ConBRepro
X CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO
EVENTO ON-LINE
02 a 04 de dezembro 2020

Regressão logística aplicada a dados de saúde e segurança do trabalho

Ageu de Araujo Machado
Departamento de Engenharia de Produção – Universidade Estadual de Maringá
Beatriz Lavezo dos Reis
Departamento de Engenharia de Produção – Universidade Estadual de Maringá
Daniele Cristina Tita Granzotto
Departamento de Estatística - Universidade Estadual de Maringá
Edwin Vladimir Cardoza Galdamez
Departamento de Engenharia de Produção – Universidade Estadual de Maringá

Resumo: O conceito de Saúde e Segurança do Trabalho (SST) é estudado por diversas áreas do conhecimento, e aborda o acompanhamento, mitigação e prevenção de acidentes e doenças que estão ligadas ao trabalho exercido por um indivíduo. É crescente o número de estudos que buscam investigar esse tema, assim como é grande o número de acidentes que ocorrem ao redor do mundo, registrados através de bancos de dados. Muitos desses *databases* são de domínio público, disponibilizados por organizações privadas e instituições governamentais, sendo um exemplo as informações da Previdência Social do Brasil referentes aos acidentes com abertura de Comunicação de Acidentes de Trabalho (CAT). Diante disso, o objetivo deste estudo é apresentar um panorama dos acidentes de trabalho no Brasil. Para atingir esse propósito será utilizado o banco de dados de abertura de CATs do país, englobando informações do período de julho de 2018 a junho de 2019. Com as informações levantadas, é necessário aplicar técnicas estatísticas para entender o comportamento e correlações entre as variáveis do banco de dados, utilizando, assim, o modelo de regressão logística. Portanto, espera-se como resultado dessa aplicação, verificar o cenário atual dos acidentes de trabalho no Brasil e entender a probabilidade de acontecer uma adversidade levando em consideração algumas variáveis independentes do modelo.

Palavras-chave: Saúde e segurança do trabalho, Acidentes de trabalho, Regressão logística.

Logistic regression applied to occupational health and safety data

Abstract: The concept of Occupational Safety and Health (OSH) is studied by several areas of knowledge, and approaches the monitoring, mitigation and prevention of accidents and diseases linked to the work performed by an individual. There is an increasing number of studies seek to investigate this topic, as well as a large number of accidents that occur around the world, registered through databases. Many databases are in public domain, available by private institutions and government agencies, for example, the information provided by Social Security in Brazil relative to accidents with the opening of Work Accident Communication (CAT). Therefore, the objective of this study is to present an overview of occupational accidents in Brazil. To achieve this purpose, the country's CATs opening database will be used, encompassing information from July 2018 to June 2019. With information gathered, it is necessary to apply statistical techniques to understand the

behavior and correlations between database variables, using the logistic regression model. Therefore, it is expected as an application result, to verify the current scenario of occupational accidents in Brazil and to understand a probability of an adversity taking into account some independent variables of the model.

Keywords: Occupational health and safety, Occupational accidents, Logistic regression.

1. Introdução

Estima-se que ocorram aproximadamente 374 milhões de acidentes de trabalho anualmente em uma abrangência global, totalizando mais de 2,78 milhões de homens e mulheres levados a óbito em decorrência de acidentes e doenças relacionadas ao trabalho (ILO, 2019). Este número representa que cerca de quatro por cento da riqueza obtida pelos países é consumida em custos com ausências ao trabalho, tratamentos, reabilitação, pensões e subsídios emergentes de lesões, mortes e doenças profissionais.

Prezar pela segurança e saúde do trabalho, além de ser uma obrigação legal das organizações, pode ser também um diferencial competitivo para as empresas, pois com a prevenção e mitigação de acidentes e doenças causados ao trabalhador são reduzidos os danos aos recursos humanos, assim como os custos envolvidos (CIARAPICA; GIACCHETTA, 2009). Para auxiliar o acompanhamento da SST, mundialmente são gerados bancos de dados que registram a ocorrência de eventos ao trabalhador. Alguns com acesso aberto, como as informações disponibilizadas pela Organização Internacional do Trabalho e Previdência Social do Brasil, e outros restritos à organização, como os registros particulares de empresas.

Para analisar e reduzir as ocorrências de doenças, acidentes e óbitos, algumas ferramentas podem ser utilizadas, sendo a mineração de dados representativa nesse âmbito, pois possibilita a avaliação de informações em grande escala. O conceito de *data mining* está associado ao processamento de dados para extração de padrões e se relaciona, mas não somente, com as etapas de preparação dos dados, execução do modelo e análise dos resultados obtidos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Dentre as técnicas utilizadas na mineração de dados, como árvores de decisão, regras de associação, *Support Vector Machine* (SVM), entre outras, a regressão logística também é empregada. Suas aplicações são amplas, como na detecção de pacientes com câncer (BELCIUG, 2020), em pesquisas de nanotecnologia (SHEHADEH; EBRAHIMI; OCHIGBO, 2012) ou ainda sobre o retorno de trabalhadores para sua função após um acidente (LEE; KIM, 2018).

Nesta pesquisa, o objetivo é apresentar um panorama dos acidentes de trabalho no Brasil, utilizando como técnica de mineração de dados a regressão logística, aplicada a um banco de dados de acidentes com abertura de CAT. Para isso, este trabalho apresenta cinco seções, onde esta descreve uma introdução ao tema, seguida pelo referencial teórico da pesquisa, que apresenta a literatura referente a SST e regressão logística. Em sequência é apontada a metodologia aplicada, os resultados obtidos e discussões acerca deles, e, por fim, a pesquisa se encerra com as considerações finais.

2. Referencial teórico

2.1 Saúde e segurança do trabalho

A saúde e segurança do trabalhador não era considerado um fator relevante no início da industrialização, mas com as revoluções industrias e evoluções da sociedade, passou a ser um aspecto decisivo no ambiente de trabalho (CIARAPICA; GIACCHETTA, 2009). O conceito de SST está relacionado a práticas que sejam capazes de reduzir ou prevenir acidentes e doenças que são causados em vista do trabalho que um indivíduo executa.

Estima-se que a cada 15 segundos um trabalhador é morto e outros 160 estão envolvidos em acidentes ocupacionais ao redor do mundo (CHEN *et al*., 2020). Esses números estão principalmente associados a países subdesenvolvidos ou em desenvolvimento, pois com os avanços tecnológicos e alterações nas leis, os países desenvolvidos possuíam subsídios para acompanhar as mudanças também no âmbito de gestão de riscos e saúde e segurança do trabalhador, aspecto este que não pôde ser observado nos demais países (BADRI; BOUDREAU-TRUDEL; SOUISSI, 2018).

Os acidentes e doenças associados ao trabalho são responsáveis por ocasionar perdas ao capital humano, mas não apenas isso, também são causadores de prejuízos financeiros às organizações e ao governo. Dessa forma, desenvolver ações para análise dos problemas relacionados a SST e propor mecanismos para redução dos eventos, é fundamental para a estratégia das organizações, evitando danos aos recursos humanos, interrupções nos processos ou ainda problemas com a reputação da organização (FERNÁNDEZ-MUÑIZ; MONTES-PEÓN; VÁZQUEZ-ORDÁS, 2012).

Para reduzir os danos causados por acidentes e doenças, tem sido desenvolvidas pesquisas sobre as legislações e normas, aplicações em empresas, análises de bancos de dados, entre outras práticas. O surgimento da gestão de SST está associada ao Reino Unido, que em 1991 desenvolveu um guia a fim de auxiliar a criação de melhorias pelos próprios colaboradores, relacionadas a sua saúde e segurança (YOON *et al*., 2013).

O estudo e acompanhamento de aspectos psicossociais no ambiente de trabalho também está associado à saúde e segurança dos colaboradores, como analisado por Hohnen e Hasle (2018) o impacto, dificuldades e aplicações da OHSAS 18001 em organizações espanholas. Yanar, Lay e Smith (2019) avaliaram o impacto que um supervisor pode ter nas lesões ocupacionais, considerando seu apoio aos colaboradores, concluindo que essa supervisão resulta em um ambiente de trabalho mais seguro e auxilia na redução do risco de acidentes e doenças. Christie e Ward (2019) pesquisaram sobre os riscos associados a trabalhadores da economia GIG, profissão que apresenta rotinas mais flexíveis, onde os trabalhadores são diariamente expostos a acidentes de trânsito.

Outra abordagem que apoia o trabalho de prevenção e mitigação de acidentes e doenças é o estudo dos registros de eventos que já aconteceram, avaliando casos em específico ou bancos de dados abertos. Gerassis *et al*. (2017) utilizaram informações da construção civil e mineração para entender as causas de acidentes que ocorriam em obras de aterro. Shirali, Noroozi e Malehi (2018) utilizam informações de lesões ocupacionais de baixa, média gravidade e algumas fatais, que ocorreram em uma indústria siderúrgica do Irã. Lee e Kim (2018) utilizaram técnicas de mineração de dados, como regressão logística, floresta aleatória e SVM para analisar o retorno de colaboradores ao trabalho após um acidente industrial.

2.2 Regressão logística

A regressão logística é uma ferramenta dentro das estatísticas aplicadas, que pode ser utilizada em diversas áreas e possui um papel importante nas pesquisas desenvolvidas por estes ramos (DAS; MAITI; PRADHAN, 2010). Também é conhecida como uma das técnicas direcionadas para a tarefa de classificação de dados, que consiste em uma aprendizagem supervisionada (JIANG; JOSSE; LAVIELLE, 2019).

A regressão logística não é a única na estatística, também é encontrada a regressão linear, que pode ser simples ou múltipla. No primeiro caso, considerando apenas um único preditor (x) e uma variável dependente (y), no segundo caso são utilizados mais de um preditor (x) no modelo (MONTGOMERY; RUNGER, 2010). Na regressão logística, ocorre a modelagem do logaritmo natural das chances, onde a variável dependente (y) expressa a probabilidade logarítmica. Essa probabilidade é expressa pela equação *logit* (LEVIN; FOX;

FORDE, 2012). Mesmo sendo uma área consolidada na estatística, há poucas aplicações e soluções para modelos de regressão logística, ainda mais considerando modelos binários, com duas ou mais variáveis dependentes (HAINES; KABERA; NDLOVU, 2018; JIANG; JOSSE; LAVIELLE, 2019).

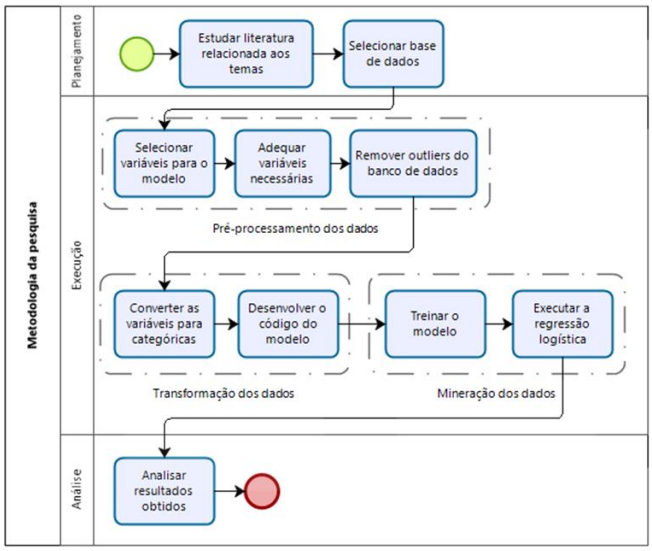
Algumas pesquisas são desenvolvidas com foco em melhorar os modelos logísticos ou ainda os algoritmos associados a eles. Li e Lederer (2019) propõem um novo modelo para calibração da regressão logística, considerando o parâmetro ℓ_{1-} penalizado, apresentando um esquema de fácil implementação e competitivo com os que já foram desenvolvidos. Haines, Kabera e Ndlovu (2018) desenvolvem um modelo ideal para aplicação em projetos de regressão logística binária, onde as variáveis estão envolvidas com o contexto de medicamentos.

Outros estudos são dedicados a aplicações de dados a regressão logística, informações essas advindas de diversas áreas. Le, Tran e Huynh (2018) utilizam a regressão logística multivariada para entender o comportamento de rotação interna de espécies químicas. Wei e Ghosal (2020) aplicam o modelo estatístico para estudar as propriedades de contração relacionadas ao encolhimento de componentes. Outros autores utilizam os métodos logísticos para aplicações em saúde e segurança do trabalho, como na comparação de técnicas de mineração de dados para avaliar o comportamento de trabalhadores (LEE; KIM, 2018) ou ainda, na avaliação dos riscos de acidentes a funcionários de estações de esqui (BOHANEC; DELIBAŠIĆ, 2015).

3. Metodologia

A pesquisa foi conduzida em três etapas, representadas pelo planejamento, execução e análise do estudo, conforme descrito na Figura 1. Dentro da etapa de execução da pesquisa existem, ainda, subdivisões que representam as fases do processo associado a mineração de dados, que envolvem o pré-processamento dos dados, sua transformação e a mineração.

Figura 1 - Metodologia de pesquisa



Fonte: Autores (2020)

Para iniciar a pesquisa, em seu planejamento, inicialmente foi conduzida uma revisão de literatura, estudando publicações representativas em relação aos temas de saúde e segurança no trabalho e regressão logística. Em seguida, mapeando as bases de dados de acesso público, foi escolhida aquela que mais se adequava ao objetivo da pesquisa: o banco de dados de acidentes com abertura de CAT, disponibilizado pela Previdência Social do Brasil.

Com as informações escolhidas, a etapa seguinte foi a seleção de variáveis considerando o montante escolhido, que em seguida foram adequadas as variáveis necessárias. Também foram removidos os outliers, ou seja, os dados que não eram condizentes com o banco de dados e que causariam um desvio errôneo na pesquisa, encerrando o pré-processamento dos dados.

Na transformação foram realizadas a categorização das variáveis, para facilitar o desempenho do modelo e também o desenvolvimento do código computacional, no *software RStudio*. Na fase de mineração dos dados, incialmente foi escolhido um banco de dados aleatório para treinamento do modelo e em seguida foi aplicada a regressão logística no banco de dados completo. Com os resultados da regressão foi possível avaliar as saídas do modelo, as correlações e probabilidades entre as variáveis e assim, concluir a pesquisa.

4. Resultados e discussões

4.1 Coleta e pré-processamento dos dados

As informações selecionadas para a pesquisa foram os bancos de dados disponibilizados pela Previdência Social do Brasil, de acidentes ocorridos com abertura de CAT. As informações são agrupadas pelas ocorrências a cada trimestre e para este estudo foram selecionados os dados relativos à um ano completo, período relativo a julho de 2018 a junho de 2019, ou seja, quatro trimestres de dados alocados em quatro planilhas diferentes.

Essas informações foram agrupadas em apenas uma planilha que continha 475.440 ocorrências de acidentes, dispostas em 25 colunas. Algumas dessas colunas apresentavam informações repetidas ou não relevantes para a pesquisa, portanto o primeiro passo foi a seleção das variáveis que iriam compor o modelo. A variável dependente escolhida para o estudo foi o óbito, que representava, dentro do montante de acidentes, aqueles que estavam associados a ocorrência de morte do trabalhador.

Como variáveis independentes foram escolhidas seis, sendo elas: idade, sexo, Classificação Nacional de Atividades Econômicas (CNAE), parte do corpo atingida, tipo de acidente e espécie de benefício que o colaborador, ou sua família, recebeu. Em seguida o banco de dados passou por adequações, pois a idade precisava ser calculada. Para isso foram utilizadas as datas de nascimento e de ocorrência dos acidentes que estavam disponíveis no banco de dados, definindo com quantos anos o colaborador sofreu aquele acidente.

Também foi realizada uma limpeza nos dados, onde foram removidos os outliers que poderiam atrapalhar a execução do modelo e o desempenho do resultado. Foram retiradas inicialmente as idades que correspondiam a mais de 85 anos ou menos de 16, considerando que são faixas etárias com pouco possibilidade de trabalho ou que não são permitidos perante a legislação. Além disso, algumas ocorrências não apresentavam a data de nascimento da vítima e também foram retiradas.

Em relação ao sexo, foram considerados apenas masculino e feminino, pois haviam poucos eventos registrados com indeterminado ou indefinido. Os campos de CNAE e parte do corpo atingida apresentavam como resposta "*n class*" ou não classificado, que também foram retirados da amostra, assim como os tipos de acidente que apresentavam a descrição "ignorado". Com a retirada dos outliers o banco de dados passou a apresentar 472.671 linhas, 2.769 a menos ocorrências do que a amostra inicial.

Após o pré-processamento dos dados, a etapa seguinte foi de conversão das variáveis, que estavam na forma texto e deveriam ser representadas em variáveis categóricas. A primeira alteração foi feita na variável dependente, onde as ocorrências de acidentes sem morte são associadas a $y = 0$ e acidentes com óbito são $y = 1$. A idade do colaborador foi considerada como a quantidade inteira de anos completos até a data do acidente. O sexo

do indivíduo foi considerado como $x = 0$ para os casos em que é feminino e $x = 1$ para masculino.

Considerando o CNAE, foram definidas 21 variáveis para expressar cada seção de atividade econômica, variando o $x = 0, 1, 2 \ldots 20$. Para essa divisão considerou-se a especificação da Comissão Nacional de Classificação (CONCLA). Para as partes do corpo foram definidos seis conjunto, variando $x = 0, 1, 2, 3, 4, 5, 6$ que correspondem respectivamente à: cabeça, tronco, membros superiores, membros inferiores, sistemas e aparelhos, e partes múltiplas. Esses grupos foram selecionados agrupando 41 possibilidades de membros que apresentava o banco dados.

Para os tipos de acidentes foi considerado como $x = 0$ para as doenças, $x = 1$ para os acidentes típicos e $x = 2$ para os acidentes de trajeto. A última variável corresponde a espécie de benefício que o funcionário ou sua família recebeu após o acontecimento do acidente, sendo $x = 0$ para o tipo PA, $x = 1$ para afastamentos de até 15 dias, $x = 2$ para auxílio doença e $x = 3$ para pensão por morte.

4.2 Análise descritiva dos dados

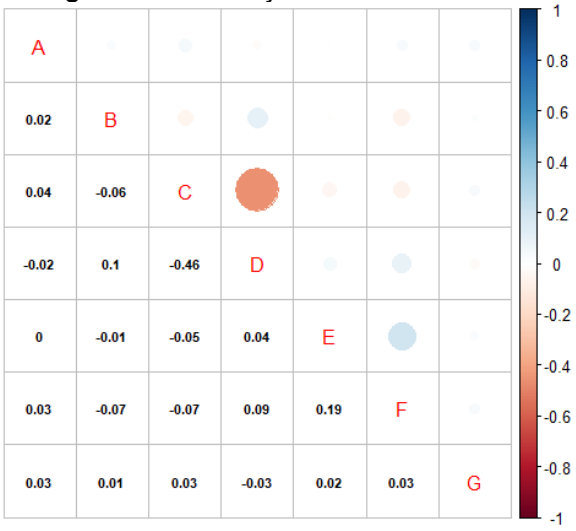
A amostra selecionada, contendo 472.671 registros de acidentes e doenças do trabalho, apresenta menos de 0,5% de óbitos, representados por 2.162 destes eventos. O sexo mais representativo é o masculino, descrito por 66,3% dos casos registrados, no entanto, quando avaliados apenas os casos de óbito, o percentual equivale à aproximadamente 91%. Também é possível associar a ocorrência de acidentes e mortes com as idades dos trabalhadores, onde o maior número de registros está em torno de 20 a 50 anos de idade, para homens e mulheres. Para os casos de óbito, o pico está na faixa etária de 30 anos para mulheres e 35 anos para homens, nos acidentes sem morte os picos são, respectivamente, 35 e 25 anos.

Considerando os CNAEs, os que apresentam maior incidência na amostra de forma geral são indústria de transformação (26,9%), comércio e reparação de veículos automotores e motocicletas (15,7%), saúde humana e serviços sociais (14,9%), atividades administrativas e serviços complementares (7,5%) e transporte, armazenagem e correio (7,3%). No entanto, quando analisados apenas os casos de óbito essa ordenação não se mantém. A mais expressiva continua sendo a indústria de transformação (18,4%) seguida por comércio e reparação de veículos automotores e motocicletas (15,6%), na terceira colocação está transporte, armazenagem e correio (15,0%), em quarto lugar a construção (13,4%) e em quinto estão as atividades administrativas e serviços complementares (8,4%).

Em relação as partes do corpo atingidas, os membros superiores (47,8%) e membros inferiores (28,9%) são os mais atingidos pelas doenças e acidentes, mas os que mais causam o óbito são relacionados a cabeça (33,6%) e partes múltiplas (28,2%). Considerando os tipos de ocorrência, a grande maioria corresponde a acidente típico (75,5%) e também é maior em relação aos casos com morte (58,9%). A espécie do benefício é principalmente voltada para PA de forma geral (98,2%), assim como nos casos de óbito (95,1%), o auxílio doença representa 1,8% dos registros e os demais tem percentual pouco significativo. Para os casos de óbito, pensão por morte representa 4,7% da amostra, ou 101 casos.

Para entender a relação das variáveis independentes entre si e considerando a variável dependente, foi desenvolvido um gráfico de correlação das variáveis (Figura 2). Esse gráfico expressa de forma numérica a relação forte ou fraca entre as dimensões, além de representar por círculos a intensidade da ligação. O nome das variáveis foi substituído pelas letras de A à F, a fim de facilitar a apresentação visual, e representam, respectivamente as variáveis: óbito, idade, sexo, CNAE, parte do corpo atingida, tipo de acidente e espécie de benefício.

Figura 2 - Correlação entre as variáveis



Fonte: Autores (2020)

As medidas de correlação são definidas para um intervalo de -1 a 1 de forma que, quanto mais distante de 0, tanto negativamente quanto positivamente, maior a correlação entre as variáveis. A maior correlação negativa, descrita pelo coeficiente de $-0{,}46$, é representada pelas variáveis C e D, que são respetivamente, sexo e CNAE. Isso representa que as duas variáveis, dentre todas as possíveis correlações, são as que apresentam menor ligação. Em contrapartida, os aspectos E e F, parte do corpo atingida e tipo de acidente, descrevem a maior correlação positiva $(0{,}19)$. As demais correlações são pouco significativas, pois todas estão muito próximas de 0.

4.3 Regressão logística

Como a amostra selecionada apresentava apenas 0,5% de casos de óbito, foi necessário treinar o modelo, inicialmente com uma amostra equilibrada, ou seja, 50% correspondente a casos de acidentes sem óbito e o restante de casos com morte. Esse treinamento foi realizado para detectar possíveis erros do modelo, além de que, ao utilizar uma amostra equilibrada, o resultado da regressão não seria tendencioso. O primeiro resultado encontrado com o modelo de regressão logística está apresentado pela Tabela 1.

Tabela 1 - Estimativas de máxima verossimilhança do modelo

Variáveis	Estimativa	Erro padrão	Valor de z	Pr(>|z|)
(Intercept)	-8,78735	0,132787	-66,176	<2e-16
idade	0,031312	0,001813	17,271	<2e-16
sexo	1,562319	0,078067	20,012	<2e-16
CNAE	-0,033995	0,004976	-6,832	8,40E-12
parte_corpo	-0,067172	0,019052	-3,526	0,000422
tipo_ac	1,007624	0,044972	22,406	<2e-16
especie_ben	0,595924	0,04198	14,195	<2e-16

Fonte: Autores (2020)

O primeiro valor observado é a estimativa do intercepto, ou seja, o valor de β_0, que corresponde a aproximadamente -8,78 quando todas as variáveis possuem valor 0. Os demais coeficientes que acompanham as variáveis independentes na equação são apresentados na coluna de estimativa. A coluna seguinte, que representa o erro padrão,

está associada aos erros dos coeficientes de estimativa. Para avaliar se todas as variáveis apresentam significância para o modelo, o valor de Pr(>|z|) ou p-valor deve ser maior que 0,05, simbolizando que todas as variáveis selecionadas para esse modelo são significativas no teste, conforme apresentado pela última coluna.

Outro resultado da regressão logística são as razões de chance ou *odds ratio* do modelo, que estão relacionadas com a probabilidade de ocorrência do evento em função das variáveis estipuladas. Além disso, quando os valores de razão de chance são próximos a 1 são considerados praticamente insignificantes para análise, pois indicam que não tem ou tem pouca associação entre as duas variáveis (dependente e independente), como é o caso do CNAE e idade, apresentados na Tabela 2.

Tabela 2 - Estimativa de razão de chance (*odds ratio*)

Variáveis	Razão de chance	Erro padrão	Valor de z	Pr(>|z|)
idade	1,0318078	0,0018707	17,2707	<2,2e-16
sexo	4,769871	0,3723706	20,0125	<2,2e-16
CNAE	0,9665761	0,0048099	-6,8316	8,40E-12
parte_corpo	0,9350348	0,0178139	-3,5258	0,0004223
tipo_ac	2,7390853	0,1231819	22,4056	<2,2e-16
especie_ben	1,8147077	0,076182	14,1953	<2,2e-16

Fonte: Autores (2020)

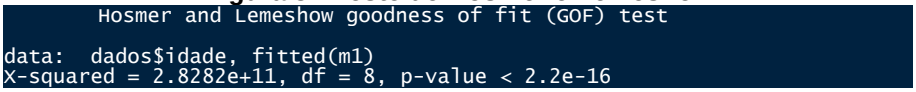
Os coeficientes com valor positivos indicam uma razão de chance maior de acontecer o óbito do que o valor ao qual estão sendo comparados ($x = 0$), como são as variáveis descritas na segunda coluna da Tabela 2, e os valores negativos apresentam uma queda na chance de ocorrência do óbito, em função das variáveis que estão sendo comparadas com $x = 0$.

Para a idade do trabalhador, a interpretação é que para cada ano de vida que se passa, aumenta em 3,18% a probabilidade de ocorrer óbito devido a algum acidente ou doença de trabalho. A chance de uma pessoa do sexo masculino vir a óbito é de 4,77 vezes maior que o sexo feminino.

Em relação a parte do corpo, o tronco tem 6,5% de chances a mais de ocorrência do óbito, quando comparado as demais partes. Quanto ao tipo de acidente, o acidente típico apresenta aproximadamente 2,74 mais chances de óbito em relação as doenças do trabalho. Nos casos de afastamento por mais de 15 dias, as chances acontecer a morte do trabalhador aumentam em 81% se comparadas com PA.

A fim de verificar o ajuste do modelo, alguns testes podem ser realizados após a aplicação da regressão logística. Neste caso foi escolhido o teste Hosmer e Lemeshow, conforme apresentado pela Figuras 3.

Figura 3 - Teste de Hosmer e Lemeshow



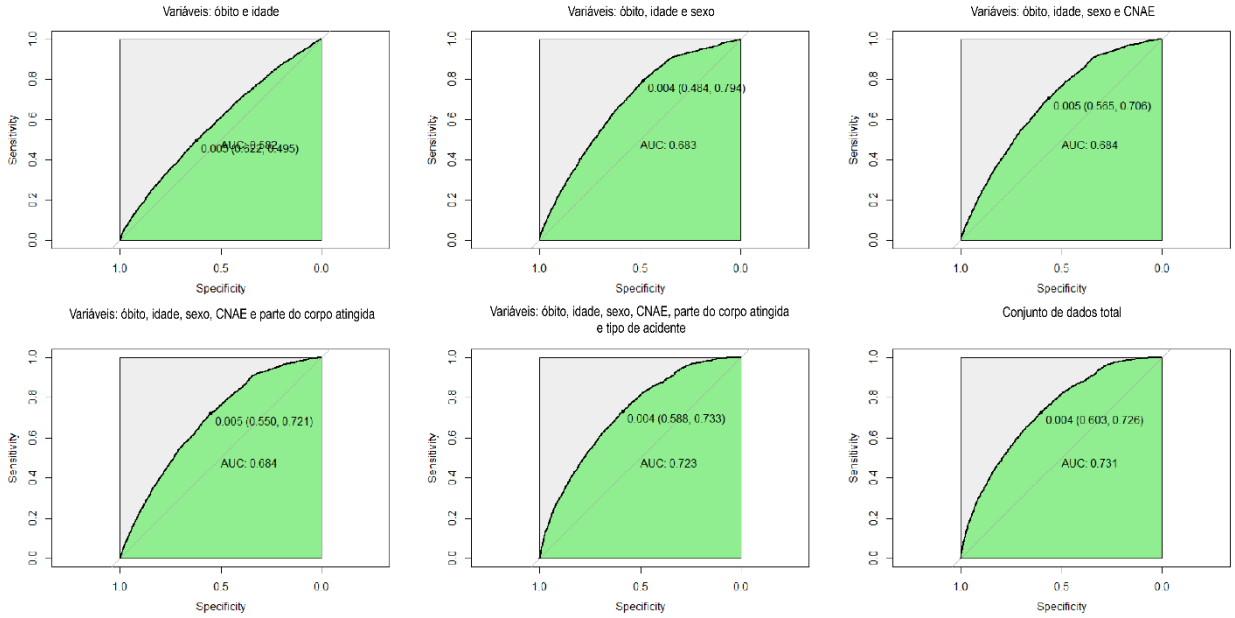
Fonte: Autores (2020)

O teste apresentado na Figura 3 busca determinar a qualidade do ajuste do modelo, onde são realizados comparativos entre os eventos observados e esperados, dividindo a base de dados em 10 grupos, utilizando também a estatística de qui-quadrado. Como resultado deste teste, deve ser observado o p-valor, que analisa se há diferenças significativas entre

o que foi observado e o que é esperado, para significância de 0,05. No teste de Hosmer e Lemeshow aplicado, o p-valor foi extremamente baixo e distante de 0,05, representando que não há diferenças significativas entre o que foi observado e o que é esperado do modelo.

A análise final realizada neste artigo foi em relação a curva ROC, que representa o nível de sensibilidade do modelo, pois quanto mais alta a curva e mais distante da diagonal central, mais sensível é o modelo e mais próximo ao esperado. Para avaliar a sensibilidade da regressão logística realizada e entender a importância das variáveis escolhidas, foi elaborada uma curva ROC para o modelo com a amostra escolhida, mas também foi elaborada uma curva com menor número de variáveis independentes, conforme apresentado nas Figura 4.

Figura 4 - Curvas ROC e variáveis do conjunto de dados



Fonte: Autores (2020)

As curvas que apresentam seu valor de área menor ou igual a 0,5 são considerados modelos sem discriminação, e acima disso até 0,8 são regressões com discriminação aceitável. Em todas as simulações o valor de área foi superior a 0,6, representando discriminações aceitáveis para curva ROC. Além disso, nota-se que com o aumento das variáveis a curva torna-se mais inclinada e mais próxima da área superior do gráfico, assim como o valor da área abaixo do gráfico aumenta de 0,528 para 0,731. No entanto, mesmo apresentando um valor crescente de área com a inserção das variáveis, a curva de ROC ainda não é ideal, necessitando de adaptações ao modelo para a curva se eleve e sua área abaixo da curva se aproxime de 1.

5. Considerações finais

A análise de dados relacionados a saúde e segurança do trabalhador são um fator decisivo e uma estratégia para organizações públicas e privadas. Evitar acidentes, doenças e óbitos são uma forma de reduzir perdas humanas, físicas e financeiras para essas instituições, por isso exigem atenção e cuidado. Uma forma de lidar com o tratamento desses dados é utilizando a regressão logística como técnica de mineração de dados, a fim de entender as correlações entre as variáveis para ocorrência de um evento.

Neste estudo, com o objetivo de entender o cenário da saúde e segurança do trabalho no Brasil, foi utilizado um banco de dados de acesso público, com registros de doenças, acidentes e óbitos. A amostra selecionada possuía quase 500.000 ocorrências, com

algumas variáveis associadas que passaram pela etapa de pré-processamento dos dados. Em seguida, as informações foram processadas pelo *software RStudio*, utilizando a técnica de regressão logística, para analisar como a idade, sexo, CNAE, parte do corpo atingida, tipo de acidente e espécie de benefício poderiam interferir nas probabilidades de ocorrência de óbito a um trabalhador.

Com a aplicação do modelo de regressão logística foi possível avaliar que as variáveis escolhidas eram todas significativas para o modelo. Além disso, pela definição das razões de chance, foi apresentado que os homens possuem 4,77 vezes mais chances de morrer por um acidente ou doença do trabalho do que mulheres, assim como ocorre o aumento em 3,18% da chance de óbito para cada ano de vida que o trabalhador completa. Avaliando a disposição das curvas ROC é possível analisar que a inserção do maior número de variáveis é significativa, no entanto, as curvas ainda estão distantes do ponto (1;0), representando pouca sensibilidade do modelo.

Como sugestão para trabalhos futuros, estão aplicação voltadas a SST no cenário brasileiro, pois poucas publicações foram encontradas na literatura. Além disso, dentro das técnicas utilizadas para dados de SST a regressão logística não é comumente encontrada, portanto também é uma área recomendada para estudos. Diante dos resultados apresentados por esse modelo, com baixa sensibilidade, seria necessário estudar novamente o modelo proposto e amostra selecionada, em busca de resultados mais significativos e melhores correlações entre as variáveis dependente e independentes.

Agradecimentos

Este estudo foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001 e Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq).

Referências

BADRI, A.; BOUDREAU-TRUDEL, B.; SOUISSI, A. S. Occupational health and safety in the industry 4.0 era: A cause for major concern?. **Safety Science**, v. 109, p. 403-411, 2018.

BELCIUG, S. Logistic regression paradigm for training a single-hidden layer feedforward neural network. Application to gene expression datasets for cancer research. **Journal of Biomedical Informatics**, v. 102, p. 103373, 2020.

BOHANEC, M.; DELIBAŠIĆ, B. Data-mining and expert models for predicting injury risk in ski resorts. In: **International conference on decision support system technology**. Springer, Cham, 2015. p. 46-60.

CHEN, H. *et al*. Comparative study on the strands of research on the governance model of international occupational safety and health issues. **Safety Science**, v. 122, p. 104513, 2020.

CHRISTIE, N.; WARD, H. The health and safety risks for people who drive for work in the gig economy. **Journal of Transport & Health**, v. 13, p. 115-127, 2019.

CIARAPICA, F. E.; GIACCHETTA, G. Classification and prediction of occupational injury risk using soft computing techniques: An Italian study. **Safety Science**, v. 47, n. 1, p. 36-49, 2009.

DAS, U.; MAITI, T.; PRADHAN, V. Bias correction in logistic regression with missing categorical covariates. **Journal of Statistical Planning and Inference**, v. 140, n. 9, p. 2478-2485, 2010.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37-37, 1996.

FERNÁNDEZ-MUÑIZ, B.; MONTES-PEÓN, J. M.; VÁZQUEZ-ORDÁS, C. J. Occupational risk management under the OHSAS 18001 standard: analysis of perceptions and attitudes of certified firms. **Journal of Cleaner Production**, v. 24, p. 36-47, 2012.

GERASSIS, S. *et al*. Bayesian decision tool for the analysis of occupational accidents in the construction of embankments. **Journal of Construction Engineering and Management**, v. 143, n. 2, p. 04016093, 2017.

HAINES, L. M.; KABERA, G. M. D-optimal designs for the two-variable binary logistic regression model with interaction. **Journal of Statistical Planning and Inference**, v. 193, p. 136-150, 2018.

HOHNEN, P.; HASLE, P. Third party audits of the psychosocial work environment in occupational health and safety management systems. **Safety Science**, v. 109, p. 76-85, 2018.

JIANG, W.; JOSSE, J.; LAVIELLE, M. Logistic regression with missing covariates—Parameter estimation, model selection and prediction within a joint-modeling framework. **Computational Statistics & Data Analysis**, v. 145, p. 106907, 2020.

LE, T. H. M.; TRAN, T. T.; HUYNH, L. K. Identification of hindered internal rotational mode for complex chemical species: A data mining approach with multivariate logistic regression model. **Chemometrics and Intelligent Laboratory Systems**, v. 172, p. 10-16, 2018.

LEE, J.; KIM, H. R. Prediction of return-to-original-work after an industrial accident using machine learning and comparison of techniques. **Journal of Korean Medical Science**, v. 33, n. 19, 2018.

LI, W.; LEDERER, J. Tuning parameter calibration for ℓ1-regularized logistic regression. **Journal of Statistical Planning and Inference**, v. 202, p. 80-98, 2019.

SHEHADEH, M.; EBRAHIMI, N.; OCHIGBO, A. Predicting the Type of Nanostructure Using Data Mining Techniques and Multinomial Logistic Regression. **Procedia Computer Science**, v. 12, p. 392-397, 2012.

SHIRALI, G. A.; NOROOZI, M. V.; MALEHI, A. S. Predicting the outcome of occupational accidents by CART and CHAID methods at a steel factory in Iran. **Journal of Public Health Research**, v. 7, n. 2, 2018.

WEI, R.; GHOSAL, S. Contraction properties of shrinkage priors in logistic regression. **Journal of Statistical Planning and Inference**, v. 207, p. 215-229, 2020.

YANAR, B.; LAY, M.; SMITH, P. M. The interplay between supervisor safety support and occupational health and safety vulnerability on work injury. **Safety and Health at Work**, v. 10, n. 2, p. 172-179, 2019.

YOON, S. J. *et al*. Effect of occupational health and safety management system on work-related accident rate and differences of occupational health and safety management system awareness between managers in South Korea's construction industry. **Safety and Health at Work**, v. 4, n. 4, p. 201-209, 2013.