



Desenvolvimento e análise de algoritmos aplicados na predição de acidentes em ambiente fabril

Brayan C. Ksenhuck; Mário E. M. Vieira;
Thiago A. Lechuga
Advanced Analytics – Supply Chain
Cervejaria Ambev
São Paulo, Brasil
e-mail: brayan.ksenhuck; mario.vieira;
thiago.lechuga@ambev.com.br

Fabrcio O. Bezerra
Supply Chain Analytics
HBSIS
Blumenau, Brasil
e-mail: fabricio.bezerra@hbsis.com.br

Pedro L. P. Corrêa
Departamento de Computação e Sistemas Digitais
Universidade de São Paulo
São Paulo, Brasil
e-mail: pedro.correa@usp.br

Resumo - Acidentes de trabalho é um tema relevante, principalmente quando temos uma realidade no Brasil onde ocorreram mais de meio milhão de ocorrências no ano de 2017. Na empresa onde este projeto foi realizado da área de Alimentação e Bebidas, não diferente do cenário nacional, é abordado como um tema crítico, necessitando-se de uma abordagem focada para resolver este problema. Este trabalho apresenta um caso de estudo de um planejamento de 10 anos para diminuição nos acidentes de trabalho, concluindo-se na entrega de um algoritmo de predição de acidentes. O objetivo deste trabalho foi desenvolver um modelo de predição e analisar os resultados de cinco diferentes algoritmos para indicar a probabilidade de possíveis acidentes nas cervejarias, sendo possível tomar ações preventivas. Como escopo de análise, foram abordados os algoritmos, *Random Forest*, *Regressão Logística*, *Classificador XGB*, *Linear SVC*, *SVC*. O algoritmo que obteve a melhor desempenho, utilizando a métrica *ROC_AUC* como determinante, foi o *Support Vector Clustering (SVC)* com 73,75%. Com este, o modelo realiza a predição com aproximadamente 62% de precisão e 74% de recall. O principal diferencial deste trabalho é a publicação do código fonte como software livre na plataforma GitHub para aplicação em outras empresas de manufatura, bens de consumo, entre outras.

Palavras-Chave; código fonte livre; aprendizado de máquina; análise de desempenho; acidentes de trabalho.

I. INTRODUÇÃO

Para Cervejaria Ambev, o principal ativo são as pessoas! Então, como deixá-las mais seguras dentro do nosso ambiente de trabalho? A receita foi encontrada baseada em três ingredientes fundamentais: mentalidade, processo e tecnologia. A mentalidade de segurança no trabalho é um bem negociável, prioritário, essencial e não hierárquico! O processo, é a chave para aplicarmos essa mentalidade e implantarmos a tecnologia em todos os níveis das nossas operações, cervejarias, centros de distribuição e nas ruas com nossos vendedores. Por último, mas não menos importante, a tecnologia. Percebemos que por mais refinado que o processo estava, os resultados não eram suficientes. Logo, iniciativas foram realizadas para inclusão de tecnologia no dia a dia dos colaboradores.

Dados do ministério do governo indicam que, apenas para o ano de 2017, 549.405 acidentes de trabalho foram registrados em todo o Brasil. Apesar que os números indicam uma queda de aproximadamente 7% em relação ao ano anterior, ainda assim foram computados aproximadamente 2.900 casos de mortes originadas por acidentes em ambientes de trabalho [1].

A primeira iniciativa para mudar este curso dentro da empresa estudada, foi mudar a mentalidade. Atualmente há diversos avisos e alertas tanto na área administrativa, quanto na área fabril, como pode ser visto na Figura 1. Nestes, há informações sobre segurança e dicas de como proceder em certas situações cotidianas. As regras aplicadas são rígidas e servem para todos os que frequentam os ambientes: colaboradores, visitantes ou terceiros, todos estão sob a mesma regra.

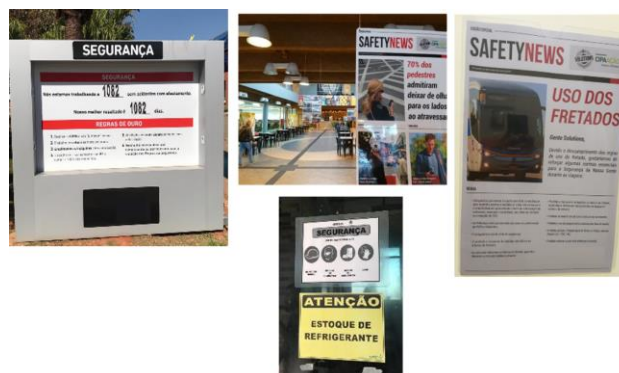


Figura 1. Avisos cotidianos sobre segurança de trabalho. Fonte: Autoria própria.

Com a vontade gerada de incentivar aplicações tecnológicas, foi-se instaurado um aplicativo (Figura 2) para que todos os colaboradores pudessem, de maneira fácil e rápida, identificar e informar a equipe de segurança local, qualquer problema relacionado a condições inseguras ou atos inseguros. Condições inseguras remetem para situações onde, por exemplo, existe uma goteira causando risco em piso molhado. Já atos inseguros, são situações onde, por exemplo, o colaborador está andando e digitando no celular, que é proibido em qualquer ambiente e situação.



Figura 2. Tela inicial e opções do aplicativo de segurança. Fonte: Autoria própria.

Este aplicativo foi a primeira iniciativa de aplicar a tecnologia para auxiliar na diminuição do número de casos. Com este, foi possível ter informações padronizadas e com qualidade, para que fosse possível juntando dados de outros sistemas, realizar análises mais profundas. Contudo, apesar de estas informações já possibilitarem a visibilidade da situação, buscou-se utilizando dados/tecnologia como fonte, realizar alterações dos processos e resultar em predição, não somente correção.

Existe uma padronização para classificação de acidentes: quando houve atendimento médico local e o colaborador voltou ao trabalho; quando houve atendimento médico local e o colaborador foi dispensado do dia; quando houve atendimento médico local, mas o colaborador foi levado para atendimento em hospital.

O próximo passo foi o desenvolvimento de um algoritmo que utiliza registros de atos e condições inseguras, nível de demanda, jornadas de trabalho e mais de 180 variáveis com as quais conseguimos prever acidentes em nossas unidades, alertando os times de segurança local para que tomem as ações necessárias buscando um ambiente mais seguro.

O algoritmo é dinâmico, atualizado diariamente, aprendendo novos cenários de classificação para realizar o principal sonho: zero acidentes. Essa mentalidade de seguir o guia de segurança não permite atalhos e garante que todos estejam em ordem com suas rotinas. Nesse trabalho iremos apresentar o algoritmo, resultados e como o usuário interage com essa solução.

O algoritmo já está em seu quarto ano em ambiente produtivo e esse ano, seu código fonte foi aberto para toda a comunidade de ciência de dados, tecnologia da informação e segurança do trabalho (Open Source).

Este trabalho busca explicar todo o desenvolvimento de um algoritmo para prever as chances de acidentes dentro do ambiente fabril e avaliar diferentes algoritmos de aprendizado de máquina.

Através deste trabalho buscamos disponibilizar uma ferramenta complementar ao ambiente de trabalho melhorando a segurança, através do desenvolvimento de um algoritmo que gera saída por unidade de negócio (neste caso cervejaria), a probabilidade de um acidente acontecer em breve.

Esse tipo de projeto não é uma vantagem competitiva, mas uma ação para deixar o mundo um lugar mais seguro para se trabalhar.

II. DESENVOLVIMENTO

Este projeto foi apoiado por uma visão de longo-prazo desta companhia (planejamento de dez anos), onde por direcionamento, todos os acidentes são devidamente cadastrados com a maior granularidade possível, viabilizando conforme os anos, possibilidades de aplicação de algoritmos e análises de dados. Além disso, esta visão de longo prazo também abordou preocupações questões técnicas de captura, organização, visualização e plano de ação, possibilitando garantir um nível de maturidade sólido para realizar a predição.

Segundo [2] a ciência de dados, como por exemplo a predição de acidentes, não é possível ser aplicada sem antes ter uma revisão e adequação do processo e dados suficientes que reproduzam o ambiente real. Por isso no planejamento de 10 anos mencionado, o primeiro passo realizado foi a alteração da mentalidade no momento do registro dos acidentes de trabalho e posteriormente, a agregação dos dados apontados pelo aplicativo de segurança, para que por último, os modelos de predição fossem possíveis.

A. Dados

O projeto foi iniciado mediante diversos debates para entender qual seria direcionamento dos trabalhos, pois a questão de segurança dentro de um ambiente fabril poderia ser norteadada por diversas frentes e metodologias. Foram selecionadas então cinco fontes principais de dados para o algoritmo (Figura 3): informações dos colaboradores, estrutura da cervejaria com seu volume de produção, dados meteorológicos do local da unidade, acidentes já ocorridos e os reportes apontados no aplicativo já mencionado.



Figura 3. Fonte de informações para os algoritmos. Fonte: Autoria própria.

Um ponto de atenção que foi tomado durante todo o desenvolvimento, é que as bases não contêm nomes ou nada que relacione um funcionário com as informações nominais. Todas tabelas e informações são anônimas.

Este procedimento foi tomado para já adequar este algoritmo aos novos temas de privacidade de dados, seguindo a Lei Geral de Proteção de Dados (LGPD) - lei nº 13.709, de 14 de agosto de 2018.

O trabalho de [3] aborda justamente o tema dos dados com a utilização de algoritmos. O autor cita que qualquer solução no âmbito de análise de dados, ou desenvolvimento de algoritmos, a empresa deve assumir responsabilidade de acompanhar inclusive a capacidade de autoconstrução destes e conforme os

algoritmos/inteligência artificial evoluem. Por isso, o autor também destaca, que é crucial que desde o começo de projetos utilizando de tais tecnologias e durante toda sua concepção e evolução, deve-se sempre ter como coluna central a questão da privacidade dos dados.

Buscando adequação justamente a esta nova lei sobre a privacidade dos dados, todas tabelas e informações são anônimas. Alguns exemplos de informações que foram utilizadas:

- Informações de Pessoas
 - Ano;
 - Faixa Salarial;
 - Tempo de companhia;
 - Estado Civil;
 - Pacote de Benefícios;
- Características Fabris
 - Quantidade de paradas das linhas de produção;
 - Turnover;
 - Quantidade de horas extras;
 - Engajamento;
- Ocorrências
 - Acidentes de trabalhos;
 - Incidentes durante o período de trabalho;
 - Ações positivas;
 - Análises Descritivas;
- Dados de Ambiente
 - Temperatura;
 - Precipitação;
 - Dia da semana;
 - Sazonalidade;
- Dados de Aplicativo de Segurança
 - Atos inseguros;
 - Condições inseguras.

Conforme mencionado, o ponto de partida para análise foi entender o lado pessoal dos colaboradores. Idealmente pretende-se entender os aspectos que mais aumentam um risco de acidente.

No total, iniciamos a análise com mais de 180 variáveis. Portanto, após conseguir reunir as informações, buscou-se iniciar o entendimento destas, focando na busca mais qualitativa do que quantitativa.

B. Análise Exploratória dos Dados

Para o desenvolvimento deste trabalho foram aplicadas ferramentas computacionais utilizando bibliotecas de acesso livre. Abaixo seguem descritas as ferramentas, versões e bibliotecas utilizadas:

- 1) Ferramentas para programação:
 - Windows 10 – Versão 1803 ou mais recente.
 - Jupyter notebook – Versão 5.7.8 ou mais recente
 - Python – Versão 3.6 ou mais recente
- 2) Bibliotecas:
 - Pandas – Versão 0.24.2 ou mais recente.

Biblioteca que oferece uma estrutura de dados fácil de ser utilizada e de alta performance, além de diversas ferramentas para análises.

- *NumPy* – Versão 1.16.3 ou mais recente.
Pacote fundamental para manipulação de dados.
- *Matplotlib* – Versão 3.1.0 ou mais recente.
Biblioteca 2D em Python para facilitar plotar gráficos.
- *XGBoost* – Versão 0.9.0 ou mais recente.
Biblioteca que contempla modelos de classificação baseados no método *GradientNBoosting*.
- *MLxtend* – Versão 0.16.0 ou mais recente.
Biblioteca amplamente utilizada para normalizar dados.
- *Scikit-learn* – Versão 0.21.2 ou mais recente.
Biblioteca que oferece ferramentas e métodos simples e fácil para data mining e análise de dados.

Por meio destas bibliotecas, elaboramos uma análise exploratória dos dados. Esta é uma metodologia para iniciar aplicações de modelos de ciência de dados. O objetivo principal desta etapa está, além de entender a viabilidade e gerar insights, poder entender correlações e grandezas entre as variáveis [4].

Neste trabalho iniciamos a análise observando registros totais e distribuição de acidentes por ano. Um importante passo foi analisar a relação entre registros com e sem acidentes. Esta visão foi importante, pois indicou o claro desbalanceamento da variável dependente.

Esta é uma questão recorrente em por exemplo, casos de análise de fraude de cartões de crédito ou estudo para classificação de doenças raras, onde a quantidade de amostras é maior para uma classe do que em relação a outra. Deste modo, os cenários apresentados para o modelo de aprendizado de máquina, ficará enviesado de forma que ele tenderá a prever com maior probabilidade a classe majoritária. Conforme pode ser visto na Figura 4, o número de registros com acidente é apenas 4% do total.

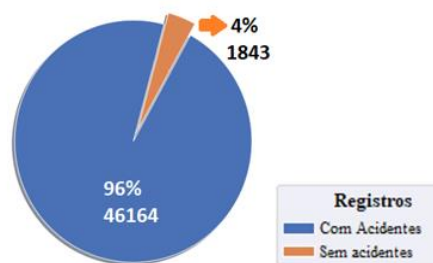


Figura 4. Comparação de registros com registros com e sem acidentes.
Fonte: Autoria própria.

Após esta visão inicial dos dados e suas distribuições, entramos na próxima etapa da análise que visa tentar balancear as variáveis e entender mais a fundo estatisticamente suas relações com a predição objetivo de prever acidentes de trabalho.

C. Pré-Processamento

A etapa de pré-processamento é uma das mais importantes dentro de todo o escopo para aplicação de algoritmos de aprendizado de máquina [5]. Utilizamos este momento para realizar algumas ações, tais quais:

tratar dados corrompidos ou com valores distintos, descartar atributos irrelevantes e alterar estrutura dos dados.

Segundo [6] esta etapa pode consumir até 90% de todo o tempo e esforço necessários para desenvolvimento de uma inteligência artificial. Principalmente pois os dados afetam diretamente a qualidade das previsões e variáveis utilizadas para atingir a variável resposta.

Para realizar estas e outras ações, realizamos algumas análises:

1) Limpeza dos dados

Nessa etapa, procurou-se entender possíveis variáveis categóricas e eliminar registros contendo variáveis com valores em branco ou nulos.

A Figura 5 apresenta a porcentagem de registros observados que estavam nulos em algumas variáveis analisadas.

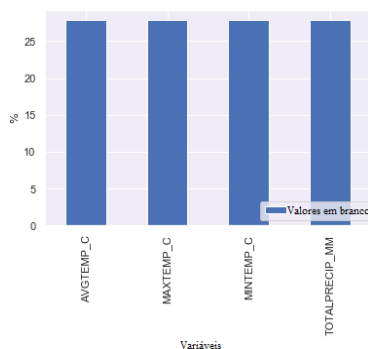


Figura 5. Variáveis analisadas e a porcentagem de registros em branco. Fonte: Autoria própria.

Onde:

- AVGTEMP_C = Média da temperatura em determinada cidade;
- MINTEMP_C = Valor mínimo da temperatura em determinada cidade;
- MAXTEMP_C = Valor máximo da temperatura em determinada cidade;
- TOTALPRECIP_MM = Valor total do volume de precipitação em uma determinada cidade.

2) Normalização dos dados

Esta etapa consiste em aplicar uma escala padrão nas variáveis, de modo que estas apresentem valores entre 0 e 1. Deste modo, deixando os dados nesta escala, os modelos não são impactados de acordo com a magnitude dos valores. Usamos esse método como medida preventiva para evitar dados enviesados, pois a ideia consiste em analisar diversos modelos para o mesmo conjunto de dados [7].

3) Separação treino e teste

Neste trabalho utilizamos 20% para base de treino e 80% para teste. Conforme mencionado anteriormente sobre a questão do desbalanceamento dos dados, utilizamos um método de validação cruzada, que consiste em balancear o número das variáveis a cada iteração. Este será mais bem detalhado ao longo do trabalho. Deste modo, minimiza-se as chances de algum dado importante para a classificação ser deixado de fora durante o treinamento [7].

4) Seleção de variáveis

É nesta etapa que são determinadas as variáveis que possuem correlação com a variável dependente, ou seja, neste ponto entendemos quais variáveis realmente importam para a predição de acidentes. Contudo, as variáveis não devem ser correlacionadas entre si, logo, independentes [7].

Para isso realizamos estas seguintes etapas:

a) *Remoção de variáveis constantes*: Um importante passo para a seleção de variáveis consiste na análise dos seus valores e sua variância conforme a série temporal. Esta etapa consiste em analisar se o valor de uma variável altera ou não ao longo do tempo, determinando se ela pode ser considerada irrelevante para o modelo.

Nesta etapa foi removida 1 variável que não alterou seu valor de acordo com o tempo.

b) *Remoção de variáveis quase constantes*: Esta etapa consiste na possibilidade de determinar uma faixa de corte para aceitar sua remoção ou não.

No caso deste trabalho, removemos todas as variáveis que não houve variância significativa nos dados maior que 98% do tempo, ou seja, se em 98% dos registros o valor da variável não alterou, ela foi descartada.

O valor adotado para este trabalho foi arbitrário, mas foi baseado em melhores práticas de acordo com outros algoritmos onde esta faixa de corte fica entre 96% - 99% [8]. Nesta etapa foram removidas 4 variáveis que não alteram seu valor de acordo com o tempo.

c) *Remoção de Variáveis Correlacionadas*: Nesta etapa foi realizada uma análise para determinar outro ponto muito importante quando se deseja aplicar algoritmos para previsões: as variáveis têm que ser altamente correlacionadas com a variável resposta, contudo, devem ser independentes entre si.

Com isso, objetiva-se utilizar apenas variáveis que impacta negativamente ou positivamente a predição de acidentes. Contudo, estes impactos não podem ser vistos entre as variáveis, indicado assim possíveis vies.

d) *Seleção Recursiva de variáveis*: Utilizamos o método de Random Forest para analisar a importância da variável na predição. Quanto maior o valor da taxa, maior sua importância e consequentemente esta variável deve estar no modelo. A Figura 6 mostra uma parte das variáveis analisadas.

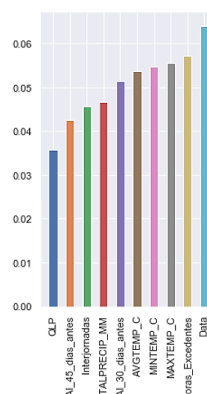


Figura 6. Exemplo da saída da análise da seleção recursiva. Fonte: Autoria própria.

Onde:

- QLP = Número de Colaboradores
- QTD_FAI_45_dias_antes = Número de registros registrados com acidentes sem afastamento retroativos 45 dias.
- Interjornadas = Tempo de descanso do colaborador entre jornadas de trabalho.
- QTD_FAI_30_dias_antes = Número de registros registrados com acidentes sem afastamento retroativos 30 dias.
- Horas_Excedentes = quantidade de horas excedentes realizadas.
- Data = dia e mês do ano (sazonalidade).

Nesta etapa foram removidas 35 variáveis que não atenderam o padrão adotado.

e) *Remoção de Variáveis Correlacionadas:* Neste trabalho adotamos a métrica de avaliação do índice Pearson de Correlação. Esta é uma escala amplamente utilizada na ciência dos dados. Este método foi desenvolvido por Karl Pearson [9] e foi o primeiro a desenvolver uma métrica capaz e representar correlação entre variáveis. O trabalho de [10] apresentou 13 maneiras distintas de se olhar correlações e destas, apenas o índice de Pearson e uma outra métrica apresentou resultados satisfatórios e estáveis de acordo com os diversos testes realizados.

Esta escala possui a faixa de valores entre -1 até 1. Sendo em seu menor índice, esta métrica indica variáveis que possuem correlação negativa com a variável resposta. Portanto quanto mais perto de 1 as variáveis esta positivamente correlacionadas.

Para este caso, adotamos variáveis que possuem valor abaixo que 0,8 nesta escala.

f) *Remoção de variáveis através da métrica ROC_AUC:* Também analisamos a importância das variáveis passando pelo filtro de ROC_AUC. Neste método descartamos variáveis que não apresentaram grau maior do que 0,5. Contudo após todas as análises anteriores realizadas, não foi encontrado nenhuma variável que seria necessário ser retirada [11].

g) *Output da etapa de pré Processamento:* Após todas as etapas de pré-processamento descritas acima, eliminamos diversas variáveis que, juntamente com o time de negócios, não impactavam o modelo ou durante as análises mais profundas fomos as retirando. Na Figura 7 temos uma ilustração que mostra a quantidade de variáveis por etapa.

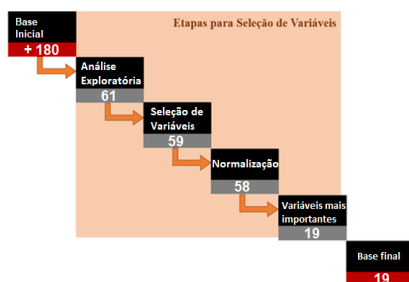


Figura 7. Ilustração da quantidade de variáveis restantes ao longo das análises das variáveis. Fonte: Autoria própria.

III. RESULTADOS

Para determinar o melhor modelo de acordo com o conjunto de dados neste trabalho, elaboramos testes com os algoritmos: *Random Forest*, Regressão Logística, Classificador XGB, *Linear SVC* e *SVC*

A. Método de Validação / Teste

Utilizamos o método de validação cruzada (*cross validation*) denominada *K-fold* para validação dos algoritmos. Este método foi utilizado dado a natureza desbalanceada dos dados e que, automaticamente este método cria inicialmente k partições (*folds*), mais próximas possíveis em número e quantidade de registros contendo as variáveis a serem avaliadas. Conforme as interações de treino e validações (estimativas das predições) forem ocorrendo, o próprio método cria a partir da base em momento k-1, uma nova quebra para continuar o treino e validação [12]. Assim ocorrendo até o número de interações determinado. A predição é realizada e validada a partir do momento k-1. A Figura 10 ilustra o funcionamento das interações e partições.

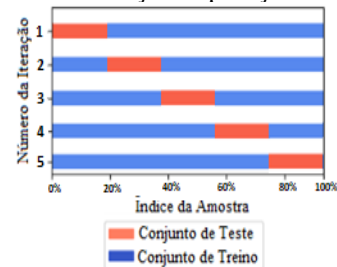


Figura 8. Ilustração de como o método de validação cruzada utilizando o método *K-Fold*. Fonte: Adaptada de [12].

O processo para validação cruzada foi realizado 5 vezes para cada modelo testado. Este número de interações foi escolhido de acordo com os testes realizados. Na literatura, encontram-se discussões sobre o número ideal de interações, como por exemplo em [7] onde ele indica que o número ideal seria k=10, mesmo valor indicado pelo trabalho de [13], que indica este valor mesmo que o poder computacional permita realizar maiores valores. O trabalho de [14] aponta que 5 ou 10 iterações são as melhores possibilidades. Contudo, de acordo com os testes realizados na nossa realidade dos conjuntos de dados, percebemos que o número ideal foi k = 5 interações.

A Figura 9 apresenta os resultados em forma de visualização dos resultados em momento k = 5.

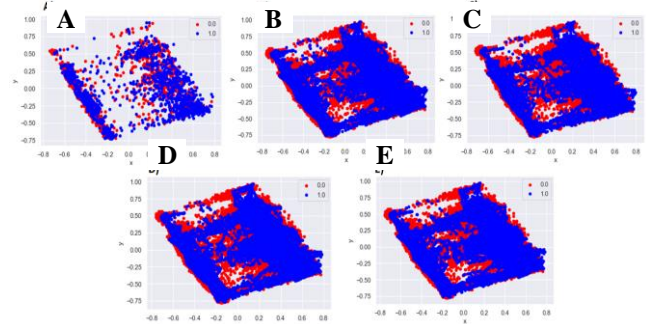


Figura 9. Resultados para K=1. Onde: A) *Random Forest*; B) Regressão Logística; C) Classificador XGB; D) *Linear SVC*; E) *SVC*. Fonte: Autoria própria.

No final, foi obtido a média dos resultados de cada iteração do método. Os resultados podem ser vistos na seção C deste capítulo.

B. Métricas para Validação

Devido à natureza do desbalanceamento dos dados, utilizamos outras formas de métricas para chegar uma conclusão ideal. Seguem abaixo as métricas utilizadas juntamente com a representação gráfica dos resultados.

1) *Acurácia*: é o valor da diferença entre uma medida e seu valor real [15]. O termo foi inicialmente cunhado por Mikhail e Ackermann, que também, durante seu trabalho, incluí que a precisão neste contexto, o quão regular esta medida com relação à sua média [16].

Na Figura 10 temos os resultados obtidos nos testes com os diferentes modelos.

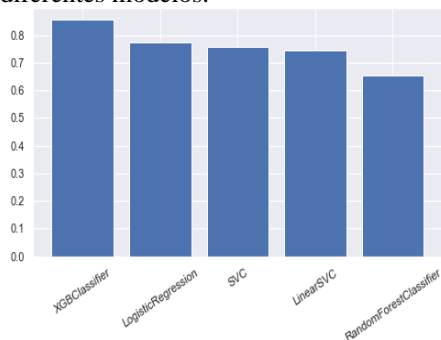


Figura 10. Acurácia por algoritmo. Fonte: Autoria própria.

Onde:

- *XGBClassifier*= Algoritmo *XGB*;
- *LogisticRegression*= Algoritmo Regressão Logística;
- *SVC* = Algoritmo *SVC*;
- *LinearSVC* = Algoritmo *Linear SVC*;
- *RandomForestClassifier*= Algoritmo *Random Forest*.

2) *Precisão*: Acurácia e precisão são termos que na cultura popular, podem se sobrepor. Contudo, ambas possuem significados próprios e especialmente quando aplicados em modelos estatísticos, são cruciais para avaliação.

É possível observar que a acurácia está relacionada com as medidas de valores da sua média, informando se a medida deste ponto médio está correta ou não. Já a precisão está relacionada com o quanto estas medidas variam [16].

Na Figura 11 temos os resultados obtidos nos testes com os diferentes modelos. O melhor valor é 1 e o pior valor é 0.

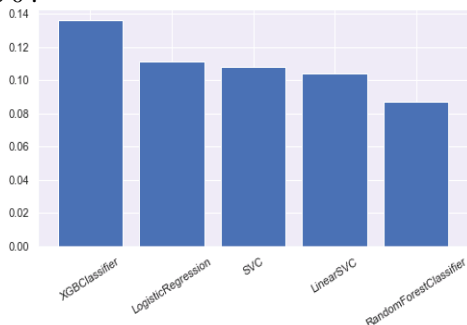


Figura 11. Precisão por algoritmo. Fonte: Autoria própria.

3) *Recall*: é a capacidade do classificador de classificar corretamente amostras positivas dentro do ambiente de todas as amostras observadas. Abaixo temos os resultados observados nos testes dos cinco modelos. O melhor valor é 1 e o pior valor é 0 (Figura 13).

Na Figura 12 temos os resultados obtidos nos testes com os diferentes modelos.

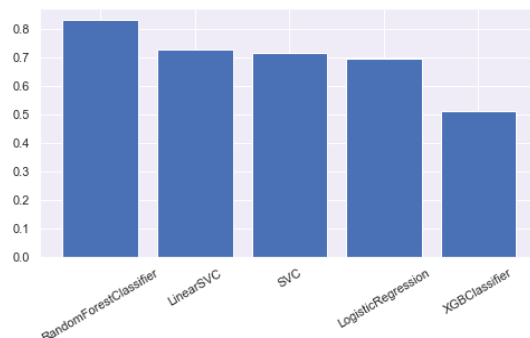


Figura 12. Recall por algoritmo. Fonte: Autoria própria.

4) *F1*: pode ser interpretado como uma média ponderada da precisão e do *Recall*, em que uma pontuação F1 atinge seu melhor valor em 1 e a pior pontuação em 0 (Figura 14). Abaixo temos os resultados observados nos testes dos cinco modelos.

Na Figura 13 temos os resultados obtidos nos testes com os diferentes modelos.

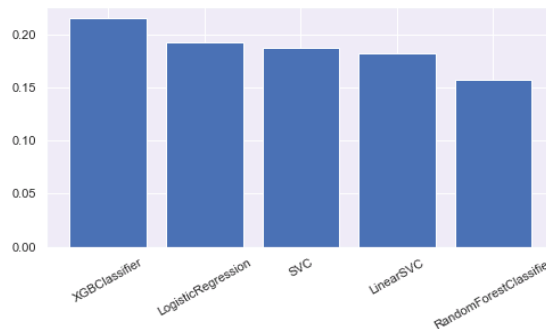


Figura 13. F1 por algoritmo. Fonte: Autoria própria.

5) *ROC AUC*: é originada a partir do Receiver Operating Characteristic (ROC).

O ROC é um modo estatístico que apresenta graficamente uma avaliação de pontos entre eixos de verdadeiros positivos e falsos positivos de modelos desenvolvidos a partir de uma massa de dados [4].

Na Figura 14, é mostrado um exemplo de um gráfico ROC com cinco pontos aleatórios representando cinco modelos de classificação diferentes (A, B, C, D e E).

Neste exemplo demonstrado, A representa um comportamento mais conservativo e D é o mais liberal. Além disso, é possível perceber que um ponto no espaço ROC é melhor que outro se e somente se ele está acima e à esquerda do outro ponto, pois representa que tem uma maior taxa de verdadeiros positivos e uma menor taxa de falsos positivos [17].

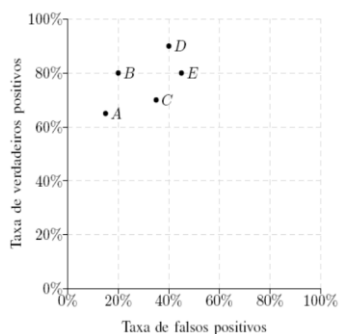


Figura 14. Exemplo de cinco pontos dentro da curva ROC. Fonte: Adaptado de [17].

A Área sob a curva, ou em inglês, *Area under Curve* (AUC), executa uma curva a partir de todos pontos dentro da curva ROC. Logo, a curva mostra todos os pontos possíveis deste os mais restritivos até os mais liberais. Utilizando o exemplo da figura anterior, a curva total ROC seria:

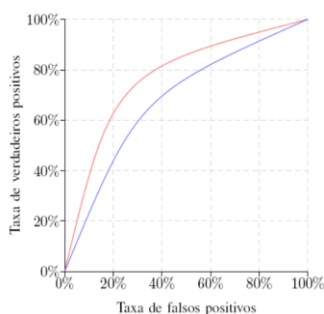


Figura 15. Exemplo de cinco pontos dentro da curva ROC. Fonte: Adaptado de [17].

A partir deste gráfico, foi-se desenvolvida a métrica ROC_AUC para representar, em uma escala de 0 até 1, os valores desta área [17].

Na Figura 16 temos os resultados obtidos nos testes com os diferentes modelos. Valores próximos de 1 são melhores, enquanto os mais próximos de 0 são os piores.

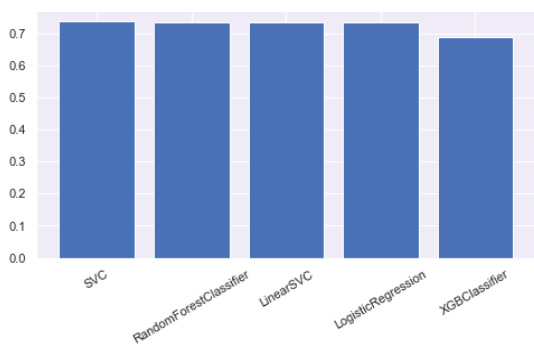


Figura 16. ROC_AUC por algoritmo. Fonte: Autoria própria.

A métrica utilizada como principal fonte de avaliação dos algoritmos testados neste trabalho é a ROC AUC, dado que cada vez mais, diversos trabalhos e análise são feitas utilizando esta métrica dentro do ambiente de mineração e análise de dados [18].

6) *MCC: Matthews Correlation Coeficient* (MCC) é, em essência, um valor de coeficiente de correlação que um coeficiente de +1 representa uma previsão perfeita, 0

uma previsão aleatória média e -1 uma previsão inversa [19].

O MCC considera os casos verdadeiros e falsos positivos e negativos, se tornando um indicador balanceado apresentado um equilíbrio entre performance, comportamento e coerência dentro de situações em que os dados representam situações com diversas classes, desbalanceados e randômicos

Na Figura 17 temos os resultados obtidos nos testes com os diferentes modelos [20].

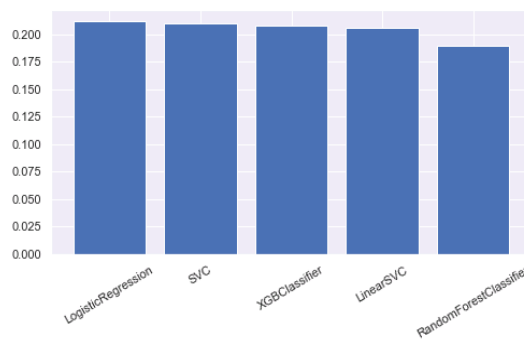


Figura 17. MCC por algoritmo. Fonte: Autoria própria.

C. Análise dos Resultados

Após todas as estimativas das métricas realizadas, obtivemos os seguintes valores apresentados na Tabela 1.

TABELA I. TABELA COM AS MÉDIAS DAS MÉTRICAS ANALISADAS APÓS OS TESTES DOS ALGORITMOS NA VALIDAÇÃO CRUZADA

model	accuracy	precision	recall	f1	roc_auc	mcc
SVC	0.757902	0.107929	0.715333	0.187553	0.737483	0.210085
RandomForestClassifier	0.652451	0.086792	0.828667	0.157105	0.736978	0.189951
LinearSVC	0.746186	0.104337	0.724667	0.182400	0.735863	0.205561
LogisticRegression	0.772145	0.111441	0.692667	0.191981	0.734021	0.211506
XGBClassifier	0.853955	0.136014	0.511333	0.214814	0.689607	0.207564

É possível analisar a partir das informações da tabela que a Regressão Logística apresentou o melhor balanço entre Precisão e Recall, o que é desejado conforme foi explicado no capítulo 2.

Conforme mencionado anteriormente, a métrica ROC AUC foi a considerada como principal variável para decisão da escolha do melhor modelo de acordo com a realidade dos dados apresentada.

Portanto, como pode ser visto na tabela acima, o modelo que obteve o maior valor desta métrica foi o SVC. Posteriormente a escolha do modelo, foi realizado um trabalho de validação deste modelo, quando este foi apresentado para a base de treino.

D. Validação do Modelo Selecionado

O objetivo desta etapa é gerar uma série de gráficos, análises e ajustes para garantir a validade no modelo aplicado na base histórica e posteriormente, para predição. Esta etapa permite identificar caso alguma variável que não se ajustaram bem ao modelo, ou identificar variáveis mais importantes que influenciam diretamente o resultado da predição.

O primeiro foi estudar o resultado matriz de confusão do modelo SVC aplicado a nossa base de teste. O resultado pode ser visto na Figura 18:

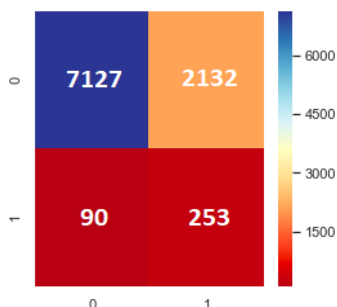


Figura 18. Figura com resultados das predições. Fonte: Autoria própria.

Podemos observar que temos 2.132 casos em que a predição apontou que teria um risco de acidente alto e, segundo as informações históricas, não ocorreu de fato. Indicando possível persistência de *overfitting*.

O segundo teste foi ver avaliar a curva ROC. O resultado pode ser visto na Figura 19:

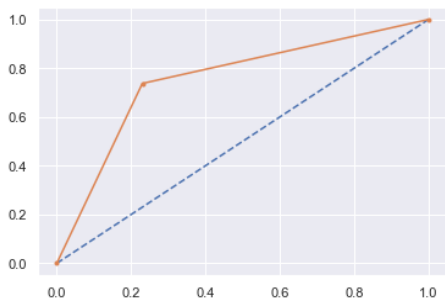


Figura 19. Curva ROC modelo SVC. Fonte: Autoria própria.

O terceiro teste foi analisar as curvas de precisão e recall do modelo. O resultado desta análise pode ser visto na Figura 20:

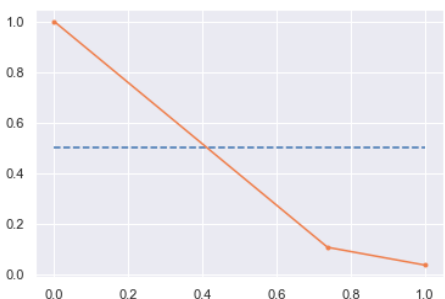


Figura 20. Precisão e Recall modelo SVC. Fonte: Autoria própria.

A Figura 21 apresenta o relatório da classificação realizada. O objetivo é avaliar principalmente os valores de *recall* e precisão entre as classificações maioritárias (“0.0” = Registro sem acidente) em comparação com classificações minoritárias (“1.0” = Registros com acidente). Sendo que se busca um valor de *recall* alto para classe minoritária e precisão alta para classe majoritária.

Esta visão traduz que o algoritmo precisa entender e conseguir prever com uma alta taxa de confiabilidade acidentes, sendo estes são minoritários dentro do conjunto de dados.

	precision	recall	f1-score	support
0.0	0.99	0.77	0.87	9259
1.0	0.11	0.74	0.19	343
micro avg	0.77	0.77	0.77	9602
macro avg	0.55	0.75	0.53	9602
weighted avg	0.96	0.77	0.84	9602
accuracy:	0.76859			
precision:	0.10608			
recall:	0.73761			
f1:	0.18548			
auc:	0.75367			
mcc:	0.21793			

Figura 21. Resultado das métricas utilizando modelo SVC. Fonte: Autoria própria.

E. Resultados da predição

O modelo escolhido após toda a análise no item anterior foi através do modelo de algoritmo (SVC). Este método apresentou melhor desempenho dentro das métricas analisadas. Os resultados da predição podem ser vistos na tabela a seguir:

TABELA II. TABELA CONTENDO A TAXA DE RISCO PREVISTA PELO ALGORITMO.

Data	Unidade	Scored Probabilities
2018-11-09	Brewery EE	71.69%
2018-11-09	Brewery F	45.66%
2018-11-09	Brewery MM	34.54%
2018-11-09	Brewery HH	21.07%
2018-11-09	Brewery NN	19.14%

Onde:

- Data = Dia da predição;
- Unidade = Cervejaria;
- Scored Probabilities = Risco de Acidente (predição).

F. Interação com o usuário

Os valores previstos alimentam um *dashboard* (Figura 22). Este serve de alertar para o campo e auxiliar a tomada de decisão. Neste contém as seguintes informações:

- Mapa da taxa de Riscos de acidente por cidade;
- Variáveis que mais contribuem para o modelo;
- Histórico de registros com acidentes;
- Riscos em um gráfico de série temporal.

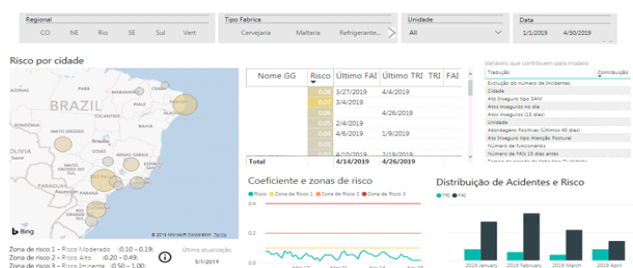


Figura 22. *Dashboard* analítico desenvolvido para dar visibilidade das predições. Fonte: Autoria própria.

As previsões indicam a probabilidade de acidentes, possibilitando realizar uma classificação de:

- Risco Moderado: 0,10 – 0,19
- Risco Alto: 0,20 – 0,49
- Risco Iminente: 0,50 – 1,00

O objetivo desta classificação é, caso a probabilidade gerada for alta, ocorram medidas para evitar acidentes, o que será descrito na seção abaixo.

G. Ecosistema

Tão importante quanto a previsão, os próximos passos são uma grande parte do ecossistema. O que fazer com as informações e quais ações precisam ser tomadas são cruciais para serem aplicadas ao campo

O algoritmo tem por funcionalidade prever os riscos de acidentes e posteriormente, alimentar uma visualização de dados, para que assim, o time responsável consiga monitorar onde possíveis acidentes de trabalho ocorrerão. Logo, este sinal é enviado para a equipe de segurança local e assim, medidas podem ser realizadas para identificação do problema, para que seja possível realizar um plano de ação para prevenção desses acidentes.

A Figura 23 apresenta o croqui de funcionamento desse processo preditivo de segurança, evidenciando as principais bases de dados, fluxo de trabalho e possíveis ações do usuário foco.

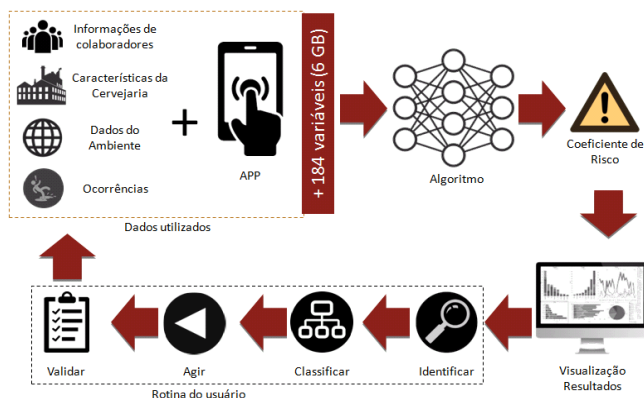


Figura 23. Ilustração geral do trabalho. Fonte: Autoria própria.

Após a visualização dos resultados do algoritmo como mencionado acima, o time de segurança do trabalho atua nas etapas apresentadas na Figura 24 com o intuito de prevenir que o acidente aconteça.

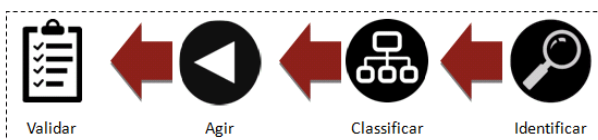


Figura 24. Etapas de atuação após análise dos resultados do algoritmo. Fonte: Autoria própria.

H. Impacto gerado na redução dos acidentes

A Figura 25 apresenta a redução percentual de acidentes de trabalho em cada ano se comparado ao ano anterior. Esses acidentes foram reportados ao longo de oito anos e tem como limitação de tema a área cervejeira. Tanto acidentes envolvendo colaboradores próprios, como

também terceirizados são tratados da mesma forma e são exigidas as mesmas regras de segurança.

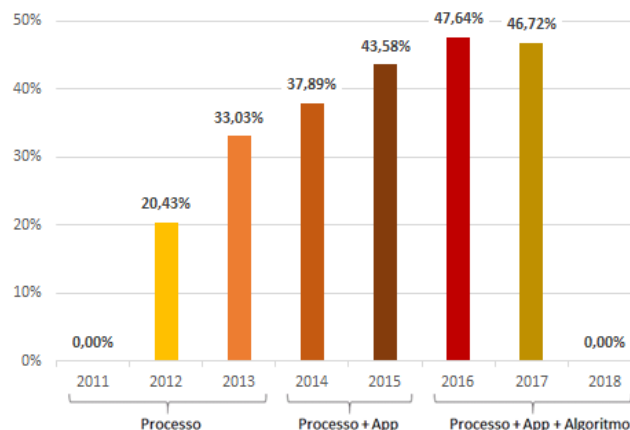


Figura 25. Redução de acidentes por ano (%). Fonte: Autoria própria.

IV. SOFTWARE LIVRE

Uma das principais características de disponibilizar um código aberto ao público, é a possibilidade de diferentes usuários indicarem melhorias. Ter a comunidade proativa, sugerindo alterações e mudanças no código, é algo extremamente visado para contínua evolução deste trabalho.

A Figura 26 mostra a primeira página da plataforma GitHub, onde versão inicial deste projeto pode ser encontrada em:

<https://github.com/cervejaria-ambev/safety>

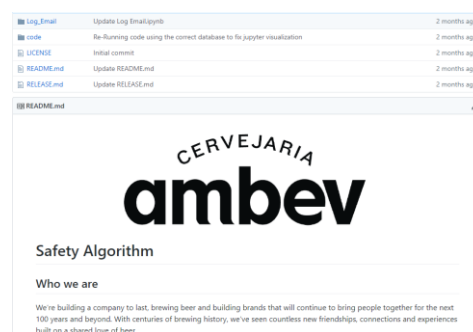


Figura 26. Página inicial do repositório criado contendo o código de acesso livre. Fonte: Autoria própria.

Nossa expectativa é que outras empresas de manufatura possam aplicar esses passos e desenvolverem seus próprios modelos de dados para melhorar a experiência de seus colaboradores no ambiente de trabalho.

Cada algoritmo responde diferentemente de acordo com o comportamento e distribuição dos dados ao longo do tempo. Dado que este modelo será disponibilizado ao público e, portanto, a novas bases de dados e para atender a possíveis alterações futuras em bases próprias, colocamos estes diferentes algoritmos para teste.

Os modelos utilizados são amplamente utilizados no mundo de ciência de dados. É importante saber que aplicações futuras em outras empresas ou setores, irão obter diferentes resultados.

V. CONCLUSÃO E PRÓXIMOS PASSOS

Foi discutido o desenvolvimento e análises realizadas para desenvolver um algoritmo para predição diária do risco de acidente em ambientes fabris, no caso deste trabalho, cervejarias. Além disso, foram demonstrados métodos para avaliação de cinco modelos aplicados em nossos dados.

O modelo escolhido após toda a análise foi o algoritmo SVC. Este apresentou melhor desempenho dentro da métrica ROC_AUC utilizada como determinante para tomada de decisão. Foi também realizada uma validação deste modelo escolhido na base de teste.

Após foi realizada a predição. O resultado da predição apontou que, para cervejaria aqui anonimamente denominada EE, esta será classificada como um alto risco de acidente, dado que a predição apontou um valor de 0,7169 ou 71,69%. Portanto, o próximo passo consiste em enviar esta predição com um alerta para o gerente de fábrica e equipe de segurança local, para assim realizar a identificação da causa raiz deste índice elevado e tomar providências desenvolvendo um plano de ação para mitigar possíveis situações ou condições inseguras.

Um passo adiante seria incluir dados de processos para este modelo. O objetivo seria incluir, como por exemplo, chances de uma garrafa estourar na linha de processo, durante a produção do produto. Por mais que atualmente não se tenha este indicador cadastrado como uma variável dentro de sistemas, outras variáveis podem me indicador uma probabilidade deste evento de estouro ocorrer, como pressões e temperaturas aplicados ao vidro.

Outro ponto de melhoria é que com a visibilidade das informações colhidas durante, por exemplo, a etapa de análise exploratória, possa ser suficiente para o processo possa ser alterado ou ter uma supervisão com maior vigor para inclusão de todos acidentes ou incidentes e com maior detalhamento possível. Uma questão sobre a variável resposta dos nossos dados é que esta é originada de inclusões manuais em sistema. Isso acarreta casos de, possivelmente omissões de registros ou detalhes de pequenos incidentes ou até de casos mais graves. Isto está diretamente ligado com a qualidade das predições realizadas, como por exemplo a quantidade de registros em branco.

REFERÊNCIAS

- [1] LORENA, Talita. . In: MINISTÉRIO DA ECONOMIA. Acidentes de trabalho apresentam queda de 6,2% em 2017. [S. l.], 28 set. 2018. Disponível em: <http://www.previdencia.gov.br/2018/09/acidentes-de-trabalho-apresentam-queda-de-62-em-2017/>. Acesso em: 22 mar. 2019.
- [2] PENN, Chris. . In: IBM. The evolution of the data-driven company. [S. l.], 24 fev. 2017. Disponível em: <https://www.ibm.com/blogs/business-analytics/evolution-data-driven-company/>. Acesso em: 11 fev. 2019.
- [3] JUNIOR, Sergio Ricardo Correia de Sá. A regulação jurídica da proteção de dados pessoais no Brasil. In: JUNIOR, Sergio Ricardo Correia de Sá. A regulação jurídica da proteção de dados pessoais no Brasil. Orientador: Eduardo Magrani. 2018. Monografia (Programa de Pós-Graduação) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2018. f. 47. Disponível em: <https://www.maxwell.vrac.puc-rio.br/37295/37295.PDF>. Acesso em: 19 ago. 2019.
- [4] BATANERO C.; ESTEPA A.; GODINO J.D. Análisis exploratorio de datos: sus posibilidades en la enseñanza secundaria. *Suma*, 9, 25-31. 1991.
- [5] Gustavo Enrique de Almeida Prado Alves. Pré-processamento de dados em aprendizado de máquina supervisionado. 2003. Monografia (Doutorado em Instituto de Ciências Matemáticas e de Computação) - USP. São Paulo.
- [6] CHAPMAN, Pete et al. CRISP-DM 1.0 Step-by-step data mining guide, USA: SPSS Inc. CRISPWP-0800, 2000.
- [7] WITTEN, Ian H. et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.
- [8] PÉREZ, Joaquín et al. A data preparation methodology in data mining applied to mortality population databases. In: *New Contributions in Information Systems and Technologies*. Springer, Cham, 2015. p. 1173-1182.
- [9] CASTRO, JORGE ALBERTO. VINHETA HISTÓRICA KARL PEARSON. SESQUICENTENÁRIO DE SEU NASCIMENTO. Portal de Periódicos Científicos - FURG, [S. l.], ano 2007, 2007. Disponível em: <https://periodicos.furg.br/vittalle/article/download/982/414>. Acesso em: 19 jul. 2019.
- [10] Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1), 59-66.
- [11] PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Curvas ROC para avaliação de classificadores. *The Institute of Electrical and Electronics Engineers (IEEE)*, [S. l.], 2008. Disponível em: http://conteudo.icmc.usp.br/pessoas/gbatista/files/ieee_la2008.pdf. Acesso em: 27 ago. 2019.
- [12] REFAEILZADEH, PAYAM; TANG, LEI; LIU, HUAN. Cross-Validation. *Cross-Validation*, Arizona State University, 2008. Disponível em: <http://leitang.net/papers/ency-cross-validation.pdf>. Acesso em: 22 jul. 2019.
- [13] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1137-1145, 1995.
- [14] Fushiki, T. (2009). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, 21(2), 137-146. doi:10.1007/s11222-009-9153-8
- [15] Menditto, A., Patriarca, M., & Magnusson, B. (2006). Understanding the meaning of accuracy, trueness and precision. *Accreditation and Quality Assurance*, 12(1), 45-47. doi:10.1007/s00769-006-0191-z.
- [16] Monico, Galera & Dal Poz, Aluir & Galo, Mauricio & Santos, Marcelo & Oliveira, Leonardo. (2009). Acurácia e precisão: revendo os conceitos de forma acurada / Accuracy and Precision: Reviewing the concepts by means of an accurate procedure. *Boletim de Ciências Geodesicas*. 15. 469-483. 10.13140/2.1.1815.8086.
- [17] PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Curvas ROC para avaliação de classificadores. *The Institute of Electrical and Electronics Engineers (IEEE)*, [S. l.], 2008. Disponível em: http://conteudo.icmc.usp.br/pessoas/gbatista/files/ieee_la2008.pdf. Acesso em: 27 ago. 2019.
- [18] Sá, Clebiano da Costa. Métodos de validação tradicional e temporal aplicados à avaliação de classificadores de RNAs codificantes e não codificantes. 2018. Dissertação (Mestrado em Bioinformática) - Bioinformática, Universidade de São Paulo, São Paulo, 2018. doi:10.11606/D.95.2018.tde-19052018-122805. Acesso em: 2019-09-04.
- [19] MATTHEWS, Brian W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, v. 405, n. 2, p. 442-451, 1975.
- [20] Jurman G, Riccadonna S, Furlanello C (2012) A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. *PLOS ONE* 7(8): e41882. <https://doi.org/10.1371/journal.pone.0041882>