

## **CATEGORIA 3**

# **APLICAÇÃO DE MODELOS PREDITIVOS DE APRENDIZADO DE MÁQUINA PARA ANTECIPAR FALHAS DE BITOLA ABERTA**

## **AUTOR**

*Marcelo Ferreira Soares*

*Engenheiro Civil, Mestre em Engenharia e Gestão de Processos e Sistemas e Analista Sênior do  
PCM de Via Permanente da VLI*

## **1. INTRODUÇÃO**

Um dos maiores gargalos para o crescimento do país, ao longo das últimas décadas, e ainda hoje, é a estrutura logística. Retomar o planejamento sistêmico de longo prazo do sistema logístico nacional é uma condição necessária para se viabilizar o crescimento continuado da economia brasileira.

Em países de dimensões continentais e com mercados internos e externos crescentes, como o Brasil, é razoável que o transporte no modal ferroviário tenha um papel importante em função da carga preponderantemente transportada (granéis sólidos, tais como, minérios, grãos e açúcar), vantagem competitiva quando são consideradas as distâncias e volumes a

## 26ª SEMANA DE TECNOLOGIA METROFERROVIÁRIA 7º PRÊMIO TECNOLOGIA E DESENVOLVIMENTO METROFERROVIÁRIOS



serem transportados, menor emissão de poluentes atmosféricos e nível de segurança operacional e pessoal próprio e terceiros.

A confiabilidade das ferrovias advém de inspeções realizadas dentro da periodicidade e qualidade requerida por normas brasileiras, internacionais, definições de estratégias de manutenção dos operadores logísticos e tendem a contribuir também para a regularidade à programação de entrega de cargas.

Portanto, as inspeções são a base fundamental para a identificação de anomalias em seus mais diversos tipos e gravidade, e conseqüente planejamento e programação da intervenção requerida.

Os modelos preditivos, por sua vez, indicando possíveis falhas de forma antecipada, propiciam às equipes de manutenção o tempo de resposta necessário, que é fundamental às ações de manutenção preventivas, intervindo nos ativos ferroviários de forma programada, de forma compatível com volume do tráfego, reduzindo impactos operacionais.

Este uso contínuo dos ativos, conforme citado por Hunter Harrison em sua estratégia de negócio conhecida por PSR (Precision Scheduled Railroading), baseia-se na eliminação de ineficiências e manutenção de vagões e locomotivas em movimento.

Tal estratégia também pode ser aplicada à via permanente, visto que, trata-se de um ativo expressivo neste contexto e, conforme Hunter Harrison, "Se um ativo não é usado, passa a ser um passivo, devido aos custos associados à sua posse. As ferrovias só ganham dinheiro quando as composições estão em movimento" (BARROW, 2019, p. 24, tradução do autor).

Neste sentido, as técnicas de aprendizado de máquina têm se mostrado eficientes em estudos e aplicações práticas objetivando antecipações de falhas.

Um bom exemplo advém do trabalho desenvolvido por Yella, Dougherty e Gupta (2008) onde a identificação das condições dos dormentes de madeira (bom ou inservível) na Suécia através de imagens digitais tem obtido uma acurácia em torno de 90% com a combinação de perceptrons multicamadas e SVM.

## **2. DEFEITOS GEOMÉTRICOS – BITOLA ABERTA**

Segundo Rives et al. (1977, p. 29, tradução do autor), “ferrovia é um sistema de transporte terrestre no qual os veículos se apoiam sobre uma superfície de rolamento por intermédio de elementos rotativos metálicos”.

A NBR 5462 (ABNT, 1994) define manutenção, na norma ABNT NBR 5462:1994, como sendo a combinação de todas as ações técnicas e administrativas, incluindo as de supervisão, destinadas a manter ou recolocar um item em um estado no qual possa desempenhar uma função requerida.

Quão mais intenso e frequente forem os esforços verticais e horizontais provenientes do tráfego ferroviário sobre a via, seja no peso ou carga imposta em cada eixo de cada um dos veículos ferroviários, seja pelo fluxo de cargas sobre a via, e quão mais agressivo for o clima no qual estiver inserida tal via, maior será seu ritmo de degradação, sujeitando-a a defeitos e, conseqüentemente, falhas.

Portanto, é primordial que sejam conhecidos os defeitos, onde os mesmos ocorrem com maior frequência, como evoluem e consequências prováveis se não forem tratados a tempo. As inspeções, sejam mecanizadas, sejam visuais (inclusive a pé sobre a via), são fundamentais para a identificação e mensuração destes defeitos.

Ildfonso (2013) classifica os defeitos em duas grandes categorias, ou seja, estruturais ou geométricos, sendo que os estruturais estão relacionados aos materiais que constituem a superestrutura ferroviária, e os geométricos à posição espacial da via. Neste artigo será dada ênfase aos defeitos geométricos.

## **2.1. Defeitos geométricos**

A geometria da via permanente é sensível, visto que anomalias nos mais diversos componentes, seja da superestrutura, seja da infraestrutura, normalmente afetam diretamente a condição da grade ferroviária (SCHRAMM, 1977) e, conseqüentemente da geometria da via.

Castello Branco et al. (2002) ratifica esta preocupação citando que é natural que os defeitos evoluam de maneira descompassada, muito em função das vidas úteis distintas de cada componente, sendo a geometria da via o componente com menor vida útil.

Alguns autores demonstram atenção especial com a geometria, como por exemplo, Brina (1979), quando define que o objetivo da conservação da via é garantir que seu traçado, através

do perfeito nivelamento, alinhamento e superelevação, se mantenha conforme projetado, de tal forma a que não venha a impactar negativamente o tráfego ferroviário.

Ildefonso (2013) subdivide os defeitos geométricos em:

- a. defeitos de bitola;
- b. defeitos de nivelamento;
- c. defeitos de alinhamento;
- d. empeno (ou torção).

Este artigo está direcionado aos defeitos de bitola, em especial bitola aberta, portanto este tópico será detalhado a seguir.

## **2.2. Defeitos de bitola**

Nabais (2014, p.26) define bitola como “o comprimento do segmento retilíneo ortogonal aos trilhos, paralelo ao plano de rolamento da via, cujas extremidades tocam as faces internas dos boletos e cujo afastamento desse segmento em relação ao plano de rolamento é de 15,88 mm”, onde “o plano de rolamento é aquele que tangencia simultaneamente a superfície superior do boleto de cada trilho”.

Hungria (2017) cita que em função da inscrição nas curvas dos rodeiros dos vagões ferroviários, há a necessidade que haja conicidade nas rodas, objetivando-se reduzir o arraste da roda sobre o trilho externo.

## 26ª SEMANA DE TECNOLOGIA METROFERROVIÁRIA 7º PRÊMIO TECNOLOGIA E DESENVOLVIMENTO METROFERROVIÁRIOS



Porém esta conicidade traz um deslocamento lateral do rodeiro em trechos de tangente, provocando um movimento conhecido por “hunting”, onde as rodas do vagão, segundo Steffler (2013), “caçam” o trilho.

A situação ideal é que a roda fique em contato com o trilho em até metade da largura do boleto, pois desta forma, tanto as tensões nas rodas, quanto nos trilhos, não ultrapassam a tensão de escoamento do aço que as compõe (HUNGRIA, 2017).

Quando há uma condição de bitola aberta, esta situação citada anteriormente, não acontece, gerando, em um primeiro momento, apenas escoamentos nas rodas e trilhos. Porém, com bitolas abertas acima dos limites definidos, a roda tende a se concentrar em uma pequena faixa do boleto, gerando tensões muito altas, sobrecarregando o trilho e suas fixações aos dormentes, podendo ocasionar a deformação do trilho e conseqüente queda do rodeiro, gerando um descarrilamento.

Os esforços transversais atuam diretamente nas condições de via e anomalias, potencializando o risco de ocorrências ferroviárias, principalmente em curvas, ainda mais havendo problemas estruturais que levem a abertura da bitola.

Os esforços transversais considerados por Hungria (2017) são:

- a. aceleração não compensada;
- b. velocidade e estado de manutenção;
- c. esforços de frenagem e tração;
- d. força transversal causada pela variação de temperatura.

A fórmula geral para o cálculo do esforço lateral é (HUNGRIA, 2017):

$$L = 1,1 P_{Le} \left( \frac{V^2}{381 R} \right) + \frac{P_{Le} V}{1000} + 2 E_t \left( \cos \frac{\delta}{2} \right) + 2 P_o \operatorname{sen} \left( \frac{l}{2 R} \right) \quad (1)$$

em que:

$P_{Le}$  = Peso por eixo da locomotiva, tf

$V$  = Velocidade, km/h

$R$  = Raio da curva, m

$E_t$  = Esforço trator, kgf

$\delta$  = Ângulo entre os engates do vagão e da locomotiva, graus

$P_o$  = Esforço longitudinal devido à variação de temperatura, tf

$l$  = Comprimento da barra do trilho, m

Hungria (2017) cita que os esforços transversais devem ser mantidos em níveis adequados, buscando-se maior segurança na operação ferroviária, através da melhora dos parâmetros de qualidade da via (foco em confiabilidade e previsibilidade) ou limitando-se o número de locomotivas em tração múltipla.

### **3. Aprendizado de máquina**

#### **3.1. Surgimento e evolução**

A história do Aprendizado de Máquina (AM) tem estreita ligação com a trajetória da Inteligência Artificial (IA). Em 1842, Ada Lovelace escreve, aquele que é considerado o primeiro algoritmo a ser processado por uma máquina, a máquina analítica desenvolvida por Charles Babbage, gerando o questionamento se estas máquinas poderiam vir a se tornar inteligentes (GOODFELLOW; BENGIO; COURVILLE, 2016).

A IA se inicia nos anos 1950, com o foco em máquinas pensantes, cujos principais personagens são Alan Turing e John von Neumann. Os avanços ocorridos nas décadas seguintes até os dias de hoje, levaram a evoluções, expansões de potencial e novas tecnologias.

A partir da década de 1980, o aprendizado de máquina se apresenta como uma nova área de pesquisas, com o propósito de capacitar os computadores em aprender e desenvolver modelos para realizar atividades em domínios específicos. Até que, no início do século XXI, surgem também o deep learning e a computação cognitiva (JONES, 2017).

Tom Mitchell (1997, p.2, tradução do autor) cita que “um programa de computador aprende através de uma experiência  $E$  a respeito de alguma classe de tarefas  $T$  e métrica de performance  $P$ , se esta performance em realizar tais tarefas  $T$ , conforme medidas  $P$ , são aprimoradas com a experiência  $E$ ”.

### **3.2. Técnicas de aprendizado de máquina**

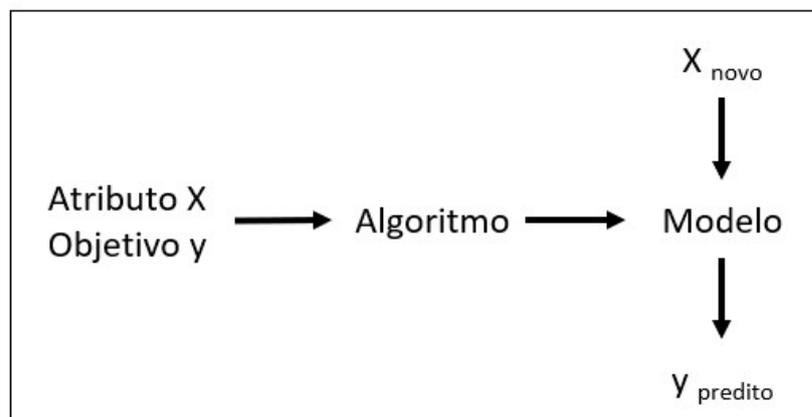
Originado no aprendizado indutivo, ao longo dos últimos 40 anos, o AM tem se desenvolvido em duas vertentes distintas, ou seja, técnicas de aprendizado supervisionado e não

supervisionado. Importante ressaltar que vários autores, ao longo dos últimos anos, têm considerado o aprendizado semi-supervisionado (CHAPELLE; SCHÖLKOPF; ZIEN, 2006).

O aprendizado supervisionado (Figura 1) tem por objetivo desenvolver e treinar modelos que, a partir de dados de entrada (variáveis independentes ou preditores) e dados de saída conhecidos, predigam respostas (variável dependente) das futuras observações (predições).

(JAMES et al., 2013; HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

**Figura 1 – Esquema básico de técnica de aprendizado supervisionado**



Fonte: Adaptado e traduzido de MOLNAR (2019).

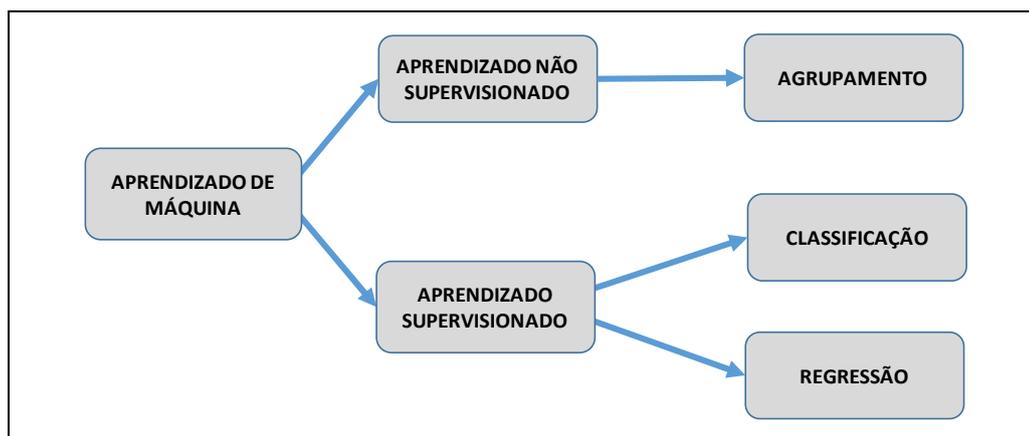
Enquanto isso, o aprendizado não-supervisionado não possui a figura metafórica do “instrutor” ou “professor” indicando as respostas corretas esperadas. Neste caso, busca-se identificar padrões que não estão explícitos ou estruturas intrínsecas nos dados de entrada, agrupando-os e, conseqüentemente, possibilitando a interpretação dos mesmos

(GOODFELLOW; BENGIO; COURVILLE, 2016; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES et al., 2013).

As técnicas de aprendizado supervisionado ainda podem ser subdivididas em dois grupos: Regressão e Classificação (Figura 2). Esta subdivisão diz respeito ao tipo de resposta que o problema estudado necessita. Se a resposta requerida está relacionada com variáveis quantitativas, que assumem valores numéricos, trata-se de um problema de regressão.

Porém, quando a resposta está relacionada às variáveis qualitativas (ou categóricas), tem-se um problema de classificação (GOODFELLOW; BENGIO; COURVILLE, 2016; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES et al., 2013).

**Figura 2 – Técnicas de aprendizado de máquina**



Fonte: Adaptado de HARRINGTON, 2012.

### **3.3. Métricas para avaliação de resultados**

As métricas utilizadas, em sua maioria, advêm das matrizes confusão (GIBERT et. al., 2016).

No caso deste estudo, trata-se de uma matriz de classificação binária (falha x não falha).

Várias são as métricas possíveis de serem utilizadas (LEVER; KRZYWINSKI; ALTMAN, 2016), de acordo com a necessidade específica de cada estudo, disponíveis a partir da matriz confusão, mas este estudo foca, principalmente, na acurácia (YIN; ZHAO, 2016; FINK; ZIO; WEIDMANN, 2013) e sensibilidade, além da especificidade (LI et. al., 2014) e eficiência, que se resume na média aritmética entre a sensibilidade e a especificidade.

Uma outra métrica utilizada neste estudo trata-se do coeficiente de concordância Kappa (COHEN, 1960), que retrata a confiabilidade e precisão dos dados classificados através do seu grau de concordância, assumindo valores menores ou igual a um e, sendo classificado, conforme Tabela 1.

**Tabela 1 – Relação do valor do coeficiente Kappa com força de concordância**

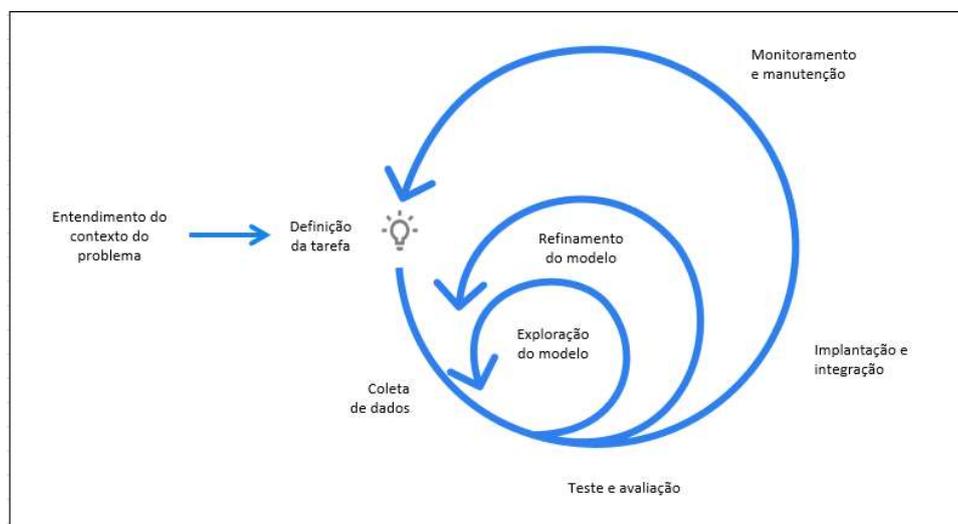
<b>Coeficiente <i>Kappa</i></b>	<b>Força de concordância</b>
<b>&lt; 0.00</b>	<b>Ausente</b>
<b>0.00-0.20</b>	<b>Mínima</b>
<b>0.21-0.40</b>	<b>Razoável</b>
<b>0.41-0.60</b>	<b>Moderada</b>
<b>0.61-0.80</b>	<b>Significativa</b>
<b>0.81-1.00</b>	<b>Quase perfeita</b>

Fonte: Traduzido de LANDIS; KOCK, 1977.

#### 4. Diagnóstico - aspectos metodológicos

A metodologia a ser utilizada neste estudo possui abordagem quantitativa e se baseia em Ciência de Dados, visto que se objetiva resolver problemas e responder a perguntas através da análise de dados, com o uso de técnicas de aprendizado de máquina. Portanto, a metodologia na qual este estudo se baseou refere-se ao Ciclo de Vida de Desenvolvimento de Aprendizado de Máquina sugerido por Jordan (2018) e apresentado na Figura 3.

Figura 3 – Ciclo de vida de desenvolvimento



Fonte: Traduzido de JORDAN (2018).

Na fase de contextualização e configuração do projeto, o conhecimento técnico acerca do problema contribui para a delimitação do escopo do projeto a ser desenvolvido. Portanto, antes de se iniciar o desenvolvimento do estudo, deve-se conhecer detalhadamente o

problema em questão, explorando-o e aprofundando-se nas análises de tal forma a garantir a identificação e caracterização precisa do problema.

Os resultados almejados através dos modelos preditores dependem intrinsecamente da quantidade e qualidade dos dados de entrada para os algoritmos, portanto a fase de Coleta e rotulagem dos dados demanda um esforço significativo, pois as bases de dados originais não necessariamente se apresentam da forma mais adequada à análise e rotulagem.

Visto que este estudo tem por objetivo principal o desenvolvimento de modelos preditores de falhas, ou seja, identificar se uma determinada observação do fenômeno indica um cenário de falha ou não falha (variável resposta qualitativa), define-se o problema por aprendizado supervisionado de classificação.

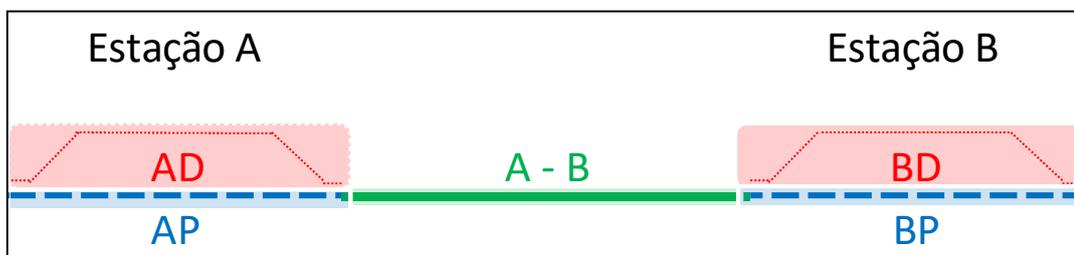
Com o objetivo de particularizar a análise e identificar o trecho crítico a ser estudado, foram consideradas todas as falhas de via permanente ocorridas entre 01 de janeiro de 2013 a 31 de dezembro de 2018, estratificando-as a partir de duas dimensões: localização geográfica das falhas e tipos de falhas.

Iniciou-se a análise pela identificação da localização geográfica das ocorrências de falhas ao longo das vias, sendo realizada a estratificação dos dados até se obter a supervisão de manutenção de via permanente crítica, ou seja, aquela com a maior quantidade de falhas por quilômetro de extensão.

#### 4.1. Definição do equipamento crítico

No caso deste estudo as linhas singelas e as duas linhas dos pátios de cruzamento são consideradas equipamentos, conforme Figura 4.

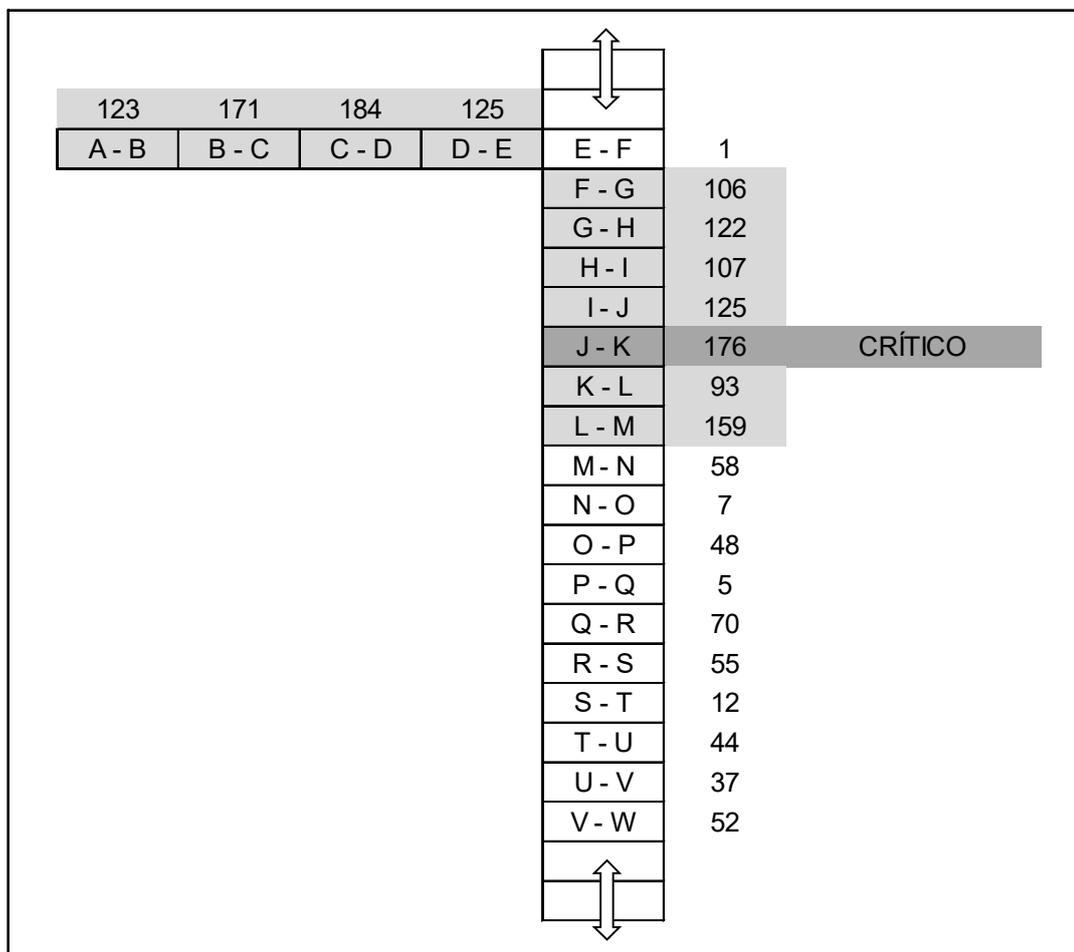
Figura 4 – Equipamentos típicos de uma ferrovia de linha singela



Fonte: Elaborado pelo autor, 2020.

Tendo por base a Figura 4 e o fato da maioria das falhas se encontrarem em equipamentos em linhas singelas, a Figura 5 apresenta somente estas linhas e respectivo quantitativo de falhas na supervisão identificada como crítica.

Figura 5 – Distribuição das falhas por equipamento na supervisão crítica



Fonte: Elaborado pelo autor, 2020.

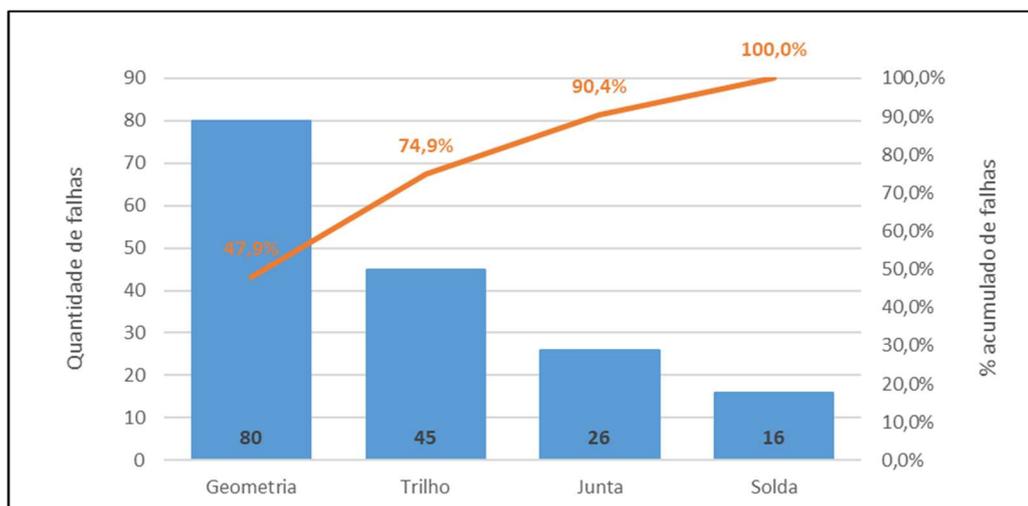
São identificados dois subtrechos distintos (ambos hachurados) com concentração de falhas. O subtrecho A–E refere-se a um ramal ferroviário que atende a um terminal integrador, enquanto o subtrecho E – W é a linha tronco. Levando-se em consideração o quantitativo de falhas, volume de tráfego mais intenso e homogeneidade na constituição da grade ferroviária (dormentes de madeira, fixação rígida e trilhos com perfil TR-37), optou-se pelo equipamento

J-K.

## 4.2. Identificação da falha crítica

O Gráfico 1 apresenta a estratificação das falhas por item, onde o item Geometria corresponde a 47,9% do total das falhas, portanto definido como item a ser detalhado.

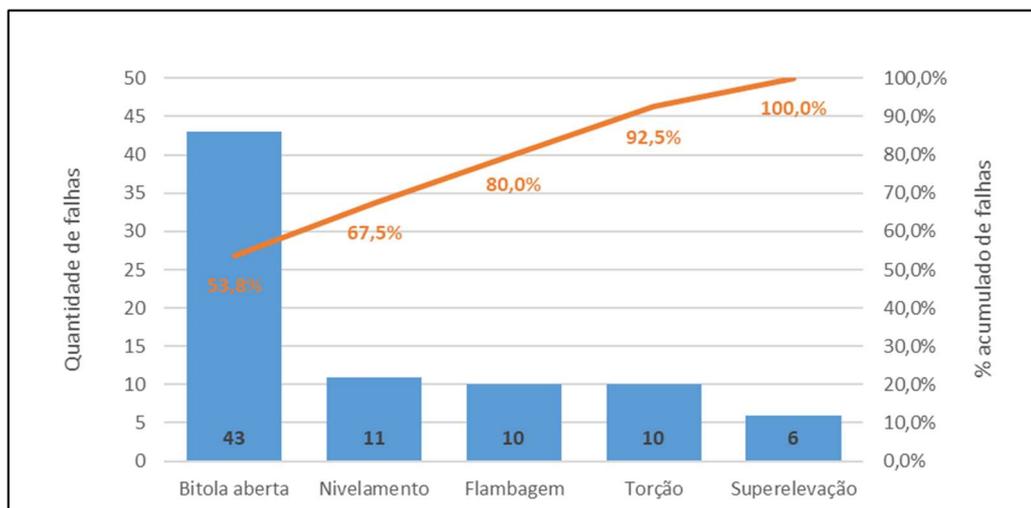
Gráfico 1 – Estratificação das falhas por item



Fonte: Elaborado pelo autor, 2020.

Detalhando-se os itens referentes aos conjuntos curva e tangente, tem-se que o item preponderante, com 53,8%, é bitola aberta (GRÁFICO 2).

Gráfico 2 – Detalhamento dos itens referentes aos conjuntos curva e tangente



Fonte: Elaborado pelo autor, 2020.

Em resumo, o estudo será direcionado para eventos de Bitola Aberta no equipamento J-K.

### 4.3. Coleta de dados

As bases de dados de defeitos e falhas de bitola aberta foram obtidas através dos sistemas informatizados onde tais dados são inseridos, sejam eles provenientes de medições manuais efetuadas pelos inspetores, técnicos ou rondas de linha, sejam eles mecanizados, obtidos através das inspeções do carro-controle.

O período estudado para este equipamento crítico compreende-se de 01 de janeiro de 2013 a 30 de novembro de 2019, onde foi efetuada a citada homogeneização dos dados.

#### **4.4. Formatação inicial da base de dados**

O equipamento estudado se trata de uma linha singela, portanto a falha em qualquer componente significa a falha do equipamento como um todo.

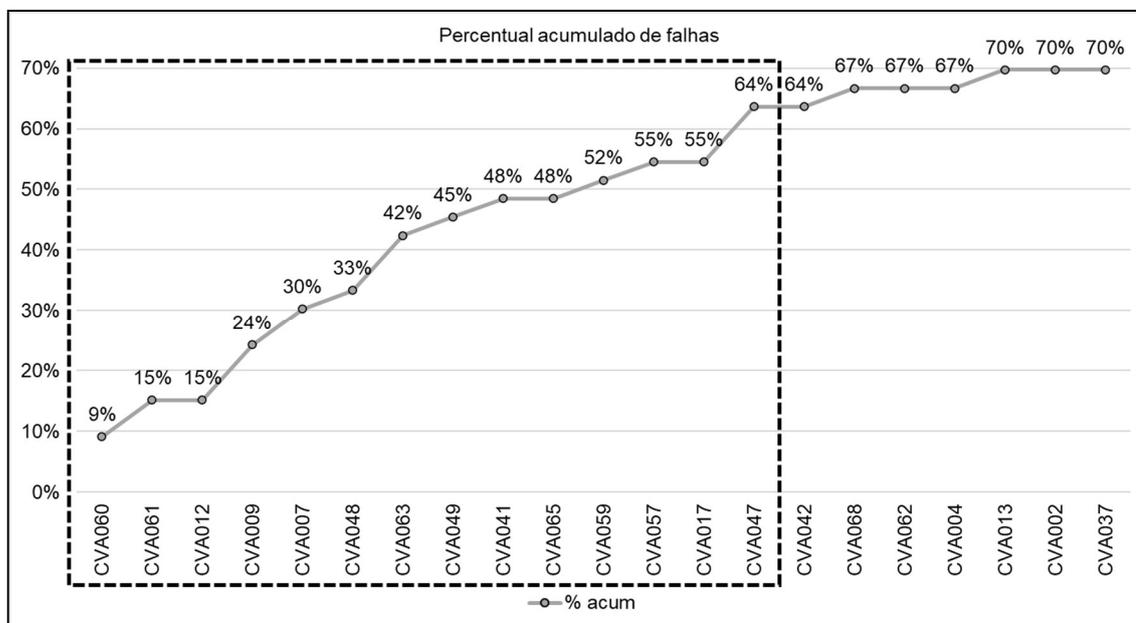
O equipamento estudado é formado por 67 curvas (denominadas CVA002 até CVA068) e 66 tangentes (denominadas TGT002 até TGT067), ou seja, 133 componentes ou características.

Segundo Yu e Liu (2003), em estudos nos quais há uma quantidade excessiva de características, faz-se necessária a escolha das características mais relevantes, evitando inclusive redundâncias, através de um determinado critério de avaliação, objetivando o aumento da eficiência na aprendizagem dos algoritmos, contribuindo em ganhos de acurácia, além de melhorar a compreensão dos resultados obtidos.

Cada um dos 133 componentes deste equipamento é considerado uma variável deste estudo, enquadrando-se na situação citada por Yu e Li (2003). Mediante este cenário, a definição da quantidade ótima de componentes ou variáveis a serem estudadas, levou-se em consideração a maior quantidade possível de observações, ou seja, datas não quais há a ocorrência de defeitos e/ou falhas no equipamento, de tal forma a garantir a representatividade do comportamento do equipamento ao longo do tempo.

Visto que um gráfico com todos os 133 componentes seria inviável de visualizar com qualidade os resultados obtidos, optou-se por estabelecer uma linha de corte no valor acumulado de 70%, onde estão representados 21 componentes.

**Gráfico 3 – Cenários analisados no equipamento J-K**



Fonte: Elaborado pelo autor, 2020.

A partir da Gráfico 3, constata-se que até o décimo quarto componente, ou seja, a curva CVA047, a inclinação da curva indica que o ritmo de acumulação de registros de falha é consistente e significativo. Porém, após o décimo quarto componente a inclinação perde força, ou seja, mesmo considerando outros componentes, o ganho percentual cumulativo de registros de falhas é menos expressivo.

Em apenas catorze componentes (10,6% do total de 133 componentes do equipamento J-K), tem-se 21 falhas (63,6% do total de 33 falhas), 913 defeitos (47,1% do total de 1940 defeitos) e 199 datas (64,8% do total de 307 datas).

Portanto estabelece-se, em função da representatividade apresentada, como base de estudo os catorze componentes indicados na área tracejada do Gráfico 3, para os quais foram considerados seus respectivos defeitos para a formatação da base de dados inicial.

#### **4.5. Formatação final da base de dados**

As inspeções, embora sigam uma frequência pré-estabelecida, não geram registros de falhas e/ou defeitos todos os dias, visto que não há ao menos uma inspeção sendo realizada diariamente. Devido a este fato, a avaliação passar a ser em períodos mensais, onde a base de dados teve que ser revista para apresentar valores mensais e não mais diário.

Mesmo com os procedimentos adotados até então, a base de dados de defeitos ainda apresentava dados faltantes. Para fins de preenchimento dos dados faltantes, utilizou-se o algoritmo missforest do R® na base de dados inicial (STEKHOVEN; BUHLMANN, 2012).

A fase seguinte corresponde à imputação dos valores das falhas nos respectivos componentes, meses e anos das ocorrências das mesmas.

Para que os resultados deste estudo viabilizem ações preventivas de manutenção, definiu-se um período de dois meses de antecedência para predição de falha no equipamento estudado. Este seria o tempo mínimo para ações de manutenção focadas principalmente em inspeções e intervenções localizadas e programadas baseado no fluxo de trabalho atual, onde os inputs são gerados, de forma genérica, no mês 0.

No início do mês 1 tais inputs seriam lidos pelo modelo e seriam realizadas as previsões. Se estas previsões indicarem que no mês 2 ocorrerá falha, já ao longo do mês 1 as inspeções podem ser intensificadas e ações emergenciais de manutenção, devidamente negociadas com o Centro de Controle Operacional, também podem ser tomadas. Além disso, ao final do mês 1, ações complementares e de caráter mais duradouros podem ser programadas para o mês 2 e efetivadas, ao longo do mês 2, garantindo-se a segurança operacional.

#### **4.6. Divisão da base de dados entre treino e teste**

A divisão da base de dados em base de treino e base de teste é uma tarefa muito particular e depende muito de cada estudo realizado, mas alguns percentuais já são comumente usados. Os valores encontrados mais comuns ficam entre 70 e 80% dos dados direcionados para treino dos algoritmos e 20 a 30% destinados para os testes dos modelos gerados (YIN; ZHAO, 2016). Neste estudo, adotou-se que a linha de corte seria entre os meses de junho e julho de 2018, portanto a relação adotada foi 76% para dados treino e 24% para dados teste, dentro do alcance citado anteriormente.

#### **4.7. Exploração de modelos**

De posse das bases treino e teste completos, iniciou-se então a etapa de exploração de modelos (CHURPEK et al., 2016; HESLOT et al., 2012), objetivando-se identificar o(s) modelo(s)

e a(s) técnica(s) que venha(m) a apresentar a melhor resposta aos objetivos deste estudo, conforme métricas definidas.

#### **4.8. Escolha dos algoritmos**

O tipo de problema a ser resolvido, em linhas gerais, direciona a escolha por alguns tipos de algoritmos a serem usados, mas, normalmente, não há uma definição clara a respeito do algoritmo ideal, portanto várias opções foram testadas e avaliadas em busca da que gera os melhores resultados conforme o problema enfrentado.

Eis os algoritmos (ou métodos) escolhidos na biblioteca caret, perfazendo um total de 16 opções com o seu respectivo nome a ser utilizado ao longo deste estudo, ambos apresentados na Tabela 2, junto à família de algoritmos a qual pertencem.

**Tabela 2 – Lista de algoritmos, métodos e modelos utilizados no estudo**

<b>Família de algoritmos</b>	<b>Método</b>	<b>Nome do modelo</b>
<b><i>Logistic Regression</i></b>	<b><i>Glm</i></b>	<b><i>GLM</i></b>
	<b><i>LogitBoost</i></b>	<b><i>LogitBoost</i></b>
	<b><i>regLogistic</i></b>	<b><i>RegLog</i></b>
<b><i>SVM</i></b>	<b><i>svmRadial</i></b>	<b><i>svmRadial</i></b>
	<b><i>svmPoly</i></b>	<b><i>svmPoly</i></b>
	<b><i>svmLinear</i></b>	<b><i>svmLinear</i></b>
	<b><i>svmLinear2</i></b>	<b><i>svmLinear2</i></b>
<b><i>Neural Network</i></b>	<b><i>Nnet</i></b>	<b><i>NeuralNet</i></b>
<b><i>Adaboost</i></b>	<b><i>Adaboost</i></b>	<b><i>ADA</i></b>
<b><i>Random Forest</i></b>	<b><i>Rf</i></b>	<b><i>RF</i></b>
	<b><i>Ranger</i></b>	<b><i>RFranger</i></b>

<b>XGBoost</b>	<b><i>xgbDART</i></b>	<b><i>xgbDART</i></b>
	<b><i>xgbLinear</i></b>	<b><i>xgbLinear</i></b>
	<b><i>xgbTree</i></b>	<b><i>xgbTree</i></b>
<b><i>Naïve Bayes</i></b>	<b><i>Nb</i></b>	<b><i>NB</i></b>
<b><i>KNN</i></b>	<b><i>Knn</i></b>	<b><i>KNN</i></b>

Fonte: Elaborado pelo autor, 2020.

#### 4.9. Análise horizontal

A análise horizontal consiste em avaliar a performance dos 16 algoritmos elencados utilizando-se o ajuste de hiperparâmetros padrão de cada um, portanto sem aprofundamento na geração dos modelos.

##### 4.9.1. Primeira abordagem: Análise inicial

A primeira abordagem consiste em aplicar a base treino composta por 14 variáveis nos 16 algoritmos escolhidos (Tabela 2), sem qualquer ajuste de hiperparâmetros, usando validação cruzada com 5 partições e 5 repetições.

##### 4.9.2. Segunda abordagem: Eliminação recursiva de características

Buscando-se o aprimoramento dos modelos, decidiu-se pela implementação da eliminação recursiva de características, cujo objetivo é obter o menor conjunto de características que satisfaçam o treinamento do modelo de predição.

**26ª SEMANA DE TECNOLOGIA METROFERROVIÁRIA**  
**7º PRÊMIO TECNOLOGIA E DESENVOLVIMENTO METROFERROVIÁRIOS**



A eliminação recursiva de características foi utilizada através da função rfe da biblioteca caret, com todas as 4 funções disponíveis (rfFuncs, lmFuncs, nbFuncs e treebagFuncs) e validação cruzada com 5 partições e 5 repetições.

Os resultados obtidos indicaram a utilização do subconjunto formado pelas variáveis CVA060, CVA017, CVA007, CVA057 e CVA059.

A Tabela 3 apresenta os resultados obtidos com os 16 modelos, usando-se cinco e catorze características e sem ajuste de hiperparâmetros.

**26ª SEMANA DE TECNOLOGIA METROFERROVIÁRIA**  
**7º PRÊMIO TECNOLOGIA E DESENVOLVIMENTO METROFERROVIÁRIOS**



**Tabela 3 – Resultados comparativos com cinco e catorze características**

Modelo	Quant. Variáveis	Acurácia modelo	kappa modelo	Acurácia	kappa	Sensitividade	Especificidade	Eficiência
GLM	14	70,90%	0,0026	64,71%	0,2388	42,86%	80,00%	61,43%
	5	83,16%	0,0578	70,59%	0,4056	71,43%	70,00%	70,71%
LogitBoost	14	84,36%	-0,0029	76,47%	0,4925	57,14%	90,00%	73,57%
	5	82,21%	0,0488	64,71%	0,2388	42,86%	80,00%	61,43%
xgbLinear	14	84,07%	0,0458	58,82%	0,0480	14,29%	90,00%	52,14%
	5	86,72%	0,1601	64,71%	0,1639	14,29%	100,00%	57,14%
RFranger	14	86,72%	0,0000	58,82%	0,0000	0,00%	100,00%	50,00%
	5	86,00%	0,0793	64,71%	0,1639	14,29%	100,00%	57,14%
NB	14	72,21%	-0,0760	58,82%	0,0000	0,00%	100,00%	50,00%
	5	78,76%	-0,0492	64,71%	0,1639	14,29%	100,00%	57,14%
svmPoly	14	86,72%	0,0000	58,82%	0,0000	0,00%	100,00%	50,00%
	5	87,09%	0,0248	64,71%	0,0000	14,29%	100,00%	57,14%
RF	14	86,72%	0,0000	58,82%	0,0000	0,00%	100,00%	50,00%
	5	85,96%	0,1078	64,71%	0,0000	14,29%	100,00%	57,14%
NeuralNet	14	86,72%	0,0000	58,82%	0,0000	0,00%	100,00%	50,00%
	5	86,94%	0,2036	47,06%	0,0129	71,43%	30,00%	50,71%
xgbTree	14	86,40%	0,0645	58,82%	0,0000	0,00%	100,00%	50,00%
	5	88,65%	0,2007	58,82%	0,0000	0,00%	100,00%	50,00%
xgbDART	14	87,12%	0,0400	58,82%	0,0000	0,00%	100,00%	50,00%
	5	88,61%	0,1900	58,82%	0,0000	0,00%	100,00%	50,00%
RegLog	14	86,72%	0,0000	58,82%	0,0000	0,00%	100,00%	50,00%
	5	86,72%	0,0000	58,82%	0,0000	0,00%	100,00%	50,00%
KNN	14	86,72%	0,0000	58,82%	0,0000	0,00%	100,00%	50,00%
	5	86,72%	0,0000	58,82%	0,0000	0,00%	100,00%	50,00%
svmRadial	14	85,56%	-0,0144	58,82%	0,0000	0,00%	100,00%	50,00%
	5	86,72%	0,0000	58,82%	0,0000	0,00%	100,00%	50,00%
svmLinear	14	85,56%	-0,0144	58,82%	0,0000	0,00%	100,00%	50,00%
	5	86,72%	0,0000	58,82%	0,0000	0,00%	100,00%	50,00%
svmLinear2	14	85,52%	-0,0133	58,82%	0,0000	0,00%	100,00%	50,00%
	5	86,72%	0,0000	58,82%	0,0000	0,00%	100,00%	50,00%
ADA	14	79,56%	-0,0398	58,82%	0,0000	0,00%	100,00%	50,00%
	5	83,23%	0,1686	58,82%	0,0000	0,00%	100,00%	50,00%

Fonte: Elaborado pelo autor, 2020.

## 4.10. Análise vertical

### 4.10.1. Refinamento dos modelos, testes e avaliações

Desta fase do estudo em diante é considerado somente o subconjunto formado por 5 características e os nove modelos de melhor performance advindos da Tabela 3, ou seja: GLM, LogitBoost, xgbLinear, RFranger, NB, svmPoly, RF, NeuralNet e xgbTree.

#### **4.10.2. Ajuste de hiperparâmetros**

A estratégia nesta fase do estudo consiste em variar um dos hiperparâmetros, enquanto todos os demais permanecem fixos, avaliando os resultados obtidos, e assim sucessivamente, até passar por todos os hiperparâmetros disponíveis para ajuste, buscando-se definir a melhor performance para cada um dos nove modelos priorizados.

O modelo GLM é o único que não possui hiperparâmetros ajustáveis no caret, portanto não apresenta alteração de performance. Todos os demais possuem ao menos um hiperparâmetro ajustável.

#### **4.10.3. Balanceamento da base treino**

Para o balanceamento da base treino foi utilizada a metodologia SMOTE (CHAWLA et al., 2002). No caso deste estudo, esta situação ocorre na base de dados de treino, onde 85,6% das observações são referentes a não falha, enquanto 13,4% são observações que correspondem a falhas. Após a aplicação desta metodologia, usando-se a função SMOTE, obtém-se uma nova base com 50% de falhas e 50% com não falhas.

#### **4.10.4. Base treino balanceada sem e com ajuste de hiperparâmetros**

De posse da base treino balanceada, foram gerados novos modelos usando-se os nove algoritmos elencados conforme Tabela 3, sem ajuste de hiperparâmetros. As métricas utilizadas nesta seção para identificar os modelos de melhor performance foram acurácia

acima de 60%, sensibilidade acima de 40% e coeficiente Kappa entre 0,2 e 0,5 (concordância variando entre mínima a moderada).

A partir desta identificação dos modelos de melhor performance e objetivando-se então obter uma melhoria na performance dos modelos, passa-se à próxima etapa, onde é realizado o ajuste dos hiperparâmetros.

#### **4.10.5. Estudo detalhado do algoritmo LogitBoost**

Na Tabela 3, identificou-se o algoritmo LogitBoost com catorze variáveis apresentava uma melhor performance quando comparado a cinco variáveis, único com este comportamento dentre os dezesseis algoritmos originalmente definidos para análise. Devido a este comportamento único, decidiu-se abordar este algoritmo, usando-se as catorze características, sob alguns cenários distintos, conforme detalhado a seguir:

- a. base treino balanceada com ajuste;
- b. base treino balanceada sem ajuste;
- c. base treino original com ajuste;
- d. base treino original sem ajuste.

#### **4.10.6. Implantação e manutenção do modelo**

As duas últimas etapas podem vir a acontecer posteriormente à finalização deste estudo, portanto não fazem parte do escopo deste estudo.

### **5. ANÁLISE DOS RESULTADOS**

#### **5.1. Compilação de resultados**

Neste tópico, objetiva-se compilar os resultados obtidos ao final deste estudo, indicando os de melhor performance diante das métricas definidas.

A Tabela 4 apresenta a compilação, comparando-se a quantidade de variáveis utilizadas para o treinamento da base treino, o uso ou não de ajuste e o uso ou não de base balanceada. Estão, portanto, destacados os oito melhores resultados considerando-se as métricas definidas na seção 4.2.5, ou seja:

- a. acurácia e sensibilidade > 75%;
- b. especificidade e eficiência > 70% e;
- c. coeficiente *Kappa* > 0,4.

**Tabela 4 – Resultados finais**

Modelo	Variáveis	Ajuste	Balanceamento	Acurácia	kappa	Sensitividade	Especificidade	Eficiência
LogitBoost	14	SIM	SIM	82,35%	0,6107	57,14%	100,00%	78,57%
RF	5	SIM	SIM	76,47%	0,5143	71,43%	80,00%	75,71%
svmPoly	5	SIM	NÃO	76,47%	0,4925	57,14%	90,00%	73,57%
LogitBoost	14	SIM	NÃO	76,47%	0,4925	57,14%	90,00%	73,57%
LogitBoost	14	NÃO	NÃO	76,47%	0,4925	57,14%	90,00%	73,57%
RFranger	5	SIM	SIM	76,47%	0,4687	42,86%	100,00%	71,43%
RF	5	NÃO	SIM	76,47%	0,4687	42,86%	100,00%	71,43%
xgbTree	5	SIM	SIM	76,47%	0,4687	42,86%	100,00%	71,43%

Fonte: Elaborado pelo autor, 2020.

## 5.2. Análise dos resultados

Conforme demonstrado na Tabela 4, oito modelos apresentaram um resultado diferenciado em relação aos demais, suplantando quatro das cinco métricas definidas. Porém, nenhum deles atingiu o mínimo de 75% de sensibilidade, indicando que os modelos apresentam dificuldade em identificar os verdadeiros positivos.

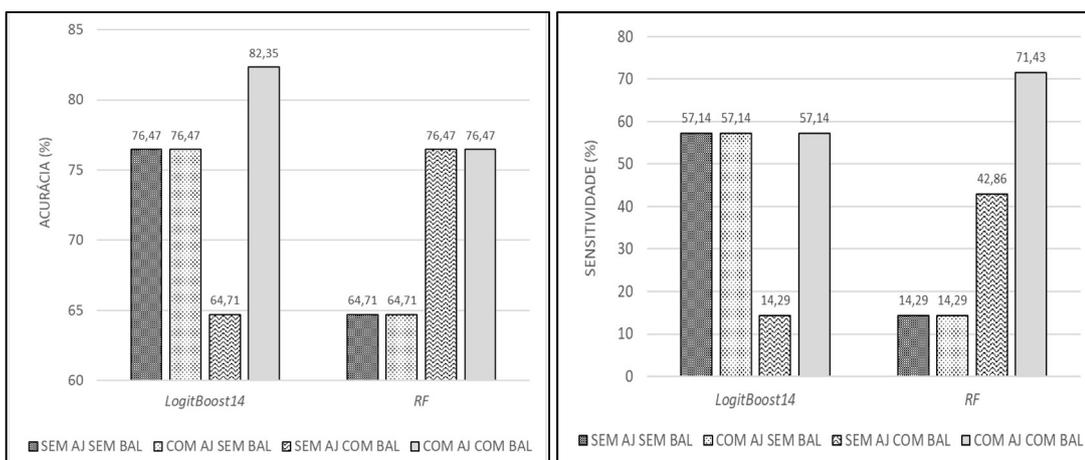
O empilhamento de modelos chegou a ser utilizado, mas não se demonstrou viável neste estudo.

Os dois modelos mais promissores são LogitBoost com 14 características, com ajuste e com balanceamento (denominado LogitBoost14) e RF com 5 características, com ajuste e com balanceamento. A evolução dos resultados obtidos para estes dois modelos pode ser melhor identificada nos Gráficos 4 a 8 apresentados.

Os valores são apresentados na ordem sequencial da evolução dos modelos, ou seja:

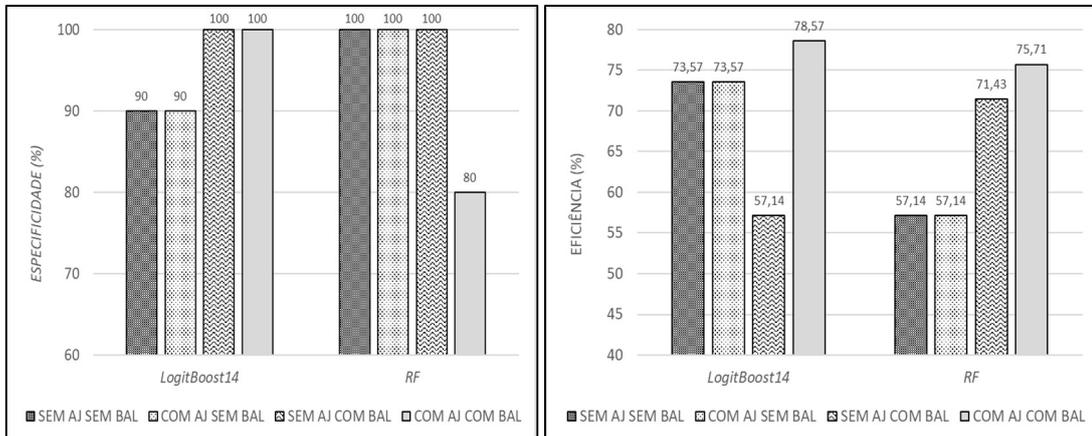
- SEM AJ SEM BAL: sem ajuste de parâmetros e sem balanceamento da base treino;
- COM AJ SEM BAL: com ajuste de parâmetros e sem balanceamento da base treino;
- SEM AJ COM BAL: sem ajuste de parâmetros e com balanceamento da base treino;
- COM AJ COM BAL: com ajuste de parâmetros e com balanceamento da base treino.

**Gráficos 4 e 5 – Evolução da acurácia e sensibilidade (%)**



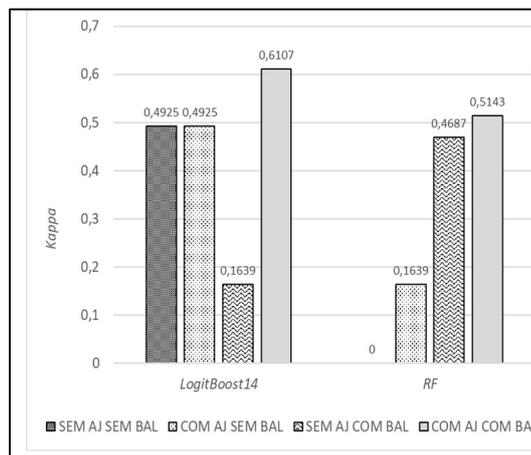
Fonte: Elaborado pelo autor, 2020.

Gráfico 6 e 7 – Evolução da especificidade e eficiência (%)



Fonte: Elaborado pelo autor, 2020.

Gráfico 8 – Evolução do coeficiente *Kappa*



Fonte: Elaborado pelo autor, 2020.

Outros modelos se mostraram igualmente viáveis neste estudo e devem ser considerados e explorados em estudos futuros. São eles: RFranger, SVMPoly (com ajuste de parâmetros,

porém com a base treino sem balanceamento) e xgbTree, ou seja, não se denota uma tendência específica para uma determinada família de algoritmos, ressaltando a importância da avaliação de uma ampla gama de algoritmos.

Na situação em que se desenvolveu este estudo é factível conviver com uma acurácia que não seja a maior possível, mas onde a sensibilidade seja privilegiada. É importante obter o equilíbrio entre estas duas métricas, visto que um excesso de falsos negativos pode comprometer o desenvolvimento futuro de uma ferramenta promissora (YELLA; DOUGHERTY; GUPTA, 2009).

A qualidade dos dados obtidos, mesmo diante dos resultados alcançados, esteve compatível com as técnicas de predição usadas em aprendizado de máquina, visto que foi possível, através de técnicas diversas, análises sucessivas e proposição de novas técnicas e cenários distintos, apresentar evolução na qualidade preditora dos modelos.

Outro ponto a ressaltar é a importância do balanceamento da base treino, onde a maioria dos modelos de melhor performance conforme Tabela 4 (cinco em um total de oito), obteve melhores resultados com a base balanceada.

Em contrapartida, o empilhamento de modelos não se demonstrou viável neste estudo.

### **5.3. Impactos operacionais e financeiros correlatos**

Um fluxo de transporte de grãos percorre o trecho no qual está localizado o equipamento crítico objeto deste estudo.

Em função do ganho em previsibilidade, obtido através da utilização dos modelos preditivos apresentados neste estudo, é possível contribuir para a redução das paradas não programadas do tráfego ferroviário, contribuindo para o incremento de receita e ganhos em CAPEX.

#### **5.4. Impactos em gestão de riscos operacionais**

Outros ganhos que devem ser considerados, são aqueles relativos à gestão de risco da empresa, visto que uma falha de bitola aberta pode acarretar ou contribuir para uma ocorrência ferroviária (descarrilamento, tombamento ou adernamento de vagões e/ou locomotivas).

Ocorrências ferroviárias podem vir a impactar negativamente tanto em termos operacionais como financeiros. Os impactos financeiros advêm de vários fatores, tais como, perda receita, perdas dos ativos envolvidos na ocorrência ou gastos na recuperação dos mesmos, gastos no atendimento às ocorrências e, em casos extremos, indenizações a terceiros impactados.

Outros impactos potenciais também podem ocorrer, tais como, impactos ao meio ambiente (APPs, APAs, cursos d'água, pontos de outorga, etc.), à segurança pessoal dos empregados e população limítrofe (a ferrovia cruza dezenas de perímetros urbanos), de aspecto regulatório (bens concedidos e arrendados pela União) e, inclusive, imagem da empresa.

## **6. CONCLUSÕES**

Este estudo teve como principal objetivo executar uma análise comparativa através do desenvolvimento de modelos de predição para falhas usando técnicas de aprendizado de máquina, tendo como direcionamento principal contribuir para a previsibilidade das ações de manutenção da via permanente, levando a um planejamento mais assertivo. Especificamente, focou-se nas falhas oriundas de bitola aberta.

Desta forma, conclui-se que os objetivos esperados para este estudo foram alcançados, mesmo com as limitações existentes, pois é apresentada a possibilidade de se desenvolver uma ferramenta de predição para a ferrovia como um todo, direcionada não somente a um único tipo de falha, buscando contribuir para um aprimoramento do planejamento das ações de manutenção. Além disso, demonstra-se a necessidade da geração de dados em maior quantidade e qualidade de tal forma a aprimorar continuamente os modelos de predição de falha, reduzindo a probabilidade de overfitting e contribuindo também nas demais métricas de avaliação, tais como, sensibilidade, especificidade e eficiência, além de trazer como consequência intervenções de cunho cada vez mais preventivo, redução de custos, menores impactos operacionais e, conseqüentemente, retornos financeiros mais expressivos.

Deve-se ressaltar que todos os ganhos apresentados até aqui consideram o estudo realizado em apenas um dos mais de quatrocentos equipamentos da empresa, e em apenas um dos fluxos de transporte de cargas, portanto as potencialidades da implantação deste estudo na empresa, são promissores e significativos.

**26ª SEMANA DE TECNOLOGIA METROFERROVIÁRIA**  
**7º PRÊMIO TECNOLOGIA E DESENVOLVIMENTO METROFERROVIÁRIOS**



Como sugestão de outro trabalho futuro, há a possibilidade de, ao invés de se focar em um único equipamento, analisar-se subtrechos com vários equipamentos, desde que estes equipamentos tenham a sua constituição e comportamentos semelhantes. Ou, avaliar a probabilidade associada à ocorrência da falha em face do conjunto de variáveis explanatórias.

Além do estudo em subtrechos homogêneos citado no parágrafo anterior, sugere-se também como trabalho futuro a análise de séries temporais nas falhas de bitola aberta. E, obtendo-se êxito, é possível também a extrapolação para outros tipos de falhas, inclusive defeitos em evolução para falhas (com uma visão mais preventiva), não somente da superestrutura ferroviária, mas da infraestrutura ferroviária também.

Outra opção de trabalho futuro seria a conjugação da base de dados de falhas e defeitos de bitola aberta com as quatro componentes dos esforços transversais (HUNGRIA, 2017), principalmente, os efeitos de temperaturas elevadas em curvas.

Usando-se o mesmo raciocínio apresentado neste estudo, pode-se analisar a contribuição dos defeitos geométricos como um todo, citados na seção 3.3.2 (não somente as bitolas abertas) em falhas críticas dos equipamentos, ou seja, ocorrências ferroviárias. Neste caso, outras informações podem ser conjugadas, tais como, temperatura no trilho ou na região de estudo e índices pluviométricos.

Recomenda-se deixar aberta a possibilidade de uso de algoritmos diferentes para equipamentos ou trechos com vários equipamentos distintos. Esta possibilidade reside no fato da via permanente ser heterogênea, se comportando de maneiras diferentes em trechos diferentes, seja em função dos materiais que a constituem e das condições destes materiais,

**26ª SEMANA DE TECNOLOGIA METROFERROVIÁRIA**  
**7º PRÊMIO TECNOLOGIA E DESENVOLVIMENTO METROFERROVIÁRIOS**



seja em função das condições climáticas, seja em função do volume de tráfego ou da intensidade das ações de manutenção executadas.

Finalmente, recomenda-se o direcionamento de estudos e desenvolvimentos correlatos, usando-se num primeiro momento algoritmos diversos, focando-se em bases treino balanceadas, ajustes precisos de parâmetros e atenção para o uso ou não de subconjuntos de variáveis para treinamento dos modelos, testando todas as alternativas possíveis.

## 7. REFERÊNCIAS BIBLIOGRÁFICAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **ABNT NBR 5462**: confiabilidade e manutenibilidade: terminologia. Rio de Janeiro: ABNT, 1994. 37 p.

BARROW, Keith. Precision Schedule Railroading: evolution or revolution? **International Railway Journal**, New York, v. 59, n. 9, p. 24-26, set. 2019.

BRINA, H. L. **Estradas de Ferro I**. Rio de Janeiro: LTC, 1979. 258 p.

CASTELLO BRANCO, J. E. S.; FERREIRA, R. **Tratado de estradas de ferro vol. II**: prevenção e investigação de descarrilamentos. Rio de Janeiro: Reflexus, 2002. 534 p.

CHAPELLE, Olivier; SCHÖLKOPF, Bernhard; ZIEN, Alexander. **Deep semi-supervised learning**. Cambridge, Massachusetts: Massachusetts Institute of Technology, 2006.

CHAWLA, N. V. *et al.* SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 1 June 2002.

CHURPEK, Matthew M. *et al.* Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. **Critical Care Medicine**, v. 44, n. 2, p. 368–374, 2016.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37-46, 1960.

COIMBRA, M. V. **Modos de falha dos componentes da via permanente ferroviária e seus efeitos no meio ambiente**. 2008. 165 f. Dissertação (Mestrado em Engenharia de Transportes) – Instituto Militar de Engenharia, Rio de Janeiro, 2008.

GIBERT, X. *et al.* Deep multitask learning for railway track inspection. **IEEE Transactions on Intelligent Transportation Systems**, v. 18, n. 1, p. 153–164, 2017.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, Massachusetts, USA: MIT Press, 2016.

HARRINGTON, P. **Machine learning in action**. Shelter Island, New York, USA: Manning Publications, 2012.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**. New York, NY: Springer New York, 2009.

HESLOT, N. *et al.* Genomic selection in plant breeding: a comparison of models. **Crop Science**, v. 52, n. 1, p. 146–160, 2012.

HUNGRIA, L. H. **Segurança Operacional de Trens de Carga**. São Paulo: All Print Editora, 2017. 816 p.

ILDEFONSO, J. S. **Proposta de um sistema de gerência para vias férreas brasileiras**. 2013. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2013.

JAMES, G. *et al.* **An introduction to statistical learning**. New York, NY: Springer New York, 2013. v. 103.

JONES, M. T. **Um guia para iniciantes sobre inteligência artificial, aprendizado de máquina e computação cognitiva**. [S. l.]: IBM, 2017. Disponível em:

<https://www.ibm.com/developerworks/br/library/guia-iniciantes-ia-maquina-computacao->

**26ª SEMANA DE TECNOLOGIA METROFERROVIÁRIA**  
**7º PRÊMIO TECNOLOGIA E DESENVOLVIMENTO METROFERROVIÁRIOS**



cognitiva/guia-iniciantes-ia-maquina-computacao-cognitiva-

pdf.pdf%0Ahttps://www.ibm.com/developerworks/br/library/guia-iniciantes-ia-maquina-computacao-cognitiva/index.html. Acesso em: 10 jul. 2019.

JORDAN, J. **Organizing machine learning projects**: project management guidelines. [S. l.]: O Autor, 2018. Disponível em: <https://www.jeremyjordan.me/ml-projects-guide/>. Acesso em: 11 ago. 2019.

LEVER, J.; KRZYWINSKI, M.; ALTMAN, N. Classification evaluation. **Nature Methods**, v. 13, n. 8, p. 603–604, 2016.

LI, H. *et al.* Improving rail network velocity: a machine learning approach to predictive maintenance. **Transportation Research Part C: Emerging Technologies**, v. 45, p. 17–26, 2014.

MITCHELL, T. M. **Machine Learning**. New York, USA: McGraw-Hill Science, 1997.

MOLNAR, C. **Interpretable Machine Learning**. A Guide for Making Black Box Models Explainable. Mountain View, California, USA: Creative Commons, 2019.

NABAIS, R. J. S. **Manual básico de engenharia ferroviária**. São Paulo: Oficina de Textos, 2014.

RIVES, F. O.; PITA, A. L.; PUENTE, M. J. M. **Tratado de ferrocarriles I – via**. Madrid: Rueda, 1977.

SCHRAMM, G. **Técnica e economia na via permanente**. Tradução por Rudy A. Volkmann. Porto Alegre: Meridional, 1977. 297 p.

STEFFLER, F. **Via permanente aplicada: guia teórico e prático.** Rio de Janeiro: LTC, 2013. 314 p.

STEKHOVEN, D. J.; BUHLMANN, P. MissForest-non-parametric missing value imputation for mixed-type data. **Bioinformatics**, v. 28, n. 1, p. 112–118, 1 jan. 2012.

STOPATTO, S. **Via permanente ferroviária: conceitos e aplicações.**, São Paulo: USP, 1987. 251 p.

YELLA, S.; DOUGHERTY, M.; GUPTA, N. K. Condition monitoring of wooden railway sleepers. **Transportation Research Part C: Emerging Technologies**, v. 17, n. 1, p. 38–55, 2009.

YIN, J.; ZHAO, W. Fault diagnosis network design for vehicle on-board equipments of high-speed railway: A deep learning approach. **Engineering Applications of Artificial Intelligence**, v. 56, n. Oct. p. 250–259, 2016.

YU, L.; LIU, H. Feature selection for high-dimensional data: a fast correlation-based filter solution. *In*: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 20., 2003, Washington, Dc, USA. **Proceedings** [...]. Washington, Dc, USA: AAAI, 2003. v. 2, p. 856–863.