurácia, precisão, *recall*, F1 score e curva ROC/AUC (*Receiver Operating Characteristic/Area Under Curve*). Levando em consideração os dados obtidos, uma análise dos resultados foi realizada ao final da pesquisa.

Resultados e Discussão

Utilizando a função *train_test_split do scikit-learn*, dividiu se o conjunto de dados em 70% treino e 30% teste. Após a caracterização do conjunto de dados de treino e teste, o conjunto de dados foi submetido às 12 técnicas de mineração, onde foi possível avaliar os resultados das métricas e tempos de execução para cada uma das técnicas. Para garantir uma maior assertividade foram gerados dez conjuntos aleatórios para serem executados, alterando apenas dos dados de “não obtido”. A tabela abaixo apresenta a média dos resultados dos dez conjuntos de dados executados.

Tabela 1. Resultados das métricas e tempo computacional das técnicas

Técnica	Acurácia	Precisão	*Recall*	F1 score	ROC/AUC	Tempo (s)
			Ensemble			
ET	0,8200	0,8426	0,7798	0,8071	0,8963	6,6986
RF	**0,8600**	**0,8520**	**0,8564**	**0,8410**	**0,9245**	5,2631
XGB	0,6500	0,7115	0,5693	0,6173	0,6933	**4,9336**
BA	0,7850	0,7678	0,7619	0,7569	0,8847	9,0730
VO	0,7600	0,7400	0,8396	0,7791	0,8514	6,2058
ST	0,7550	0,7250	0,7693	0,7283	0,8197	6,3535
			Não *ensemble*			

SVM	0,5250	0,5407	0,6851	0,5637	0,6158	5,4563
LR	0,6150	0,6763	0,6161	0,6024	0,7152	6,4906
KNN	0,6200	0,6441	0,5728	0,5914	0,6651	8,0389
NB	**0,8639**	**0,8398**	**0,9028**	**0,8720**	**0,9373**	**5,1878**
DT	0,7300	0,7717	0,6875	0,7128	0,7296	5,9266
NN	0,6950	0,6483	0,8038	0,7013	0,7650	6,3284

Fonte: Autoria própria (2022)

As técnicas categorizadas como *ensemble* obtiveram resultados melhores nas métricas de precisão e também no menor tempo computacional, aquelas elencadas como não *ensemble*, obtiveram melhor desempenho nas métricas de acurácia, *recall*, F1 score e ROC/AUC. Porém, de forma geral é possível destacar que as técnicas categorizadas como *ensemble* apresentaram uma média maior que as técnicas não *ensemble*.

Nas técnicas *ensemble*, *Random Forest* (RF) obteve os melhores resultados de métricas, com exceção do menor tempo computacional, onde o melhor resultado esteve associado à *XGBoost* (XGB). Já em relação às técnicas não *ensemble*, *Naïve Bayes* (NB) acumulou os melhores resultados de métricas, incluindo o menor tempo computacional. Tiveram exceções nas duas subdivisões, pois os resultados das métricas da técnica de *XGBoost* (XGB) foram próximos aos algoritmos não *ensemble*, assim como *Naive Bayes* (NB) obteve resultados semelhantes às técnicas caracterizadas como *ensemble*, mesmo não estando nesta categoria.

Para os algoritmos não *ensemble* as métricas acurácia, precisão e F1 score apresentaram uma média abaixo de 70%, com exceção das métricas recall e ROC/AUC. A técnica que se destacou foi *Naïve Bayes*, apresentando os melhores percentuais em todas as métricas. Apenas a técnica *Support Vector Machine* (SVM) apresentou um percentual médio inferior a 60%.

De maneira geral, os percentuais das métrica *ensemble* foram superiores a 75%, com destaque para a métrica ROC/AUC que apresentou um percentual médio acima de 80%. A técnica que se destacou foi *Random Forest* (RF), apresentando os melhores resultados em todas as métricas.

Conclusões

Respeitar as diretrizes normativas, assim como disponibilizar e inspecionar a utilização de Equipamentos de Proteção Coletiva (EPCs) ou Individual (EPIs) são práticas essenciais na SST, tanto para o setor de serviços, como para os demais setores. A utilização de técnicas de mineração de dados vinculados a dados de saúde e segurança do trabalho podem ser relevantes para definir prioridades de intervenção, além de possibilitar a avaliação dos impactos das intervenções tanto no

setor público quanto no privado, atuando na redução de ocorrências para o setor de serviço.

Agradecimentos

Agradeço a minha orientadora, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq) pela oportunidade de desenvolvimento da pesquisa.

Referências

ARBACHE, J. **Produtividade no Setor de Serviços**. In: DE NEGRI, F.; CAVALCANTE, L. R. (Orgs.). Produtividade no Brasil: desempenho e determinantes. Brasília: IPEA, vol. 2, p. 277-300, 2006.

SANTANA, V. S; MAIA, A.P.; CARVALHO C.; LUZ, G. Acidentes de trabalho não fatais: diferenças de gênero e tipo de contrato de trabalho. Cad. Saúde Pública; 19:481-93, 2003.

TOMPA, E.; MOFIDI, A.; VAN DEN HEUVEL, S.; VAN BREE, T.; MICHAELSEN, F.; JUNG, Y.; PORSCH, L.; VAN EMMERIK, M. O valor da segurança e da saúde ocupacional e os custos sociais de lesões e doenças relacionadas ao trabalho. Agência Europeia de Segurança e Saúde no Trabalho (EU-OSHA), 2019.