APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS DE ACIDENTES DE TRABALHO NA CONSTRUÇÃO CIVIL

Mariana Néspoli de Melo (PIBIC/Fundação Araucária/UEM), Gislaine Camila Lapasini Leal (Orientadora) e-mail: gclleal@uem.br, Edwin Vladimir Cardoza Galdamez (Coorientador)

Universidade Estadual de Maringá / Centro de Tecnologia, Maringá, PR.

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

Palavras-chave: Saúde e Segurança do Trabalho, Algoritmos, Setor da Construção Civil.

Resumo:

Os dados de acidentes e doenças ocupacionais são alarmantes e impactam diretamente na economia do país. O setor da construção civil é apontado como um dos mais perigosos do mundo. O objetivo deste estudo é utilizar de técnicas de mineração de dados e ferramentas estatísticas para identificar algum padrão ou correlação dentre os acidentes neste setor, de forma que seja possível diminuir consideravelmente o número de acidentes e doenças ocupacionais, para isso foram aplicadas 12 técnicas de mineração de dados. Como resultado foi possível estabelecer os atributos que possuem mais relevância quanto ao óbito do trabalhador, sendo: natureza da lesão, parte do corpo atingida e agente causador. Com isso, gestores podem realizar ações para incentivar o uso de Equipamentos de proteção, visando à segurança do trabalhador. Portanto, a utilização das técnicas de MD contribuíram para o objetivo do projeto.

Introdução

Os acidentes e doenças causados pelo trabalho estão presentes desde a antiguidade. Apesar de vários relatos, nunca foi prioridade cuidar da saúde e segurança do trabalhador. Até que, em meados do século XVIII, com a Revolução Industrial, que foi dada certa atenção para este problema. Durante muitos anos, as medidas criadas para manter a segurança do trabalhador foram ineficazes em diminuir o número de acidentes. Essa circunstância ganhou destaque com a adoção pela OIT (Organização Internacional do Trabalho), em 1988, da Convenção 167 sobre segurança e saúde na construção (Júnior, Valcárcel e Dias, 2005).
Com base nos estudos da OIT, o Brasil ocupa o 4º lugar em relação ao número de mortes, com 2503, ficando atrás apenas da China (14.924), Estados Unidos (5.764) e Rússia (3.090). Segundo o observatório de Saúde e Segurança do trabalho (SmartLab), em 2020 o Brasil teve 446,9 mil

notificações de acidentes de trabalho notificados pela CAT (INSS/CATWEB, 2020) das quais 1,9 mil resultaram no óbito do trabalhador. O setor da Construção representou 7.129 desses acidentes.

Nas últimas décadas, em que a maioria das operações e atividades das instituições públicas e privadas são registradas computacionalmente e acumulam-se em grandes bases de dados, a mineração de dados – *Data Mining* (DM) – é uma das alternativas mais eficazes para extrair o conhecimento a partir de grandes volumes de dados, descobrindo relações ocultas, padrões e gerando regras para predizer e correlacionar dados, que podem auxiliar instituições nas tomadas de decisões de forma mais rápida e com maior grau de confiança (Cardoso, 2008).

Materiais e métodos

Por meio da detalhada análise bibliométrica, foi selecionado o banco de dados da CATWEB, por apresentar os dados de forma acessível e pública. Nele, contém todos os arquivos relacionados às ocorrências ocupacionais no Brasil, independente da gravidade do acidente. Devido ao grande número de dados, foram realizadas algumas etapas de limpeza e remoção de *outliers* . A aplicação foi feita no ambiente Jupyter Notebook, na distribuição Anaconda e os modelos foram desenvolvidos através de funções do scikit-learn, na linguagem Python.

Resultados e Discussão

Ao final da etapa de remoção de *outliers*, limpeza e segmentação para o setor da Construção Civil, o conjunto de dados passou de 990.870 instâncias para 35.051. Para obter uma melhor análise e interpretação desses dados, foram feitos gráficos em relação aos atributos, como por exemplo quantidade de ocorrências ocupacionais no mês, onde foi possível observar que mais de 1 milhão de trabalhadores sofreram algum tipo de acidente ou doença ocupacional no período de junho de 2018 a março de 2020, no setor da construção civil. Além disso, o local que ocorre mais acidentes é o típico, representando 85,5% do total e pode ser visto na Figura 1.

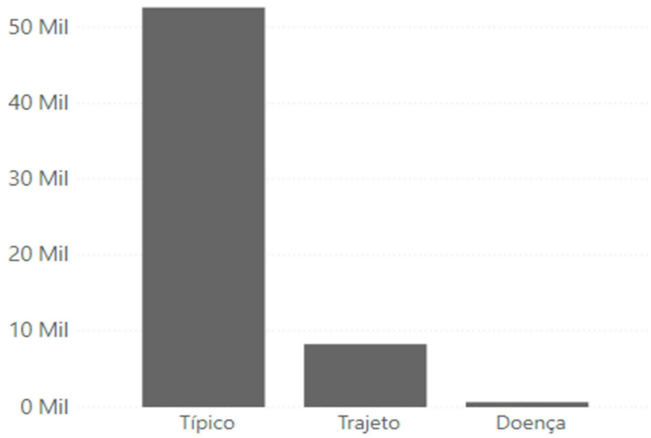

Figura 1 – Local do acidente

Com os dados preparados, seguiu-se para a etapa de execução e para isso foi preparado um conjunto balanceado com 50% de ocorrências com óbito e 50% sem óbito, selecionados de forma aleatória. Para a mineração foram utilizadas doze técnicas, seis classificadas como ensemble (Bagging, Extra Trees, Random Forest, Stacking, Voting e XGBoost) e o restante como não ensemble (Decision Trees, K-Nearest Neighbors, Logistic Regression, Naive Bayes, Neural Networks e Support Vector Machine). Os resultados das técnicas foram interpretados a partir das seguintes métricas: acurácia, precisão, recall, F1 score e curva ROC/AUC que são funções do *scikit-learn.* Utilizando a função train_test_split do scikit-learn, que divide o conjunto de dados em 70% treino e 30% teste de forma aleatória, 455 instâncias formaram o conjunto de treino e 195 o de teste, já que o conjunto balanceado continha 650 instâncias. Assim, as técnicas foram aplicadas e os resultados das métricas puderam ser comparados e estão apresentados na Tabela 1, a seguir:

Tabela 1 – Resultado das métricas e tempo computacional das técnicas
Fonte: O autor

Técnica	*Acurácia*	*Precisão*	*Recall*	*F1 score*	*ROC/AUC*	*Tempo(s)*
			Ensemble			
ET	**0,9077**	**0,9158**	0,8969	**0,9062**	**0,9527**	31,8345
RF	0,8667	0,8545	**0,9038**	0,8785	0,9249	**5,4250**
XGB	0,8872	0,8977	0,8587	0,8778	0,9475	34,6183
BA	0,8872	0,8889	0,8889	0,8889	0,9498	26,3688
VO	0,7692	0,7228	0,8111	0,7644	0,8630	20,9311
ST	0,8820	0,8835	0,8922	0,8878	0,9431	5,5571
			Não Ensemble			
SVM	0,7267	0,7333	0,6377	0,6822	0,8106	3,7272
LR	0,6461	0,6588	0,5833	0,6188	0,6627	**1,7642**
KNN	0,7744	0,8043	0,7400	0,7708	0,8290	10,6557
NB	**0,8974**	**0,8932**	**0,9109**	**0,9020**	**0,9580**	2,5300
DT	0,8051	0,7654	0,7654	0,7654	0,8148	17,7869
NN	0,7538	0,8061	0,7315	0,7670	0,8477	20,4945

Apesar da tabela de resultados mostrar as técnicas que tiveram o melhor desempenho nas métricas, isso ainda não nos permite compreender qual atributo tem mais relevância nas ocorrências de óbito. Para tanto, foram selecionadas as técnicas com os melhores desempenhos de cada grupo para compor a etapa de explicação. Assim, as técnicas selecionadas foram: Extra Trees (ET) do grupo ensemble e Naive Bayes (NB) do grupo não ensemble. Utilizou-se um algoritmo de inteligência artificial explicável, o SHapley Additive exPlanations (SHAP), que tem como objetivo facilitar o entendimento quanto ao funcionamento e as saídas de um modelo de aprendizagem de máquinas.

Analisando os gráficos gerados pelo SHAP tem-se que os atributos que contém mais relevância para a ocorrência de óbitos são a parte do corpo atingida, natureza da lesão e o agente causador.

Conclusões

Com base nos gráficos e tabelas podemos concluir que os algoritmos que tiveram a melhor performance foram o Extra Trees (ET), técnica ensemble e o Naive Bayes (NB), não ensemble. Ambos tiveram um bom desempenho, a técnica ET atingiu 90% em todas as métricas e o algoritmo (NB) variou entre 89-95%. Utilizando o algoritmo SHAP foi possível obter a relevância de cada atributo relacionado aos resultados das técnicas, e com isso podemos concluir que os que têm mais relevância no óbito do trabalhador são: parte do corpo atingida, natureza da lesão e agente causador do acidente. Em contrapartida, o sexo, idade, tipo de acidente, CBO e o UF do empregador quase não interferem. Diante disso, pode-se concluir que a técnica de mineração de dados pode contribuir para que o objetivo do projeto fosse cumprido.

Agradecimentos

Agradeço ao CNPq e Fundação Araucária pelo suporte financeiro.

Referências

CARDOSO, O.N.P., MACHADO, R.T.M. **Gestão do conhecimento usando Data mining: estudo de caso na Universidade Federal de Lavras.** Revista Adm. Pública, p.495-528, 2008.

JÚNIOR, J.M.L., VALCÁRCEL, A.L., DIAS, L.A. **Segurança e Saúde no Trabalho da Construção: Experiência brasileira e panorama internacional.** Secretaria Internacional do Trabalho, 2005.

OBSERVATÓRIO DE SAÚDE E SEGURANÇA DO TRABALHO (SMARTLAB). Disponível em: https://smartlabbr.org/sst. Acesso em: 24 de agosto de 2021.