

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO  
CENTRO TECNOLÓGICO  
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA CIVIL  
MESTRADO EM ENGENHARIA CIVIL

BRUNA DOS SANTOS NEVES

DISSERTAÇÃO DE MESTRADO

PREDIÇÃO DA NECESSIDADE DE IMPOSIÇÃO DE RESTRIÇÃO TEMPORÁRIA  
DE VELOCIDADE NA VIA PERMANENTE UTILIZANDO ALGORITMOS DE  
*MACHINE LEARNING*

VITÓRIA  
2021

BRUNA DOS SANTOS NEVES

**DISSERTAÇÃO DE MESTRADO**

**PREDIÇÃO DA NECESSIDADE DE IMPOSIÇÃO DE RESTRIÇÃO TEMPORÁRIA  
DE VELOCIDADE NA VIA PERMANENTE UTILIZANDO ALGORITMOS DE  
*MACHINE LEARNING***

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Civil do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito para qualificação.  
Orientador: Prof. Dr. Rodrigo de Alvarenga Rosa.

Coorientador: Dr. Jodelson Sabino.

VITÓRIA  
2021

Ficha catalográfica disponibilizada pelo Sistema Integrado de  
Bibliotecas - SIBI/UFES e elaborada pelo autor

---

N511p      Neves, Bruna dos Santos, 1994-  
             Predição da necessidade de imposição de restrição temporária  
             de velocidade na via permanente utilizando algoritmos de  
             machine learning / Bruna dos Santos Neves. - 2021.  
             63 f.

             Orientador: Rodrigo de Alvarenga Rosa.  
             Coorientador: Jodelson Aguilar Sabino.  
             Dissertação (Mestrado em Engenharia Ambiental) -  
             Universidade Federal do Espírito Santo, Centro Tecnológico.

             1. Ferrovias. 2. Aprendizado do Computador. 3. Ferrovias -  
             Velocidade. 4. Ferrovias - Manutenção e reparos. I. Rosa, Rodrigo  
             de Alvarenga. II. Sabino, Jodelson Aguilar. III. Universidade  
             Federal do Espírito Santo. Centro Tecnológico. IV. Título.

CDU: 628

---

# **PREDIÇÃO DA NECESSIDADE DE IMPOSIÇÃO DE RESTRIÇÃO TEMPORÁRIA DE VELOCIDADE NA VIA PERMANENTE UTILIZANDO ALGORITIMOS DE MACHINE LEARNING**

Dissertação apresentada ao Curso de Mestrado em Engenharia Civil do Programa de Pós-Graduação em Engenharia Civil da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do título de Mestre em Engenharia Civil.

Aprovado em \_\_\_\_ de \_\_\_\_\_ de 2021 por:

---

Prof. Dr. Rodrigo de Alvarenga Rosa  
Doutor em Engenharia Elétrica  
Orientador

---

Dr. Jodelson Sabino  
Vale  
Coorientador

---

Prof. Dr. Patricio Pires  
Universidade Federal do Espírito Santo  
Examinador Interno

---

Prof. Dr. Gustavo Pessim  
Instituto Tecnológico Vale  
Examinador Externo

**Vitória – ES, março de 2021**

## RESUMO

No Brasil, o transporte ferroviário é essencial para a economia do país. A via ferroviária é um sistema de transporte sobre trilhos, constituído de via permanente e outras instalações fixas, equipamentos de tráfego e demais acessórios indispensáveis à condução segura e eficiente de passageiros e/ou cargas. De acordo com a infraestrutura e a superestrutura de uma via permanente, no projeto geométrico da via é definida a Velocidade Máxima Autorizada (VMA). A VMA é velocidade máxima de projeto para cada trecho da via, podendo ser alterada pelas condições da via permanente. Quando é identificada alguma anomalia na via, a velocidade máxima do trem no trecho que apresenta a anomalia deve ser menor do que a VMA para garantir a segurança da operação e nesse caso impõe-se uma restrição de velocidade. Um dos desafios encontrados na manutenção da ferrovia é reduzir a quantidade de trechos da via permanente com restrições temporárias de velocidade. As restrições permanecem no trecho até que uma equipe da manutenção verifique e corrija a anomalia existente. Assim, as restrições podem prejudicar diretamente a capacidade de circulação da via. Um método que prediz a necessidade de imposição de uma restrição de velocidade em um trecho da ferrovia pode auxiliar o direcionamento de equipes de manutenção. Para a aplicação desse método, utilizou-se dados levantados da Estrada de Ferro Vitória a Minas (EFVM) em toda sua extensão e durante o período de 1 ano e 10 meses. Associando os dados históricos de restrições na via permanente com os dados históricos de medições de parâmetros da via, dados climáticos e a tonelagem bruta da via, foi criado um banco de dados total para realizar a predição. Esses dados foram utilizados como entrada para o treinamento de um algoritmo de *machine learning*: o *XGBoost*, que têm por objetivo prever a necessidade da imposição de uma restrição de velocidade na via no horizonte de 30 dias e 60 dias.

Palavras-chave: Restrição Temporária de Velocidade. Ferrovia. *XGBoost*. Base dados desbalanceada. Manutenção preventiva.

## **ABSTRACT**

In Brazil, rail transport is essential for the country's economy. The railroad is a rail transport system, consisting of permanent way and other fixed installations, traffic equipment and other essential accessories for the safe and efficient handling of passengers and/or cargo. According to the infrastructure and superstructure of the permanent way, the Maximum Authorized Speed (MAS) of a railway is defined on its geometric project. The MAS is the maximum speed of the project for each section of the permanent way, which can be changed by the conditions of the railway. When an anomaly is identified on the track, the maximum train speed in the section that presents an anomaly must be lower than the MAS to ensure the safety of the operation, in which case a Temporary Speed Restriction is imposed. One of the challenges encountered in railway maintenance is to reduce the number of sections of the permanent way with temporary speed restrictions. The restrictions remain on the section until a maintenance team checks and corrects an existing anomaly. Thus, restrictions can have negative impacts in the capacity of the route. A tool that predicts the occurrence of speed restrictions would be fundamental to head the maintenance teams. For the development of this tool, we used data collected from the Vitória - Minas Railway (EFVM) in its entire length and during the period of 1 year and 10 months. By associating the historical data of permanent way restrictions with the historical data of track geometry parameters, the weather conditions and the gross ton on the tracks, a total database was created to make a prediction. These data were used as input for the training of a machine learning algorithm: XGBoost, which aim to predict the need to impose a temporary speed restriction in the following 30 days or 60 days.

Key words: Temporary Speed Restriction. Railway. XGBoost. Unbalanced dataset. Preventive Maintenance.

## LISTA DE FIGURAS

|                                                                                          |    |
|------------------------------------------------------------------------------------------|----|
| Figura 1 - Superestrutura de uma ferrovia .....                                          | 9  |
| Figura 2 - Objetivo do projeto .....                                                     | 10 |
| Figura 3 - Imposição de uma restrição temporária de velocidade .....                     | 11 |
| Figura 4 - Carro-controle .....                                                          | 12 |
| Figura 5 - Funcionamento do boosting.....                                                | 18 |
| Figura 6 – Divisão do conjunto de dados por <i>k-fold cross validation</i> .....         | 23 |
| Figura 7 - Curva ROC para dois classificadores .....                                     | 26 |
| Figura 8 – Constituição da Via Permanente .....                                          | 29 |
| Figura 9 – Perfil do trilho TR-68 .....                                                  | 30 |
| Figura 10 – Desenho esquemático do desnivelamento longitudinal.....                      | 31 |
| Figura 11 – Desenho esquemático do desnivelamento transversal .....                      | 31 |
| Figura 12 – Desenho esquemático do empeno .....                                          | 32 |
| Figura 13 – Desenho esquemático do desalinhamento .....                                  | 32 |
| Figura 14 – Conceitos para o entendimento dos efeitos das restrições de velocidade ..... | 33 |
| Figura 15 - Possíveis causas de restrições.....                                          | 34 |
| Figura 16 – Mapa geral do traçado da EFVM.....                                           | 35 |
| Figura 17 – Carro-controle .....                                                         | 36 |
| Figura 18 – Localização do sistema de leitura do Carro Controle EM-80 - EFVM ....        | 36 |
| Figura 19 - Tela visualizada pelo operador do Carro Controle EM-80 .....                 | 38 |
| Figura 20 – Etapas do método proposto .....                                              | 39 |
| Figura 21 - Regiões não analisadas da Via Permanente .....                               | 41 |
| Figura 22 – Distribuição do número de medições ao longo do trecho estudado .....         | 42 |
| Figura 23 – PCA: variância explicada x componentes .....                                 | 46 |
| Figura 24 - Cenários avaliados.....                                                      | 50 |
| Figura 25 - Matriz de confusão normalizada - Teste 1 .....                               | 52 |
| Figura 26 - Matriz de confusão normalizada - Teste 2 .....                               | 52 |
| Figura 27 - Matriz de confusão normalizada - Teste 3 .....                               | 53 |
| Figura 28 - Matriz de confusão normalizada - Teste 4 .....                               | 53 |
| Figura 29 - Resultados dos modelos .....                                                 | 55 |

## LISTA DE TABELAS

|                                                                                    |    |
|------------------------------------------------------------------------------------|----|
| Tabela 1 – Matriz de confusão binária .....                                        | 24 |
| Tabela 2 - Tabela resumo referências bibliográficas.....                           | 27 |
| Tabela 3 – Identificação da local de medição em quilômetros .....                  | 42 |
| Tabela 4 – Amostra da base de dados do histórico de restrições de velocidade ..... | 43 |
| Tabela 5 – Amostra da base de dados geral.....                                     | 44 |
| Tabela 6 – Variáveis mais importantes das componentes principais .....             | 46 |
| Tabela 7 - Hiper parâmetros do XGBoost .....                                       | 49 |
| Tabela 8 - Resultados dos modelos preditivos.....                                  | 51 |

## SUMÁRIO

|                                                                       |           |
|-----------------------------------------------------------------------|-----------|
| <b>1. INTRODUÇÃO .....</b>                                            | <b>8</b>  |
| 1.1. OBJETIVOS .....                                                  | 13        |
| 1.1.1. Objetivo Geral.....                                            | 13        |
| 1.2.1. Objetivos Específicos .....                                    | 13        |
| 1.2. JUSTIFICATIVA .....                                              | 14        |
| 1.3. ESTRUTURA DA PROPOSTA DE DISSERTAÇÃO .....                       | 14        |
| <b>2. REFERENCIAL TEÓRICO .....</b>                                   | <b>16</b> |
| 2.1. ALGORITMOS DE <i>MACHINE LEARNING</i> APLICADOS A FERROVIAS .... | 16        |
| 2.2. CARACTERÍSTICAS DO <i>XGBOOST</i> .....                          | 18        |
| <b>3. DESCRIÇÃO DO PROBLEMA.....</b>                                  | <b>28</b> |
| 3.1. DESCRIÇÃO DA ESTRADA DE FERRO VITÓRIA-MINAS.....                 | 34        |
| <b>4. MÉTODO PROPOSTO .....</b>                                       | <b>39</b> |
| 4.1. LEVANTAMENTO E TRATAMENTO DE DADOS .....                         | 40        |
| <b>5. ALGORITMO PROPOSTO E RESULTADOS .....</b>                       | <b>49</b> |
| 5.1. DISCUSSÕES GERENCIAIS.....                                       | 54        |
| <b>6. CONCLUSÕES.....</b>                                             | <b>58</b> |
| 6.1. TRABALHOS FUTUROS .....                                          | 59        |
| <b>Referências .....</b>                                              | <b>60</b> |

## 1. INTRODUÇÃO

No Brasil, o transporte ferroviário é essencial para a economia do país. Em 2017, o volume total de cargas transportadas foi de 538,8 milhões de toneladas úteis enquanto o transporte de passageiros foi 1,2 milhão de pessoas (ANTT, 2018). O transporte de cargas seguro, confiável e eficiente é o objetivo desse modo de transporte. Inúmeras tecnologias avançadas e programas de segurança foram implementadas em diversas ferrovias para melhorar a eficiência ferroviária e o Ministério da Infraestrutura prevê investimentos de R\$ 30 bilhões para expandir a malha ferroviária do país nos próximos cinco anos (AGÊNCIA BRASIL, 2020).

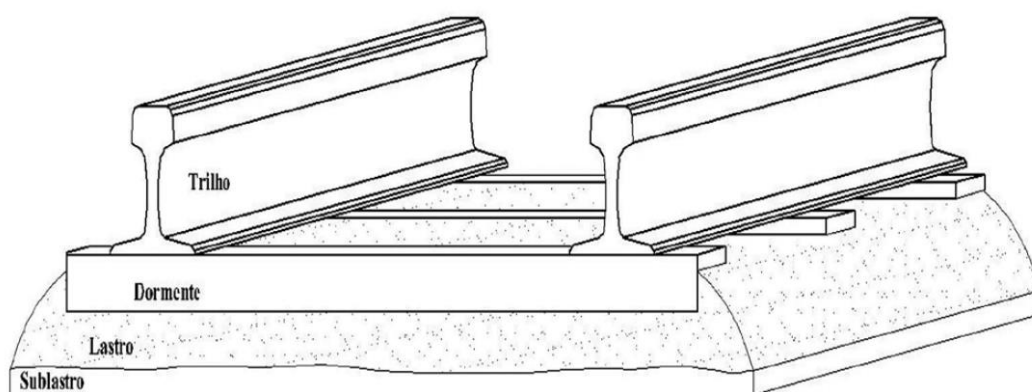
Nesse contexto, a ferrovia é um sistema de transporte sobre trilhos, constituído de via permanente e outras instalações fixas, como pátios e terminais, equipamentos de tráfego e demais acessórios indispensáveis à condução segura e eficiente de passageiros e/ou cargas. A via permanente (VP) é o conjunto das instalações e equipamentos que compõem as partes da via onde circulam os trens, citada também como linha férrea. Ela é composta de duas partes, infraestrutura e superestrutura (ROSA E RIBEIRO, 2017).

A infraestrutura é o conjunto de obras destinada a formar a plataforma da estrada e suportar a superestrutura, além de propiciar uma distribuição correta e uniforme dos esforços e impactos provenientes da superestrutura (DNIT, 2020). Ela é formada pelo conjunto de aterros, cortes, tuneis, pontes, drenagens etc., que resulta no que é denominado plataforma, ou seja, a superfície que vai receber a superestrutura (ROSA E RIBEIRO, 2017).

Já a superestrutura ferroviária (Figura 1) é a parte superior da via ferroviária, que suporta diretamente os esforços causados pela passagem dos veículos ferroviários e os transmite à infraestrutura. A superestrutura da via abrange o conjunto: trilhos, dormentes e lastro. O lastro é parte da superestrutura ferroviária, que distribui uniformemente na plataforma os esforços da via férrea transmitidos através dos dormentes, impedindo o deslocamento dos mesmos, oferecendo suficiente elasticidade à via, reduzindo impactos e garantindo-lhe eficiente drenagem e aeração (DNIT, 2020). O dormente é a peça de madeira, concreto, concreto protendido ou ferro, onde os trilhos são apoiados e fixados e que transmitem ao lastro parte dos esforços e vibrações produzidos pelos trens (DNIT, 2020). O trilho é um perfilado

metálico da seção transversal semelhante ao duplo T, com características de viga e assentada em fila dupla sobre dormente, que suporta e guia as rodas do veículo ferroviário e constitui a superfície de rolamento da via (DNIT, 2020).

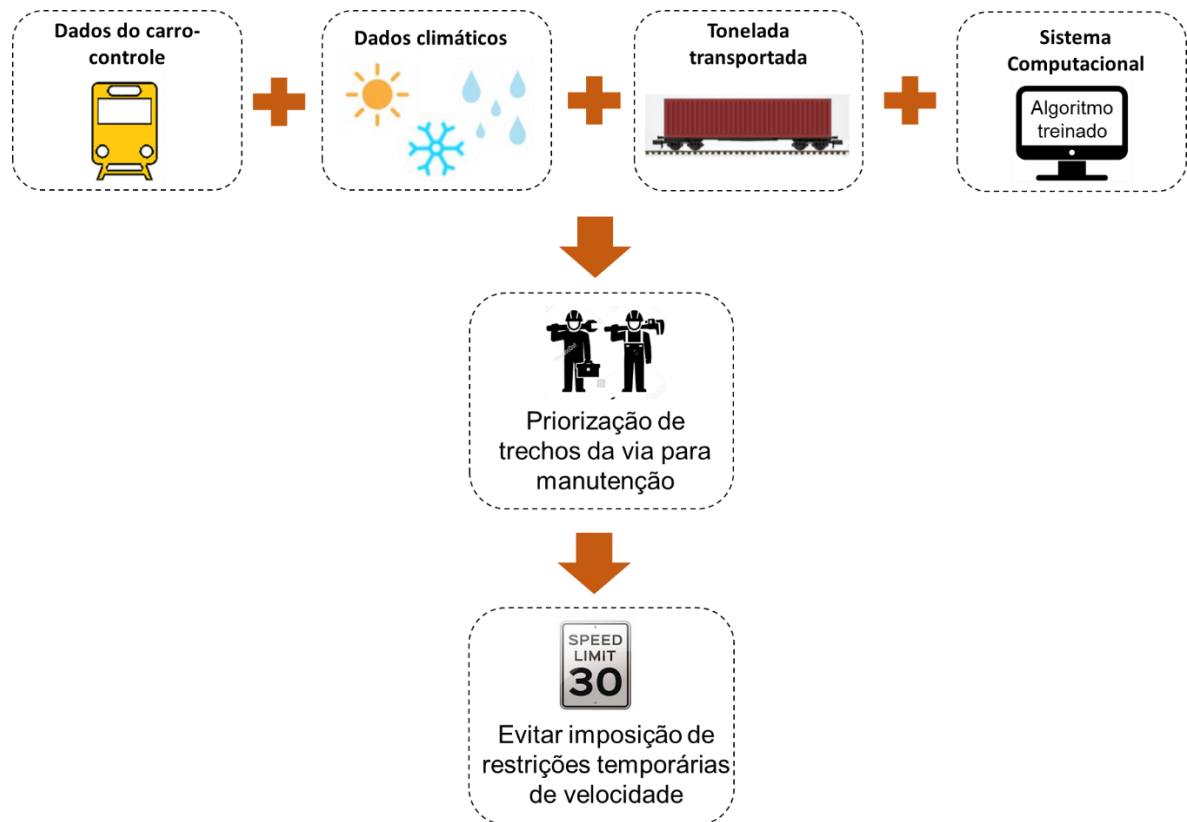
Figura 1 - Superestrutura de uma ferrovia



Fonte: Rosa e Ribeiro (2017).

As ferrovias estão constantemente buscando aumentar a quantidade de trens circulando na via. Uma das condições que leva à redução da quantidade de trens em circulação é a qualidade da via permanente, pois quando a via apresenta alguma imperfeição em sua geometria, a equipe de manutenção deve reduzir a velocidade máxima permitida para um trem. Essa redução impacta diretamente na quantidade de trens que podem circular na via e é denominada restrição temporária de velocidade. Nesse contexto, prever a necessidade de imposição dessas restrições no futuro auxiliaria significativamente o planejamento da manutenção preventiva da via permanente a fim de não haver reduções de velocidade na via (Figura 2).

Figura 2 - Objetivo do projeto



Fonte: Autor (2020).

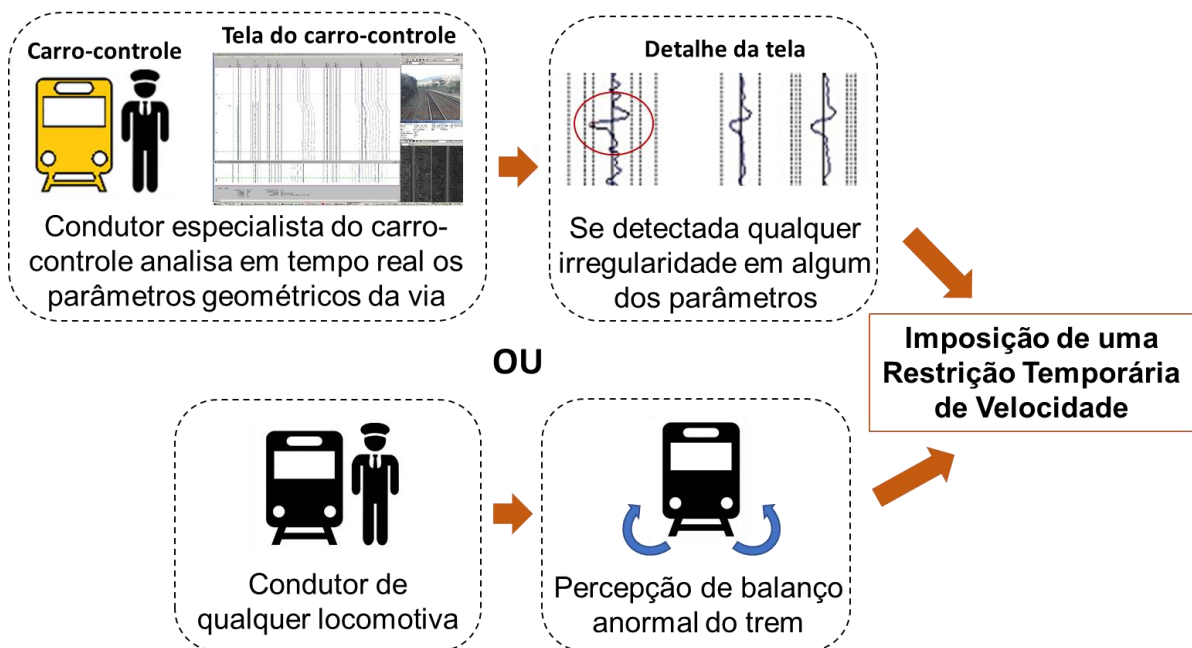
A priorização de trechos da ferrovia para a manutenção é uma atividade recorrente e que poderia ser facilitada por um modelo que prediz a necessidade da imposição de uma restrição de velocidade no futuro. Os resultados do modelo têm função de auxiliar a tomada de decisão dos especialistas visando evitar a ocorrência de restrições e melhorar a capacidade de circulação da via. Consequentemente, há um potencial ganho financeiro para a ferrovia.

As restrições temporárias de velocidade quando impostas, devem permanecer até que o reparo necessário seja realizado na via. Durante a vigência da restrição, há um impacto negativo na via, pois a redução de velocidade tende a causar aumento do tempo de viagem dos trens e consequentemente reduzir a capacidade de circulação de trens da ferrovia, afetando o cronograma de viagens da ferrovia. Portanto, devem ser impostas a menor quantidade de restrições possível, durando o menor tempo

possível, priorizando sempre a segurança e condições de operação do modo de transporte (VAN DER KOOIJ et. al, 2017).

As restrições de velocidade são impostas na via quando um maquinista de qualquer trem percebe algum balanço anormal da locomotiva ou quando são verificadas irregularidades na via enquanto são realizadas medições dos parâmetros da via (Figura 3).

Figura 3 - Imposição de uma restrição temporária de velocidade



A restrição temporária de velocidade pode ser caracterizada por parâmetros da via que são medidos pelo carro-controle (Figura 4). O carro-controle é um veículo sensoriado que se locomove sobre os trilhos realizando diversas medições de forma automatizada e a partir delas fornece os parâmetros da via. Uma restrição pode ser justificada por um parâmetro além do limite aceitável ou pela combinação de parâmetros associados, mesmo que cada um, individualmente, esteja dentro dos limites. Devido à complexidade da análise do comportamento individual e da interação entre os parâmetros, é importante recorrer a métodos computacionais para relacionar as medições da via com a ocorrência de restrições de velocidade.

Figura 4 - Carro-controle



Fonte: Autor (2019)

Para minimizar a imposição de restrições, a manutenção preventiva da via deve ser eficiente e os trechos suscetíveis a restrições futuras devem ser identificados. Usualmente, a manutenção preventiva é planejada seguindo protocolos com estimativas de vida útil dos componentes da via e a experiência dos especialistas, sem ter uma ferramenta computacional que analise o comportamento da via sinalizando possíveis problemas futuros.

Para o desenvolvimento de uma ferramenta computacional, utilizou-se dados obtidos da Estrada de Ferro Vitória a Minas (EFVM) em toda sua extensão e durante um período de 1 ano e 10 meses. Os dados referentes à geometria da via permanente foram obtidos utilizando o carro-controle, os dados climáticos foram obtidos do *Weather Channel* e as toneladas médias transportadas pela ferrovia foram obtidas por meio de relatórios de volumes transportados da EFVM.

Cabe ressaltar que os dados utilizados na realização desta dissertação foram obtidos da empresa Vale S.A. por intermédio do convênio com a Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES), número 75528452/2016. Nesse contrato, a empresa Vale S.A. autoriza a utilização de alguns dados operacionais para fins acadêmicos. A fim de viabilizar a publicação, os valores foram transformados por

uma função não divulgada com o objetivo de manutenção do sigilo dos dados apresentados.

Associando os dados históricos de restrições na via permanente com os dados históricos de medições do carro-controle, dados climáticos e a tonelagem bruta da via, foi criado um banco de dados cujos registros são compostos pelos parâmetros geométricos e características da via permanente, as informações históricas de restrições temporárias de velocidade, dados climáticos da ferrovia e a tonelagem bruta transportada na via permanente. Esses dados foram utilizados como entrada para o treinamento de um algoritmo de *machine learning*: *XGBoost*, que tem por objetivo prever a necessidade da imposição de uma restrição de velocidade na via.

## 1.1. OBJETIVOS

### 1.1.1. Objetivo Geral

Desenvolver um sistema inteligente utilizando o algoritmo *XGBoost* para prever a necessidade de imposição de restrições temporárias de velocidade nos trens a partir dos parâmetros geométricos da via permanente, da carga transportada na via permanente e das condições climáticas da ferrovia.

### 1.2.1. Objetivos Específicos

Os objetivos específicos dessa dissertação de mestrado são:

- Adquirir bases de dados das medições realizadas nos trilhos ferroviários por meio do carro controle para no período de 2017 e 2018, assim como da base histórica de registros de restrições de velocidade, dados climáticos e da tonelagem bruta transportada na via no mesmo período;
- Identificar as variáveis mais relevantes para a caracterização de uma restrição temporária de velocidade na via permanente;
- Investigar o algoritmo *XGBoost* para o problema e analisar suas configurações para eleger as que apresentam os melhores resultados;

- Aplicar o algoritmo *XGBoost* na base de dados reais da Estrada de Ferrovia Vitória a Minas (EFVM) para avaliar os cenários propostos, realizando o treinamento, validação e teste e avaliando métodos de balanceamento para a predição da necessidade de imposição de restrições temporárias de velocidade na via.

## 1.2. JUSTIFICATIVA

Na literatura científica, encontram-se vários estudos realizados referentes a utilização dos algoritmos de *machine learning* para as mais diversas áreas do conhecimento. Há pesquisas que mostram a utilização algoritmos de *machine learning* para justificar e detectar falhas na ferrovia, assim como analisar seu comportamento. Entretanto, para a predição da necessidade de imposição de restrições temporárias de velocidade na via permanente, não foi identificada nenhuma pesquisa. Logo, esta pesquisa visa preencher essa lacuna, tornando-se uma oportunidade para o desenvolvimento e avanço neste campo de pesquisa.

Além disso, as restrições temporárias de velocidade na via permanente podem impactar de forma negativa a circulação de trens na ferrovia e atualmente a manutenção preventiva da via para evitá-las é planejada com base na experiência da equipe de planejamento. Nesse contexto, o uso de algoritmos de *machine learning* como ferramenta de suporte para o planejamento da manutenção pode trazer benefícios econômicos para a empresa responsável pela ferrovia. O método desenvolvido nessa dissertação identificará, em um horizonte de 30 dias e 60 dias, os trechos nos quais haverá necessidade de imposição de restrições temporárias de velocidade, orientando a alocação de recursos para reparos na via, mantendo a confiabilidade e a segurança operacional da ferrovia.

## 1.3. ESTRUTURA DA PROPOSTA DE DISSERTAÇÃO

Para atender aos objetivos desta dissertação, o conteúdo desenvolvido foi estruturado em 6 capítulos, incluindo este primeiro que é uma introdução ao problema, e os demais distribuídos da seguinte maneira:

O Capítulo 2 apresenta a revisão da literatura sobre o *XGBoost* e uma revisão bibliográfica sobre a aplicação de machine learning em desafios envolvendo ferrovias.

O Capítulo 3 descreve o problema e onde ele ocorre.

O Capítulo 4 apresenta o levantamento e tratamento dos dados históricos para serem utilizados nesta dissertação.

O Capítulo 5 apresenta a arquitetura proposta para o algoritmo *XGBoost*, o treinamento, validação e teste do algoritmo nos cenários elaborados e os resultados alcançados, incluindo discussões gerenciais.

O Capítulo 6 apresenta as conclusões e trabalhos futuros.

## 2. REFERENCIAL TEÓRICO

Neste capítulo são apresentados os principais conceitos referentes ao *XGBoost*, à técnicas de pré-processamento de dados e o referencial teórico existente sobre a utilização de *machine learning* em ferrovias.

### 2.1. ALGORITMOS DE *MACHINE LEARNING* APLICADOS A FERROVIAS

O objetivo primordial da ciência de dados é o aprimoramento da tomada de decisão, o que geralmente provoca interesse direto para empresas (FAWCETT e PROVOST, 2016). Os algoritmos de *machine learning*, que estão inseridos na ciência de dados, podem gerar em sua saída boas sugestões e padrões dificilmente percebidos sob o olhar humano (KUBAT, 2017). Nesse contexto, diversos algoritmos de *machine learning* já foram utilizados para analisar comportamentos de ferrovias sob diferentes perspectivas.

No âmbito de detecção de ocorrência de falhas nos componentes dos trens, Krummenacher et al. (2017) propuseram dois métodos de *machine learning* para detectar automaticamente defeitos nas rodas dos trens com base na força vertical aplicada na roda, medida por sensores permanentes instalados na linha férrea. Assim, a partir de uma quantidade massiva de dados gerados automaticamente pelos sensores na ferrovia, foram treinados dois modelos: um classificador com base no *Support Vector Machine* (SVM) para classificar se a roda possui defeito ou não, e uma segunda abordagem, com redes neurais artificiais, para também detectar defeitos nas rodas, porém analisando imagens em duas dimensões dos gráficos gerados pelos sensores de força na via. Krummenacher et al. (2017) demonstraram o potencial do uso dados de sensores na linha férrea alcançando uma acurácia de até 92% na detecção de defeitos na via.

Para predição de defeitos ocorridos na ferrovia, Mohammadi et al. (2019) utilizaram medições geométricas e tonelagem da via para identificar os parâmetros que mais contribuem para a ocorrência de defeitos na via permanente. Identificados os parâmetros por meio de algoritmos de eliminação de variáveis, é utilizado o algoritmo *XGBoost* para realizar predições na via. Assim como Mohammadi et al. (2019), esta dissertação utiliza os parâmetros geométricos da via e a tonelagem bruta transportada, porém com outro propósito: o da predição de restrições de velocidade.

No mesmo contexto e usando dados de geometria da via de uma via de uma milha de amostra, Lasisi e Attoh-Okine (2018) demonstraram como combinar esses parâmetros reduzindo a dimensão da quantidade de variáveis para simplificar as análises de defeitos na via sem perder muita variabilidade nos dados. Foi aplicada a Análise de Componentes Principais para realizar a seleção dos parâmetros mais relevantes para a identificação de defeitos na via e após a seleção foram testados três algoritmos para a predição de defeitos: *Linear Discriminant Analysis*, *Support Vector Machine* e *Random Forest*. A extensão estudada foi dividida em trechos 150m que foram classificados de acordo com a presença ou não de defeitos e essa base de dados foi utilizada para a predição. Para avaliar os algoritmos foram utilizadas as taxas de Verdadeiro Positivo e de Falso Positivo. Os resultados mostraram que três componentes principais foram melhores na previsão de defeitos e com uso potencial na manutenção viária.

Lee et al. (2020) realizam a predição do *track quality index* (TQI), um indicador da via, para o planejamento da manutenção. O TQI é um indicador numérico, então eles utilizaram dois modelos de regressões diferentes para realizar a estimativa: *Online Support Vector Regression* e *Incremental Support Vector Regression*. Para a estimativa, são utilizados dados de tonelagem bruta passada na via, tipo da estrutura, curvatura, velocidade da via e desvio padrão dos dados de geometria da via por dois anos, como esta pesquisa. Os resultados apresentados mostram que a predição pode ser utilizada para priorização de trechos para manutenção.

Para aumentar a velocidade média de uma ferrovia sem comprometer a segurança viária, Li et al. (2014) utilizaram um algoritmo customizado de SVM para criar alarmes críticos dos parâmetros medidos pelo carro-controle. Usando esses parâmetros de entrada, é aplicada uma árvore de decisão para realizar a predição desses alarmes na via permanente de uma ferrovia americana em até 7 dias para direcionar a manutenção preventiva da via. Diferentemente de Li et al. (2014), que direcionam suas predições para os alarmes da via, esta pesquisa realiza a predição da necessidade de imposição de restrições de velocidade na via em um horizonte de 30 dias.

Não foi encontrado nenhum artigo na revisão bibliográfica que realiza a predição da necessidade de imposição de restrições temporárias de velocidade na via permanente em um horizonte de tempo de um mês.

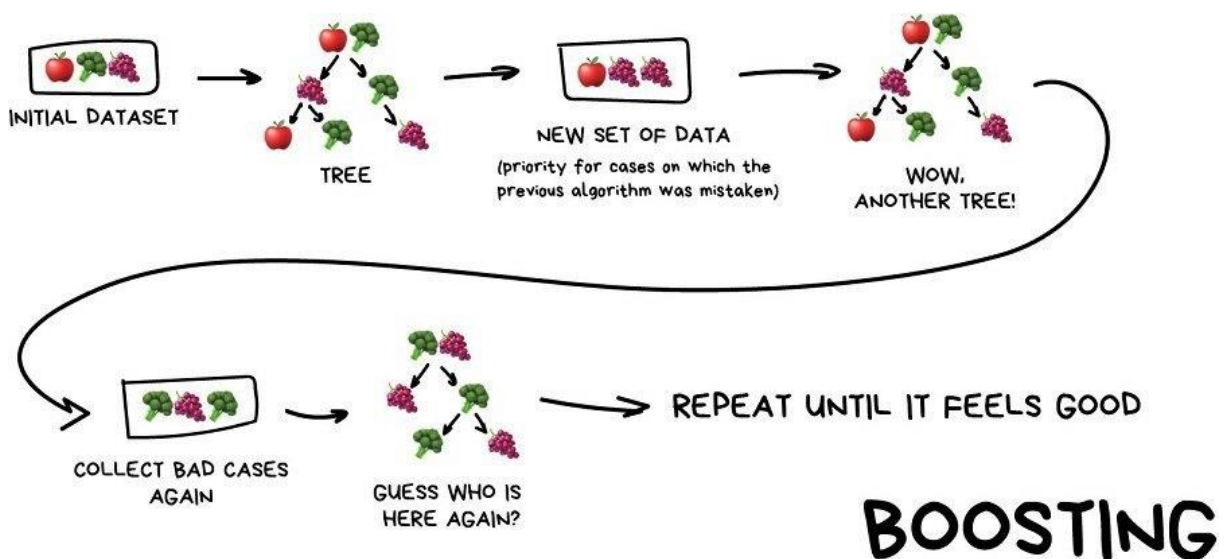
## 2.2. CARACTERÍSTICAS DO XGBOOST

O algoritmo proposto nesta dissertação é de aprendizagem supervisionada, isso é, ele possui um alvo específico a ser determinado (FAWCETT e PROVOST, 2016). Segundo Fawcett e Provost (2016), algoritmos supervisionados são geralmente utilizados para resolver problemas de classificação, como o desta dissertação, regressão e modelagem casual.

O objetivo desta dissertação é classificar, de maneira binária, se haverá necessidade de imposição de restrições temporárias de velocidade na via permanente. Assim, dentre os algoritmos de classificação, o *eXtreme Gradient Boosting* (XGBoost) foi selecionado para realizar a predição. O XGBoost é um algoritmo baseado em árvore de decisão que utiliza uma estrutura de *gradient boosting* para o aprimoramento do seu desempenho. O *gradient boosting* se refere ao processo de aprendizado relativo à construção de modelos sequenciais, convergindo em um modelo forte após seus auto ajustes a cada modelo adicionado. Entre os métodos de aprendizado de máquina utilizados na prática, o *gradient tree boosting* é uma técnica que possui resultados positivos em muitas aplicações (CHEN e GUESTRIN, 2016).

No XGBoost a predição realizada para um dado exemplo é a soma das predições fornecidas por cada árvore de decisão sequenciada no modelo (CHEN e GUESTRIN, 2016). A Figura 5 mostra como a técnica de *boosting* funciona.

Figura 5 - Funcionamento do *boosting*



Fonte: Pedro Gomes (2019).

As características que fazem do *XGBoost* um algoritmo robusto é sua otimização de paralelismo e hardware (CHEN e GUESTRIN, 2016). Além disso, esse algoritmo, diferente dos modelos de árvores de decisão puramente, inclui um termo de regularização para evitar *overfitting* e suporta diferentes equações de penalização (MITCHELL e FRANK, 2017).

Assim como outros algoritmos, o *XGBoost* possui hiper parâmetros que não são estimados pelo próprio modelo e que devem ser determinados adequadamente para que sua performance seja a melhor possível. Os principais hiper parâmetros a serem definidos são: taxa de aprendizado, profundidade máxima da árvore, número de boost, taxa de subamostra, peso mínimo de um nó filho, etapa delta máxima e gama (XIA et al., 2017).

De acordo com Chen e Guestrin (2016), a taxa de aprendizado pode variar de 0 a 1 e tem a função de diminuir a importância individual das árvores, desacelerando o aprendizado e deixando mais espaço para futuras árvores. A profundidade máxima da árvore limita o tamanho das árvores, auxiliando na redução de *overfitting*. O peso mínimo de um nó filho limita o valor da soma dos pesos das folhas da árvore para que aconteça uma divisão. A etapa delta máxima limita o peso máximo de cada iteração. E o gama é um valor que é subtraído da função de ganho do algoritmo, estabelecendo um valor mínimo de ganho a cada segmentação da árvore.

Para a determinação desses hiper parâmetros foi utilizado um algoritmo de otimização Bayesiana. A otimização Bayesiana busca encontrar um mínimo aproximado para funções que requerem muito tempo para serem avaliadas (FRAZIER, 2018). Ela utiliza o processo Gaussiano para estimar uma função genérica e ao avaliar o mínimo dessa função, ela determina novos pontos a serem testados. Com o processo Gaussiano, a cada novo teste a função genérica é aprimorada e busca-se chegar ao mínimo global da função até que algum critério de parada seja alcançado (MARTINEZ-CANTIN, 2014). No contexto de otimização de parâmetros de algoritmos de *machine learning*, a função a ser minimizada deve ser associada à uma métrica de avaliação do modelo (KLEIN et al., 2017).

Visando alcançar resultados satisfatórios nesses treinamentos, é necessário que o conjunto de dados utilizado seja devidamente tratado e seccionado (PUTRA e SAFRILAH, 2017). Assim sendo, foram aplicadas as seguintes técnicas no conjunto

de dados utilizado: normalização dos dados; seleção de variáveis de entrada; e o balanceamento da base de dados. Todos os algoritmos utilizaram o mesmo conjunto de dados tratado e seccionado e foram avaliados pelas mesmas métricas (Área sob a curva ROC (AUC), matriz de confusão, taxas de Verdadeiro Positivo e de Falso Positivo). A seguir serão apresentadas todas as técnicas citadas anteriormente e, posteriormente, cada um dos dois algoritmos será detalhado.

A análise de dados de alta dimensão é um desafio na área de *machine learning*. A seleção de variáveis fornece uma maneira eficaz de resolver esse problema removendo dados irrelevantes e redundantes, possibilitando reduzir o tempo de computação, melhorar a precisão do aprendizado e facilitar um melhor entendimento do modelo (CAI et al., 2018). A seleção de variáveis é o processo de obtenção de um subconjunto do conjunto original de variáveis do problema. Ela, a partir de um critério, seleciona as variáveis relevantes do conjunto de dados. Mohammadi et al. (2019) compararam a performance de duas técnicas de seleção de variáveis para identificar os parâmetros da geometria da via que mais contribuem para a predição de defeitos na via. Eles aplicaram *Recursive Feature Elimination* (RFE) e *Singular Value Decomposition* (SVD) para fazer a seleção de variáveis dentre mais de 60 parâmetros da geometria da via.

Para a seleção de variáveis nesta dissertação foi utilizado o *Principal Component Analysis* (PCA), assim como Fink, Zio e Weidmann (2013) aplicaram para dados de parâmetros geométricos de ferrovias. O PCA pode ser aplicado para reduzir a dimensionalidade de um conjunto de dados com muitas variáveis correlacionadas, sem perda significativa de informações, projetando os dados originais sobre um novo subespaço dimensional com ordem inferior e que contém a maioria dos recursos originais (DING, 2014). Segundo Fink, Zio e Weidmann (2013), os dados originais são convertidos por uma transformação ortogonal em um conjunto de variáveis não correlatas denominadas como componentes principais sem induzir qualquer perda significativa na variância dos dados em comparação com o conjunto de dados original. Essas componentes principais possuem uma independência maximal entre si e são a nova representação do *dataset*. Então essa técnica foi utilizada para reduzir o número de variáveis de entrada presentes na base de dados elaborada nesta dissertação e buscou-se um limite mínimo de variância explicada de 95% (PROVEZZA et al., 2021).

Antes de aplicar o PCA, é necessário normalizar as variáveis, para evitar que os algoritmos ficassem enviesados para as variáveis com maior ordem de grandeza (PROVEZZA et al., 2021). O método do *z-score* foi utilizado para realizar a normalização. A Equação (1) apresenta o cálculo do *z-score*, na qual  $x$  representa um valor da variável,  $\mu$  representa a média dos valores daquela variável e  $\sigma$  o desvio-padrão.

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

O desbalanceamento do *dataset* é outro desafio comum dos modelos classificadores de *machine learning*. Ele ocorre quando há uma grande desproporção do número de registros de uma das classes da base de dados. Como consequência, os classificadores tendem a proporcionar um grave desequilíbrio no grau de precisão, com a classe majoritária com acurácia perto de 100 por cento e a classe minoritária tendo acurácia de 0-10 por cento, por exemplo (CHAWLA et al., 2002). Por isso, é necessário um classificador que dará alta acurácia para a classe minoritária sem comprometer gravemente a acurácia da classe majoritária. Técnicas de balanceamento da base de dados são utilizadas para melhorar a performance dos modelos nessa situação.

A abordagem *sampling* tem sido amplamente usada para lidar com *datasets* desbalanceados. *Sampling* é um pré-processamento que visa minimizar as discrepâncias entre as classes por meio de uma reamostragem, isso é, balanceamento artificial do conjunto de treinamento. As categorias mais comuns de *sampling* são o *oversampling* e o *undersampling* (HASANIN et al., 2019). O *oversampling* cria observações da classe minoritária a partir das informações contidas nos dados originais. Essa geração de novas entradas pode ser feita aleatoriamente, *random oversampling*, ou sinteticamente com o auxílio de técnicas de *clustering*. O *undersampling* reduz o desbalanceamento do *dataset* focando na classe majoritária. Ou seja, elimina registros da classe com maior número de ocorrências. Além dessas duas técnicas de *sampling*, existem também os *advanced sampling methods*, como SMOTE e ADASYN, nos quais instâncias são adicionadas ou removidas de modo adaptativo.

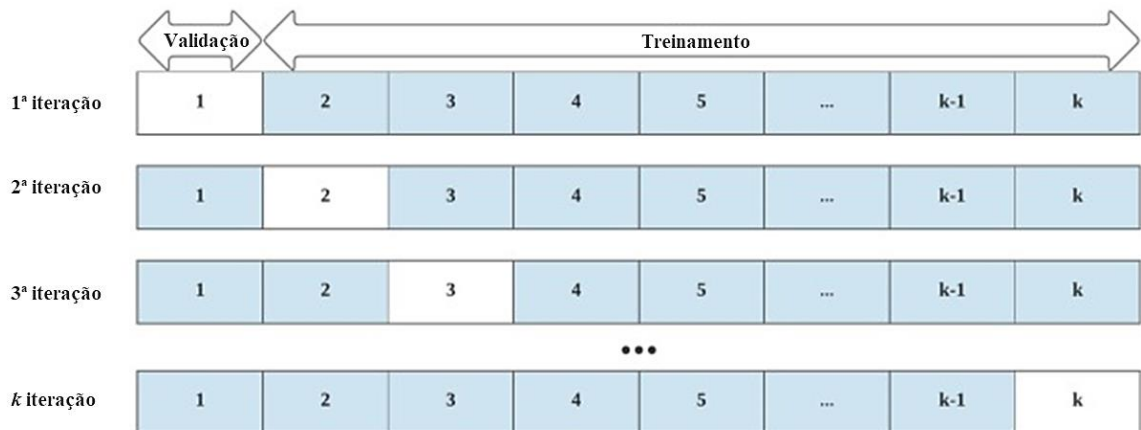
Chawla et al. (2002) desenvolveram um método diferente de se fazer *oversampling* da classe minoritária, que consiste na geração de casos artificiais para a classe de interesse a partir dos casos já existentes. Esses novos casos serão gerados na vizinhança de cada caso da classe minoritária, baseado em kNN, de forma a fazer crescer a região de decisão e, assim, aumentar o poder de generalização dos classificadores gerados para estes dados. Esse novo método é denominado SMOTE (do inglês, *Synthetic Minority Oversampling Technique*). Esses novos casos sintéticos são interpolações ao longo do segmento de reta que liga cada caso da classe minoritária a um de seus  $k$  vizinhos mais próximos, escolhidos de forma aleatória (CHAWLA et al., 2002). O ADASYN é uma técnica de balanceamento que aprimora o SMOTE. Nela, a quantidade de casos gerados é adaptativa e depende da dificuldade de aprendizado das amostras da classe minoritária, isso é, a partir de amostras que são mais difíceis de serem aprendidas são gerados mais casos artificiais (HE et al., 2008). Wang, Xi et al. (2019) mostraram como o balanceamento dos dados pode ser efetivo na melhoria do desempenho de classificadores. Eles desenvolveram um algoritmo de balanceamento para aplicação em um contexto ferroviário: determinar o momento de acionamento do freio pneumático de trens de carga, mostrando que em algumas situações, as técnicas de balanceamento já desenvolvidas podem ser aprimoradas.

Para aplicar as técnicas anteriormente mencionadas, o conjunto de dados foi dividido inicialmente em duas bases: a base de treinamento e validação e a base de teste. A proporção utilizada para essa divisão foi 80/20 e na base de treinamento e validação, foi utilizada a validação cruzada. A validação cruzada é uma técnica para avaliar a capacidade de generalização de um algoritmo, ou seja, a capacidade de responder corretamente a dados que não foram utilizados no processo de treinamento. Essa técnica é empregada em problemas nos quais o objetivo dos algoritmos é a classificação. Na validação cruzada, o conjunto de dados é dividido de forma aleatória em partes, estabelecendo-se partes para a realização do treinamento e partes para a realização da validação (SILVA, ALMEIDA e YAMAKAMI, 2012).

No método de validação cruzada, uma das técnicas utilizadas é o *k-fold cross validation*, na qual um conjunto de dados é dividido em  $k$  partes (*folds*) de forma aleatória, sendo essas partes divididas em partes de treinamento e partes de validação. Um classificador é treinado usando  $k-1$  partes e a validação desse

treinamento é feita com uma parte. Esse processo é realizado  $k$  vezes alternando de forma circular as partes estabelecidas para validação, como mostra a Figura 6.

Figura 6 – Divisão do conjunto de dados por  $k$ -fold cross validation



Fonte: Xiong (2020).

Para cada iteração do  $k$ -fold, é gerado um treinamento do algoritmo que é armazenado e a partição de validação é avaliada. Após as  $k$  iterações, são analisadas as métricas de avaliação das partições de validação, sendo que o treinamento referente à iteração cujas métricas estão mais próximas da média aritmética de todas as iterações é selecionado. Esse treinamento é utilizado posteriormente para realizar a predição da partição de teste, cujo conjunto de dados ainda não foi apresentado para os algoritmos.

A quantidade de partes na qual a base de treinamento e validação é dividida depende da quantidade de registros da base. Usualmente, o valor de  $k$  varia de 5 a 10 (XIONG et al., 2020) e quanto maior a base, maior o valor de  $k$  nesse intervalo. É importante ressaltar que a quantidade de registros disponível no conjunto de dados de treinamento e validação seja múltipla da divisão em partes que se deseja realizar, para que se possa obter resultados satisfatórios dos algoritmos.

A utilização da validação cruzada também evita a ocorrência de *overfitting*, estimando o erro de generalização do algoritmo durante o processo de treinamento. O *overfitting*

ocorre quando os algoritmos memorizam os padrões de treinamento, deixando de extrair características do modelo para se tornarem eficientes na generalização dos padrões, ou seja, para acertar padrões que os algoritmos não observaram durante o processo de treinamento (DUDA, HART e STORK, 2012; SILVA, ALMEIDA e YAMAKAMI, 2012).

Para avaliar o desempenho de todos os algoritmos propostos, foram utilizadas as seguintes métricas de avaliação: matriz de confusão, Área sob a curva ROC (AUC), taxas de Verdadeiro Positivo e de Falso Positivo. A escolha dessas métricas foi devido à distribuição desigual da variável alvo, que é uma variável binária que representa se haverá necessidade de impor uma restrição em um trecho de 100 metros.

Para classificação binária, as classificações são apresentadas na forma de uma matriz bidimensional com uma linha e uma coluna para cada classe. Cada elemento desta matriz, chamada matriz de confusão, apresenta o número de casos avaliados, na qual a classe real é a linha e a classe predita é a coluna. Bons resultados são caracterizados por valores altos na diagonal principal e valores nulos para elementos fora da diagonal principal. A Tabela 1 ilustra a nomenclatura utilizada para as quatro situações possíveis nessa matriz de confusão.

Tabela 1 – Matriz de confusão binária

|                 | Estimação positiva       | Estimação negativa       |
|-----------------|--------------------------|--------------------------|
| Classe positiva | Verdadeiro Positivo (VP) | Falso Negativo (FN)      |
| Classe negativa | Falso Positivo (FP)      | Verdadeiro Negativo (VN) |

Os Verdadeiro Positivo (VP) e Verdadeiro Negativo (VN) são os casos nos quais a predição é correspondente à realidade. O Falso Positivo (FP) representa os casos nos quais a predição aponta uma estimativa positiva enquanto a classe é na realidade negativa. O contrário ocorre no Falso Negativo (FN), a estimativa é negativa enquanto a classe é na realidade positiva. De acordo com cada problema, a importância do

Falso Negativo e do Falso Positivo pode variar, isso é, há situações nas quais a pior circunstância é FP enquanto em outras é o FN.

A Taxa de Verdadeiro Positivo (TPR) a proporção de casos positivos que foram identificados corretamente, como mostra a Equação (2). O intervalo é de 0 a 1. Um valor maior indica melhor previsão preditiva.

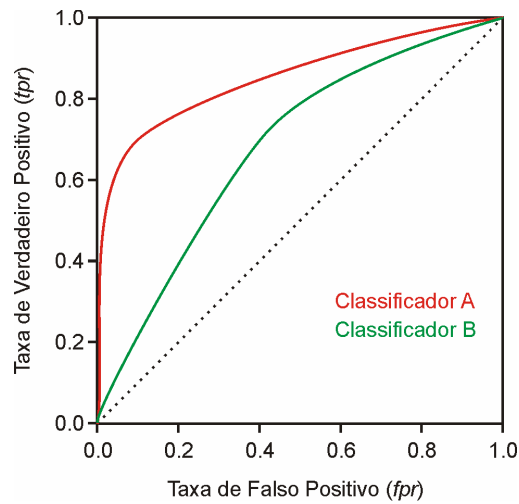
$$TPR = \frac{VP}{VP + FN} \quad (2)$$

A Taxa de Falsos Positivos (FPR) mede a taxa de alarmes falsos ou a fração de negativos reais previstos como positivos, como mostra a Equação (3). O intervalo é de 0 a 1. Um valor menor indica melhor previsão preditiva.

$$FPR = \frac{FP}{FP + VN} \quad (3)$$

Um gráfico ROC é um gráfico bidimensional, no qual a TPR é plotada no eixo das ordenadas e a taxa de falso alarme (FPR) é plotada no eixo das abscissas. Um classificador perfeito corresponderia a uma linha horizontal no topo do gráfico, porém esta dificilmente será alcançada. Na prática, curvas consideradas boas estarão entre a linha diagonal e a linha perfeita, onde quanto maior a distância da linha diagonal, melhor o sistema (FAWCETT, 2004). A linha diagonal indica uma classificação aleatória, ou seja, um sistema que aleatoriamente seleciona saídas como positivas ou negativas, como jogar uma moeda para cima e esperar cara ou coroa.

Figura 7 - Curva ROC para dois classificadores



Fonte: Fawcett, 2004.

A partir da curva ROC é calculado a *Area Under the Curve* (AUC), que é uma medida padrão para a comparação de algoritmos. O AUC pode ser obtido por métodos de integração numérica, e teoricamente, quanto maior a AUC, melhor o sistema. O valor máximo para essa métrica é 1.

A Tabela 2 sintetiza os principais artigos aplicados à ferrovia utilizados como base para essa dissertação. Apenas um artigo com o tema ferroviário apresentou todas as técnicas que serão aplicadas nessa dissertação, porém com o objetivo distinto dessa dissertação.

Tabela 2 - Tabela resumo referências bibliográficas

| Título do artigo                                                                                                           | Autores                | Periódico                                               | Ano  | Objetivo                                                                                                               | Técnicas de balanceamento | Técnicas de seleção de variáveis                                          | Validação cruzada       | Algoritmos testados                                                      | Métricas de avaliação                                                                                     |
|----------------------------------------------------------------------------------------------------------------------------|------------------------|---------------------------------------------------------|------|------------------------------------------------------------------------------------------------------------------------|---------------------------|---------------------------------------------------------------------------|-------------------------|--------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|
| Wheel defect detection with machine learning                                                                               | Krummenacher et al.    | IEEE Transactions on Intelligent Transportation Systems | 2017 | Detectar automaticamente defeitos nas rodas dos trens com base na força vertical aplicada na roda do trem              | Randomly over-sampling    | -                                                                         | 3-fold cross-validation | SVM e RNA                                                                | Acurácia, precision e recall                                                                              |
| Exploring the impact of foot-by-foot track geometry on the occurrence of rail defects                                      | MOHAMMADI, Reza et al. | Transportation research part C: emerging technologies   | 2019 | Identificar os parâmetros que mais contribuem para a ocorrência de defeitos na via permanente e prever defeitos na via | ADASYN                    | Recursive Feature Elimination (RFE) e Singular Value Decomposition (SVD). | 5-fold cross-validation | XGBoost                                                                  | True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR) e False Negative Rate (FNR) |
| Improving rail network velocity: A machine learning approach to predictive maintenance                                     | Li et al.              | Transportation research part C: emerging technologies   | 2014 | Criar alarmes críticos dos parâmetros medidos pelo carro-controlado e realizar a predição desses alarmes               | -                         | Análise de Componentes Principais                                         | 5-fold cross-validation | SVM e árvore de decisão                                                  | True positive rate (TPR) e False positive rate (FPR)                                                      |
| Principal components analysis and track quality index: A machine learning approach                                         | Lasasi e Attoh-Okine   | Transportation Research Part C                          | 2018 | Selecionar os parâmetros da via que mais impactam no TQI e usá-los para predição de defeitos na via                    | -                         | Análise de Componentes Principais                                         | -                       | Linear Discriminant Analysis, Support Vector Machine e Random Forest     | True positive rate (TPR) e False positive rate (FPR)                                                      |
| Deterioration Prediction of Track Geometry Using Periodic Measurement Data and Incremental Support Vector Regression Model | Jun S. Lee et al.      | Journal of Transportation Engineering, Part A: Systems  | 2020 | Verificar a eficiência do uso do OSVR e do ISVR para predição da deterioração dos parâmetros da via                    | -                         | -                                                                         | -                       | Online Support Vector Regression e Incremental Support Vector Regression | MSE                                                                                                       |
| Intelligent operation of heavy haul train with data imbalance: A machine learning method                                   | Wang, Xi et al.        | Knowledge-Based Systems                                 | 2019 | Uso de técnicas de machine learning para determinar o momento de acionamento do freio pneumático de trens de carga     | EasyEnsemble.M algorithm  | -                                                                         | -                       | ANN, UnderANN, OverANN, AdaBoost, UnderBoost, OverBoost, EE, EMKD        | Matriz de confusão, precision, recall e F-score                                                           |

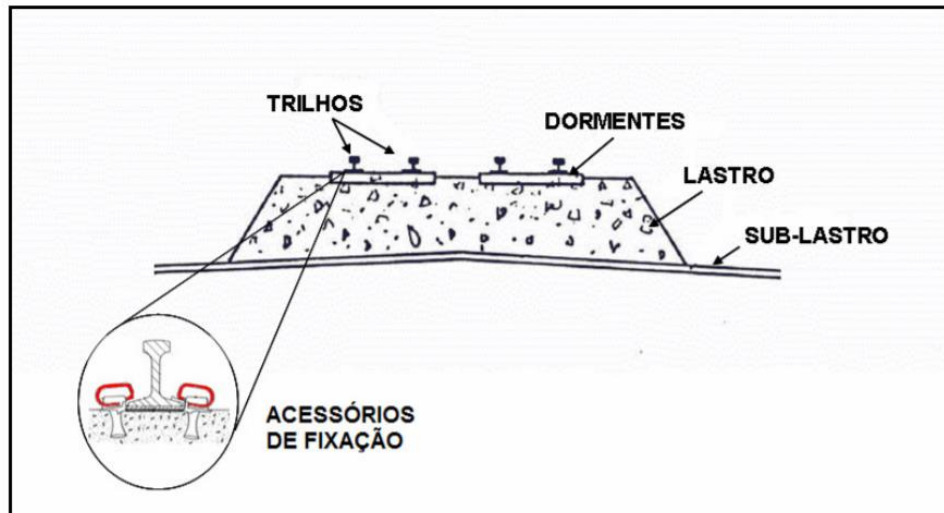
### 3. DESCRIÇÃO DO PROBLEMA

A disponibilidade da via férrea é um fator que impacta diretamente a operação ferroviária. Interrupções em trechos da via, manutenções e restrições temporárias de velocidade afetam sua disponibilidade e, portanto, busca-se evitá-las ou programá-las previamente para produzir o menor impacto no cronograma de viagem dos trens. Nesse contexto, visando aumentar disponibilidade da via férrea, almeja-se evitar a imposição de restrições temporárias de velocidade em trechos da via. Assim, é proposto um sistema computacional para auxiliar a predição da necessidade de imposição restrições de velocidade em trechos da via, direcionando as equipes de manutenção para evitá-las. O sistema computacional proposto utiliza um algoritmo de *machine learning* treinado com dados históricos da via para realizar a predição.

As restrições temporárias de velocidade são impostas quando são identificadas anomalia na via visando garantir a segurança ferroviária. Essas anomalias podem ser provenientes de defeitos na infraestrutura ou superestrutura e são refletidas nos parâmetros geométricos da via.

A infraestrutura é o complexo de obras destinada a formar a plataforma da estrada e suportar a superestrutura, além de propiciar uma distribuição correta e uniforme dos esforços e impactos provenientes da última. As características geométricas da infraestrutura se baseiam na planta (eixo), perfil longitudinal (greide) e seção transversal (ROSA e RIBEIRO, 2017). Já a superestrutura ferroviária é a parte superior da via ferroviária, que suporta diretamente os esforços causados pela passagem dos veículos ferroviários e os transmite à infraestrutura. Segundo Rives et al. (1977), a Via Permanente é constituída por duas vigas metálicas contínuas longitudinais denominadas trilhos, fixados aos apoios transversais espaçados regularmente, denominados dormentes que compõem a grade, que repousam sobre um colchão amortecedor de material granular, denominado lastro que por sua vez, absorve e transmite ao solo as pressões correspondentes às cargas suportadas pelos trilhos (Figura 8).

Figura 8 – Constituição da Via Permanente



Fonte: Rives (1977, p. 6).

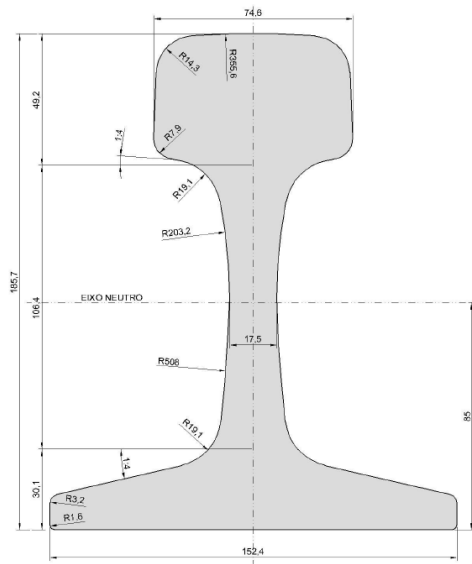
De acordo com a infraestrutura e a superestrutura de uma via permanente, no projeto geométrico da via é definida a Velocidade Máxima Autorizada (VMA). A VMA é velocidade máxima de projeto para cada trecho da via, podendo ser alterada pelas condições da via permanente, material rodante e segurança operacional. Quando é identificada alguma anomalia na via, a velocidade máxima do trem no trecho que apresenta a anomalia deve ser menor do que a VMA para garantir a segurança da operação. Quando a velocidade máxima de tráfego na via é reduzida a um valor abaixo da VMA, tem-se uma restrição temporária de velocidade.

As restrições temporárias de velocidade podem comprometer a capacidade de circulação de trens da via, podendo provocar atrasos no cronograma da ferrovia. Assim, as restrições podem impactar a confiabilidade da ferrovia e por isso sua imposição deve ser minimizada.

Como já apresentado na Figura 8, a Via Permanente é constituída por trilhos, acessórios de fixação dos trilhos, dormentes, lastro, e em alguns casos pelo sublastro. O carro-controle, que é descrito posteriormente, realiza as medições nos trilhos para calcular os parâmetros da via, incluindo a bitola. Os trilhos utilizados na EFVM são do tipo TR-68 (ABNT, 2012). A EFVM possui aproximadamente 120 mil toneladas de

trilho assentados em sua via férrea. A Figura 9 apresenta as dimensões do trilho ferroviário TR-68.

Figura 9 – Perfil do trilho TR-68

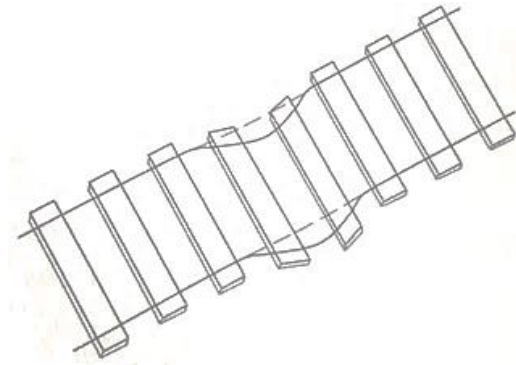


Fonte: DNIT (2015, p. 14)

Os componentes da via permanente podem apresentar anomalias e elas comprometem a geometria da via. Nessa dissertação foram levantados os defeitos geométricos que poderiam impactar as medições da geometria da via e ter relações com a imposição de restrições de velocidade temporárias na via. Os defeitos geométricos da via considerados nessa dissertação são: desnivelamento longitudinal, desnivelamento transversal, empeno, desalinhamento e abertura ou fechamento de bitola. Esses defeitos são detectados ao analisar a leitura de parâmetros geométricos da via e podem ser resultados também de outros problemas na via, como desmoronamento de ombreira de lastro.

O desnivelamento longitudinal é caracterizado pelo deslocamento vertical simultâneo dos dois trilhos da via (Figura 10). Ele gera desconforto aos passageiros durante passagem do trem e pode acarretar desengate de veículos e fracionamento do trem (ROSA e RIBEIRO, 2017).

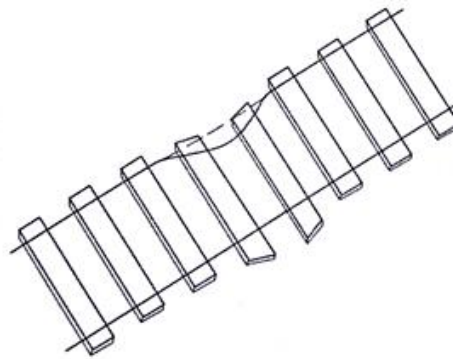
Figura 10 – Desenho esquemático do desnivelamento longitudinal



Fonte: Rosa e Ribeiro (2017, p.223).

O desnivelamento transversal (Figura 11) ocorre quando há diferença entre cotas de topo de trilho no sentido transversal, ou seja, quando um dos trilhos se desloca verticalmente e o outro não (ROSA e RIBEIRO, 2017).

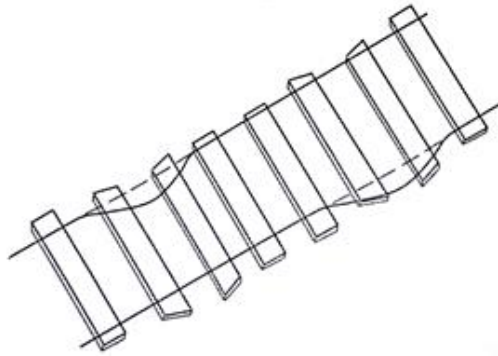
Figura 11 – Desenho esquemático do desnivelamento transversal



Fonte: Rosa e Ribeiro (2017, p. 223).

O empeno pode ser definido como uma variação do nivelamento transversal entre duas medições no sentido longitudinal da via (Figura 12). Ele é um dos defeitos mais prejudiciais à via, pois pode ocasionar o tombamento de um trem devido à torção da grade da via (ROSA e RIBEIRO, 2017).

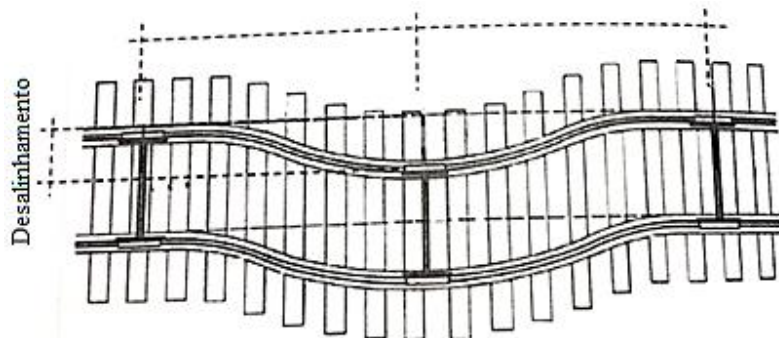
Figura 12 – Desenho esquemático do empeno



Fonte: Rosa e Ribeiro (2017, p. 224).

O desalinhamento é a diferença na medição da distância horizontal entre um ponto do trilho e uma reta formada por dois pontos: um ponto posterior e outro anterior ao ponto de referência (Figura 13). E por fim, a abertura ou fechamento de bitola são desvios na bitola da via.

Figura 13 – Desenho esquemático do desalinhamento

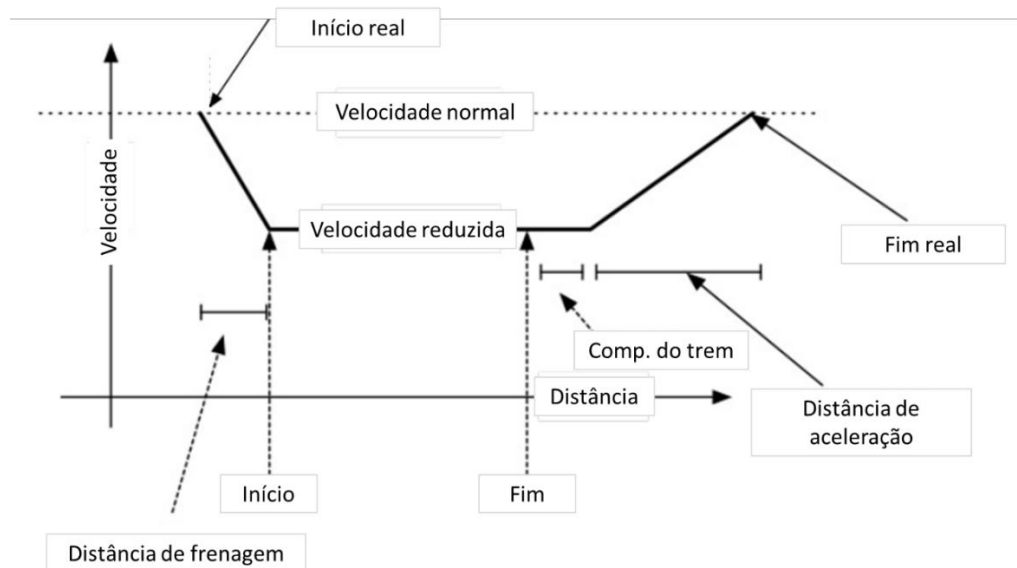


Fonte: Rosa e Ribeiro (2017, p.226).

Quando são detectadas essas irregularidades, são necessárias limitações na velocidade da via para garantir a segurança na via. Portanto, as restrições podem ser representadas como uma exceção no gráfico de velocidade normal da via (Figura 14). Os efeitos das restrições são percebidos antes do seu início e após o seu fim. Assim,

o início real da restrição é o momento no qual o trem começa a desacelerar para alcançar a velocidade reduzida, e o fim real é o momento no qual o trem volta a velocidade normal da via, que não é necessariamente a mesma de antes da restrição. Dessa maneira são caracterizados três os componentes da perda de tempo que ocorre em uma restrição: (1) o tempo de desaceleração; (2) trafegar numa velocidade mais baixa do que a normal; (3) o tempo de reaceleração. Portanto, as restrições influenciam o tempo de circulação dos trens na via e por isso devem ser retiradas com segurança o mais rápido possível (VAN DER KOOIJ et. al, 2017).

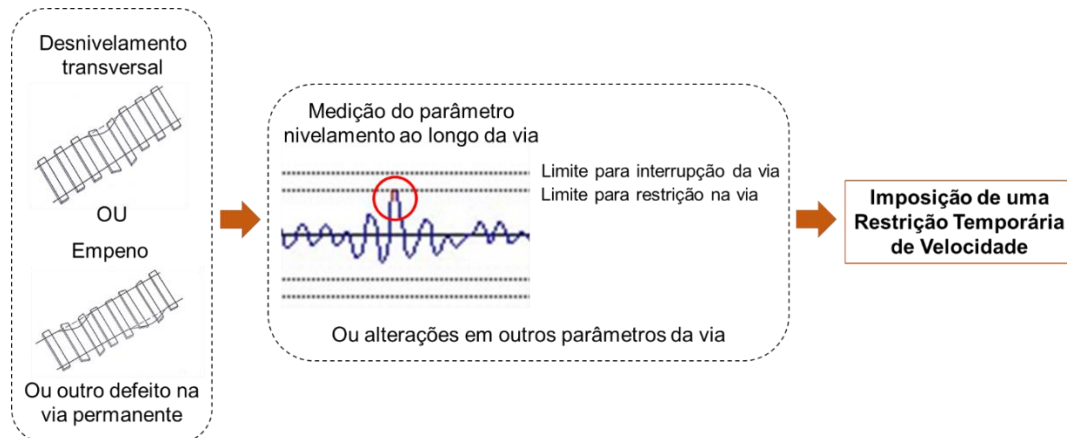
Figura 14 – Conceitos para o entendimento dos efeitos das restrições de velocidade



Fonte: Van Der Kooij (2017).

As restrições de velocidade podem ser impostas a partir da análise dos parâmetros da geometria da via (Figura 15). Esses parâmetros são medidos pelo carro-controle a cada 25 centímetros da ferrovia, gerando uma massiva quantidade de dados para ser processada. Além disso, uma restrição pode ser caracterizada pela combinação da medição de diversos parâmetros. Por exemplo, quando analisada a bitola de forma isolada, seus limites podem não indicar um problema, porém ao analisar a bitola em conjunto com o alinhamento vertical, pode indicar um desmoronamento da ombreira de lastro.

Figura 15 - Possíveis causas de restrições



Fonte: Autor (2019)

Para complementar os parâmetros geométricos, foi utilizada a tonelagem bruta transportada na via e os dados climáticos da ferrovia. As três bases de dados descritas anteriormente foram relacionadas com uma quarta: do histórico de restrições da via. Cada uma das bases possuía referenciais de localização e datas de medições diferentes. Assim, para a união foi necessário combinar e agrupar as informações, buscando as compatibilidades de espaço e tempo. Posteriormente serão detalhados o levantamento e o tratamento das bases de dados para a ferrovia na qual foi aplicado o sistema computacional proposto.

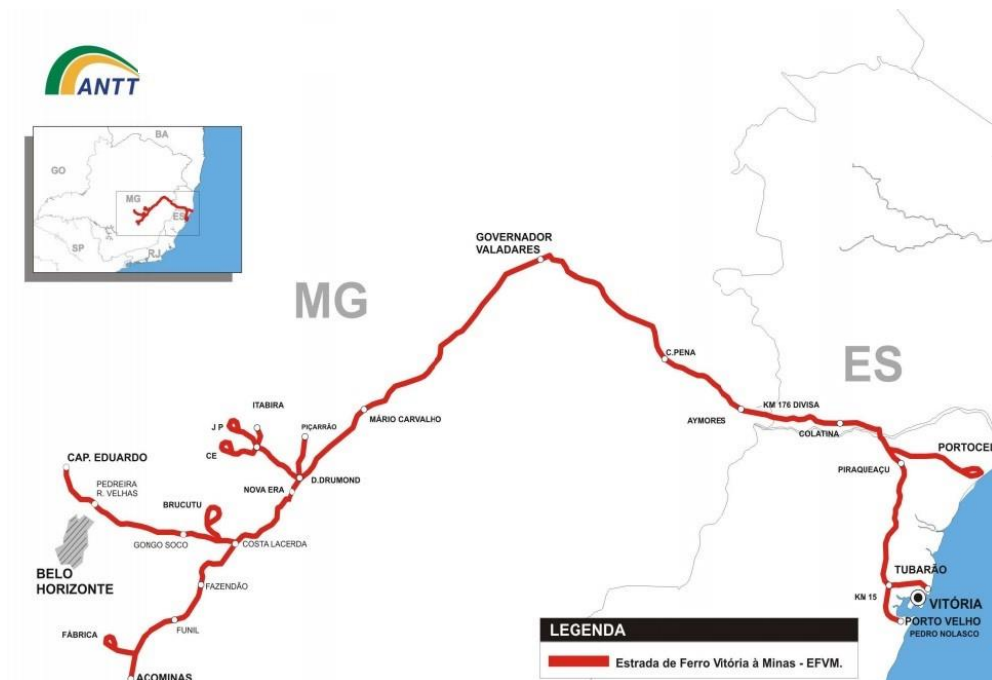
### 3.1. DESCRIÇÃO DA ESTRADA DE FERRO VITÓRIA-MINAS

Os dados coletados são provenientes da ferrovia brasileira Estrada de Ferro Vitória – Minas (EFVM) e são referentes aos meses de janeiro 2017 a outubro 2018. A Estrada de Ferro Vitória a Minas (EFVM) interliga a Região Metropolitana de Vitória, no estado do Espírito Santo, a Belo Horizonte, capital do estado de Minas Gerais, em seus 905 km de percurso, sendo essa ferrovia de concessão da empresa VALE S.A. e realiza o transporte de cargas e passageiros. Seu percurso totaliza 905km de linha férrea, sendo a maior parte composta por linha dupla. Ao longo de sua extensão, a ferrovia é dividida em supervisões, que são trechos que estão sob a responsabilidade de uma equipe de manutenção. A ferrovia foi concebida em bitola métrica e com velocidade

máxima de 65 km/h. Quando é imposta uma restrição em um trecho, sua velocidade é reduzida para 30 km/h.

Dos 905 km de extensão total, 601 km de extensão da EFVM estão em linha dupla. A realização das operações de transporte, ocorrem nas linhas duplas denominadas, linha 1 e linha 2, sendo que a linha 1 é a mais antiga. A Figura 16 apresenta o mapa geral de atuação da EFVM. Em média, são transportados pela ferrovia EFVM cerca de 145 milhões de toneladas de produtos por ano, por meio das linhas duplas, linha 1 e linha 2. O principal produto transportado na EFVM é o minério ferro, que corresponde cerca de 122 milhões de toneladas por ano, seguido pelos demais produtos que representam cerca de 23 milhões de toneladas por ano, como o carvão mineral, aço, ferro gusa, farelo de soja, derivados do petróleo, dentre outros.

Figura 16 – Mapa geral do traçado da EFVM



Fonte: Programa de Parcerias de Investimento (2019).

O carro-controlado foi a máquina de via utilizada para as medições dos parâmetros geométricos da via permanente. Ele é do modelo EM-80, fabricado pela empresa Plasser & Theurer (Figura 17), e realiza a leitura da geometria dos trilhos ferroviários, leituras das superelevações da geometria da linha, levantamento e leituras de desnivelamentos da linha, leituras de aberturas de bitolas, leituras de gabaritos

verticais e horizontais em túneis, pontes e viadutos, dentre outros. Ao transitar sobre uma linha férrea, o carro-controlê realiza suas medições a cada 25 centímetros da via e a periodicidade da inspeção na ferrovia estudada é de cerca de dois meses.

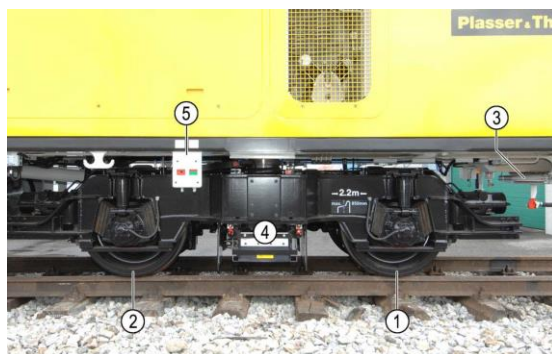
Figura 17 – Carro-controlê



Fonte: Autor (2019).

O Carro Controle EM-80 realiza as leituras da geometria dos trilhos ferroviários por meio de um sistema de equipamentos a laser e ópticos localizados em seus truques frontais, sendo estes posicionados no lado esquerdo e direito do veículo ferroviário, entre os eixos um e dois. A Figura 18 apresenta a localização dos dispositivos de leituras da geometria dos trilhos ferroviários.

Figura 18 – Localização do sistema de leitura do Carro Controle EM-80 - EFVM



- 1 = Eixo1
- 2 = Eixo2
- 3 = Luz de trabalho do truque frontal Eixo1
- 4 = Sistema KLD: Caixa do laser, lado direito
- 5 = Indicador de status de freio de estacionamento  
VERDE: Freio de estacionamento liberado  
VERMELHO: Freio de estacionamento acionado

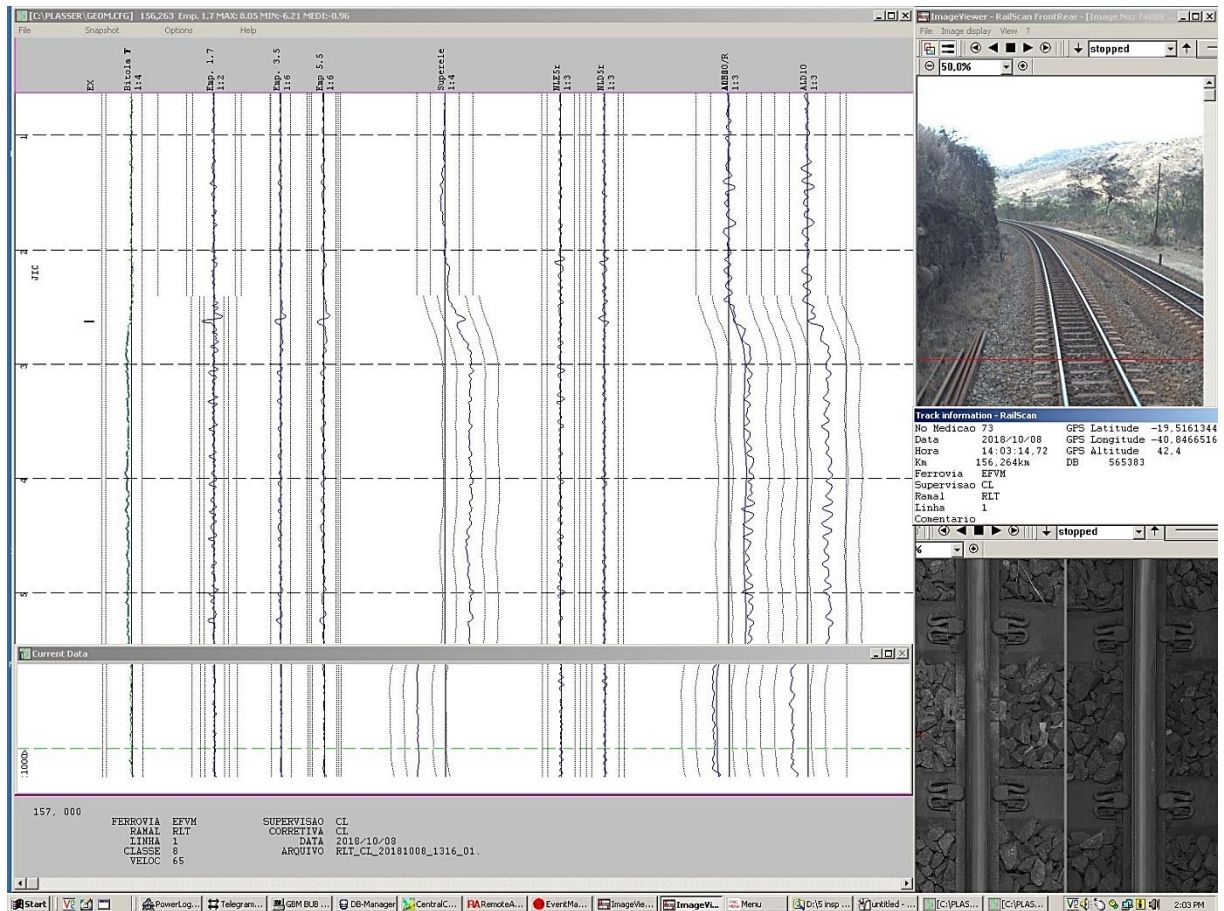
Fonte: Plasser & Theurer (2008, p. 25).

O Carro Controle EM-80, dispõe de um sistema de leituras que consiste em um processamento óptico e a laser. O sistema de processamento óptico de imagens denominado ORIAN VII é usado para fornecer os dados dimensionais em tempo real da geometria dos trilhos ferroviários com alta resolução. Sendo esses dados processados em tempo real para avaliar e monitorar as dimensões da geometria dos trilhos ferroviários.

O Carro Controle EM-80 dispõe de dois sistemas de controle para o posicionamento geográfico na via férrea, sendo um, o sistema de localização realizado pelo Sistema de Posicionamento Global (GPS) e o outro, o sistema de correção por meio de marcos geográfico disponibilizados ao longo da linha férrea da EFVM, denominados de estaqueamentos da linha. Esses sistemas estabelecem e confirmam que as leituras da geometria dos trilhos sejam realizadas continuamente nos mesmos pontos, de forma que se consiga realizar estas leituras continuamente nestes mesmos pontos.

O operador do Carro Controle EM-80, por meio de um monitor instalado na cabine de controle que se encontra interligado com o computador de bordo do veículo, consegue visualizar o processamento das leituras dos trilhos ferroviários em tempo real. A formatação da tela visualizada pelo operador do Carro Controle EM-80 (Figura 19) apresenta medições de nove parâmetros, a imagem da câmera frontal do carro e a imagem de uma câmera localizada sob o carro, que monitora a superestrutura da via. Todos os nove parâmetros foram utilizados nessa dissertação e os dados das câmeras não foram utilizados.

Figura 19 - Tela visualizada pelo operador do Carro Controle EM-80



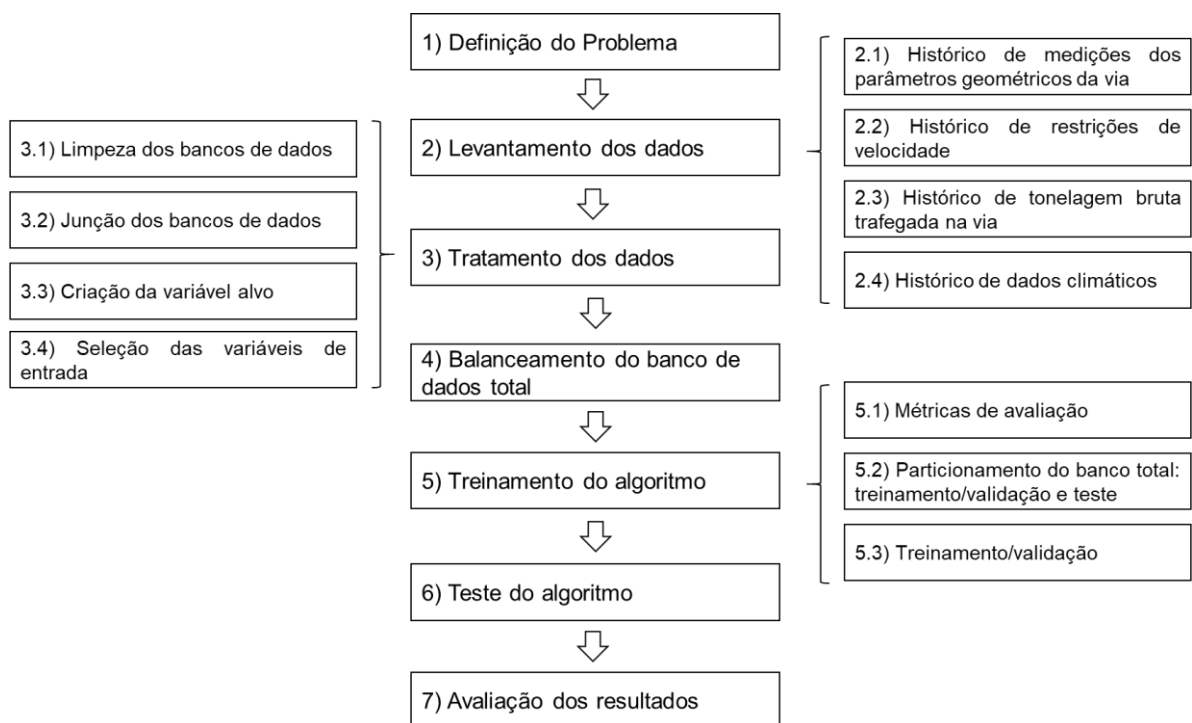
Fonte: Autor (2019).

Para os fins dessa pesquisa, foram selecionados os seguintes parâmetros da geometria da via para serem analisados: Alinhamento; Bitola; Empeno; Nivelamento; e Superelevação. Eles serão definidos no Capítulo 4 e foram selecionados em conjunto com a equipe especialista da empresa operadora da ferrovia estudada.

#### 4. MÉTODO PROPOSTO

O método proposto para desenvolver um sistema inteligente para realizar a predição da necessidade de imposição de restrições temporárias de velocidade na via permanente é composto de várias etapas descritas na Figura 20.

Figura 20 – Etapas do método proposto



Fonte: Autor (2020).

A 1) Definição do problema foi descrita no Capítulo 3, apontando todas as características do problema a ser resolvido e o objetivo da predição. O Capítulo 4 contém o 2) Levantamento, o 3) Tratamento de Dados e o 4) Balanceamento dos dados. O Levantamento de Dados consiste na obtenção e estruturação de um conjunto de dados histórico com todos os registros necessários para serem utilizados nas etapas de treinamento, validação e teste dos algoritmos de *machine learning*. Nele são especificados os três bancos de dados utilizados. O Tratamento de Dados diz respeito em analisar o conjunto de dados levantado anteriormente, verificando a qualidade dos registros e eliminando desse conjunto valores inconsistentes e erros de

leituras. Nele também é realizada a união dos bancos de dados, a criação da variável alvo e a seleção de variáveis de entrada. O Balanceamento dos dados é composto por técnicas para diminuir o desbalanceamento de registros da variável alvo, para que os modelos não fiquem tendenciosos e realizem previsões unicamente da classe majoritária da variável alvo.

O Capítulo 5 contém o 5) Treinamento do algoritmo, na qual busca-se investigar as configurações que propiciarão os melhores desempenhos para a previsão da necessidade de imposição de restrição. Para a análise e determinação das configurações que apresentaram o melhor desempenho, utilizam-se as métricas de avaliação que foram determinadas. Além disso, contém a validação cruzada realizada para avaliar o potencial de generalização dos modelos. Também contém o 6) Teste do algoritmo e a 7) Apresentação dos resultados.

#### 4.1. LEVANTAMENTO E TRATAMENTO DE DADOS

As restrições de velocidade podem ser impostas a partir da análise dos parâmetros da geometria da via. Esses parâmetros são medidos pelo carro-controle a cada 25 centímetros da ferrovia, gerando uma massiva quantidade de dados para ser processada. Além disso, uma restrição pode ser caracterizada pela combinação da medição de diversos parâmetros. Por exemplo, quando analisada a bitola de forma isolada, seus limites podem não indicar um problema, porém ao analisar a bitola em conjunto com o alinhamento vertical, pode indicar um desmoronamento da ombreira de lastro. Para complementar os parâmetros geométricos, foi utilizada a tonelagem bruta transportada na via e os dados climáticos. As três bases de dados anteriores foram relacionadas com uma quarta: do histórico de restrições da via. Cada uma das bases anteriores possuía referenciais de localização e datas de medições diferentes. Assim, para a união foi necessário combinar e agrupar as informações, buscando as compatibilidades de espaço e tempo. A diante serão detalhados o levantamento e o tratamento das bases de dados.

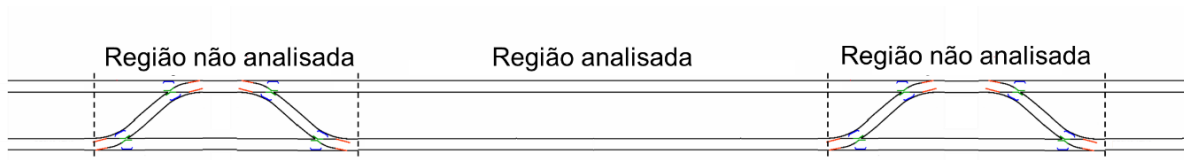
Os dados coletados são provenientes da ferrovia brasileira Estrada de Ferro Vitória – Minas (EFVM) e são referentes aos meses de janeiro 2017 a outubro 2018. Seu percurso totaliza 905km de linha férrea, sendo a maior parte composta por linha dupla.

Ao longo de sua extensão, a ferrovia é dividida em supervisões, que são trechos que estão sob a responsabilidade de uma equipe de manutenção. A ferrovia foi concebida em bitola métrica e com velocidade máxima de 65 km/h. Quando é imposta uma restrição em um trecho, sua velocidade é reduzida para 30 km/h.

As medições do carro-controle são armazenadas no banco de dados da empresa responsável pela gestão da ferrovia e dele foram levantados cerca de 70 milhões de registros referentes ao período analisado, que correspondiam a 11 inspeções. O carro-controle fornece medições de mais de cem parâmetros da via, então foi necessária uma seleção prévia dos parâmetros mais importantes para a predição. A equipe de inspeção da via selecionou 22 parâmetros para serem variáveis de entrada do modelo. Assim, a base de dados levantada continha 22 variáveis referentes aos parâmetros da via permanente e 12 variáveis referentes a identificação do trecho da leitura (dados gerais), resultando em 34 variáveis.

Após levantamento de dados do carro-controle, foi realizada uma limpeza inicial dos registros excluindo as medições realizadas nas regiões dos aparelhos de mudança de via (AMV), Figura 21. Essa primeira limpeza foi necessária pois os sensores do carro-controle perdem a referência quando passam essas regiões e realizam medições que não podem ser consideradas. Após a limpeza dos dados, o arquivo resultante apresentou uma redução de 25% em seu tamanho, contendo cerca de 53 milhões de registros. A segunda limpeza dos dados foi realizada para retirar as variáveis não relevantes para o estudo, como a variável referente ao nome do inspetor do carro-controle. Nessa limpeza, foram excluídas sete variáveis da base.

Figura 21 - Regiões não analisadas da Via Permanente



Fonte: Autor (2020).

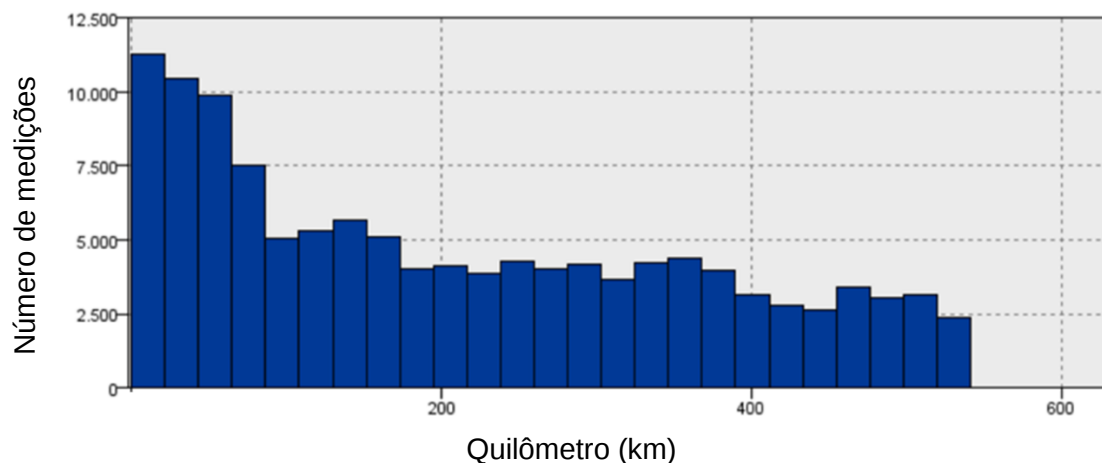
Os registros provenientes do carro-controle são referentes à toda extensão da ferrovia. Devido ao grande volume de registros e ao fato de que uma restrição de velocidade não é imposta em um trecho de apenas 25 centímetros, as medições foram agrupadas a cada 100 metros. Assim, um novo registro foi criado a partir de 400 registros de medição de 25 centímetros (Tabela 3). Cada novo registro contém média, máximo e desvio-padrão das medições dos parâmetros do carro controle nesse intervalo de 100 metros. Esse intervalo foi determinado pela equipe de planejamento da manutenção da via e pelo tamanho mínimo do trecho no qual ocorria uma restrição.

Tabela 3 – Identificação da local de medição em quilômetros

| Quilômetro do registro original | Quilômetro do novo registro |
|---------------------------------|-----------------------------|
| 2.600,00                        |                             |
| 2.600,25                        |                             |
| ⋮                               | 2.600,00                    |
| 2.699,75                        |                             |

O agrupamento reduziu a quantidade de registros para 221.200 e sua distribuição ao longo da extensão da ferrovia é mostrada no histograma da Figura 22. Os primeiros quilômetros apresentam maior quantidade de registros pois trechos de supervisão distintas possuem o mesmo valor de quilometragem. Esses trechos são diferenciados posteriormente pela supervisão.

Figura 22 – Distribuição do número de medições ao longo do trecho estudado



Fonte: Autor (2020).

Devido ao agrupamento, a quantidade de variáveis para aumentou para 72. Como o número de variáveis cresceu significativamente, foi necessária uma seleção de variáveis posterior para determinar as variáveis de entrada para o algoritmo. A técnica PCA (*Principal Component Analysis*) foi utilizada para a seleção e será detalhada posteriormente.

O banco de dados do histórico de restrições de velocidade foi obtido por meio de planilhas da equipe de manutenção da via. Ele contém as informações seguintes: supervisão, local de ocorrência, registrado em quilometro mais metro da ocorrência, linha da ocorrência e data de início da ocorrência (Tabela 4). Foram descartados os registros nos quais a ocorrência da restrição está associada a fatores não relacionados diretamente com os parâmetros da via, como o desmoronamento de um talude. Esse banco de dados foi denominado de banco de dados de restrição e possui 1.167 registros.

Tabela 4 – Amostra da base de dados do histórico de restrições de velocidade

| Supervisão | Linha | Local de ocorrência |        | Data início |
|------------|-------|---------------------|--------|-------------|
|            |       | Km Início           | Km Fim |             |
| PV         | 1     | 4.200               | 6.500  | 29/03/2018  |
| PV         | 2     | 17.771              | 17.971 | 16/03/2018  |
| CS         | 2     | 37.127              | 37.510 | 12/09/2017  |
| CS         | 2     | 44.395              | 44.636 | 15/12/2017  |
| CS         | 2     | 36.630              | 37.680 | 24/12/2017  |

Foi realizada a união dos registros dos bancos de dados da via com o banco de dados de restrição, gerando o banco de dados denominado banco de dados total, que além dos dados desses dois bancos de dados citados, contém informações sobre a tonelage bruta e dados climáticos, que são detalhados a seguir.

Tomou-se como referência o banco de dados da via para realizar a união de ambos os bancos e a chave primária utilizada foi o local de medição dos parâmetros, expresso em quilometro mais metro da medição. Nesse processo, caso existissem registros no banco de dados da via para um local e não existissem registros no banco

de restrição para o mesmo local, as variáveis referentes às restrições na base de dados total ficariam nulos. Na base de dados total descartou-se os registros que tivessem a data da restrição menor do que a data da medição dos parâmetros. Dessa forma, cada registro do banco de dados total que será utilizado para treinamento, teste e validação dos modelos, contém os dados das medições do carro-controle e a ocorrência, ou não, de uma restrição naquele local em uma data posterior (Tabela 5). Além disso, na base de dados total foi criada uma variável binária para a variável alvo do problema representando a ocorrência de restrição, referente a ocorrência ou não de uma restrição de velocidade em até 30 dias e uma segunda variável alvo referente a ocorrência ou não de uma restrição de velocidade em até 360 dias. Essa variável assume valor 1 caso no registro a diferença entre data de início da ocorrência da restrição e a data da medição do carro-controle seja menor ou igual a 30 dias para o alvo de 30 dias. O mesmo ocorre para a variável alvo de restrições em até 60 dias, porém com a diferença de datas menor ou igual a 60 dias. E assume valor zero caso contrário.

Tabela 5 – Amostra da base de dados geral

| CORRETIVA | LINHA | DATA       | Km_metros | Bit Fre_Mean |     | Alvo30dias | Alvo60dias |
|-----------|-------|------------|-----------|--------------|-----|------------|------------|
| NE        | 1     | 17/07/2018 | 0         | 1.738        |     | 0          | 0          |
| NE        | 1     | 17/07/2018 | 0.1       | 2.555        |     | 0          | 0          |
| NE        | 1     | 17/07/2018 | 0.2       | 6.076        |     | 0          | 1          |
| NE        | 1     | 17/07/2018 | 0.3       | 0.403        | ... | 0          | 0          |
| CP        | 2     | 27/11/2017 | 229.2     | 5.246        |     | 0          | 0          |
| CP        | 2     | 29/01/2018 | 229.2     | 6.205        |     | 1          | 1          |
| CP        | 2     | 13/03/2018 | 229.2     | 5.245        |     | 0          | 0          |
| CP        | 2     | 02/04/2018 | 229.2     | 4.975        |     | 0          | 0          |

Conforme mencionado anteriormente, no banco de dados total foi acrescentado mais uma variável referente a tonelagem bruta total transportada na via. A tonelagem bruta foi informada pela equipe de manutenção da via e representa a soma do peso total da carga movimentada na via com a tara dos vagões e locomotivas para o mesmo período das inspeções do carro-controle. Ela é informada por dia para cada supervisão da ferrovia de acordo com a linha férrea e compõe a base de dados de tonelagem. Para a inserção da base de tonelagem na base de dados total foi criado a

variável tonelada bruta. Essa variável é calculada como a soma da tonelada bruta diária no período de 60 dias anteriores a medição do carro-controle.

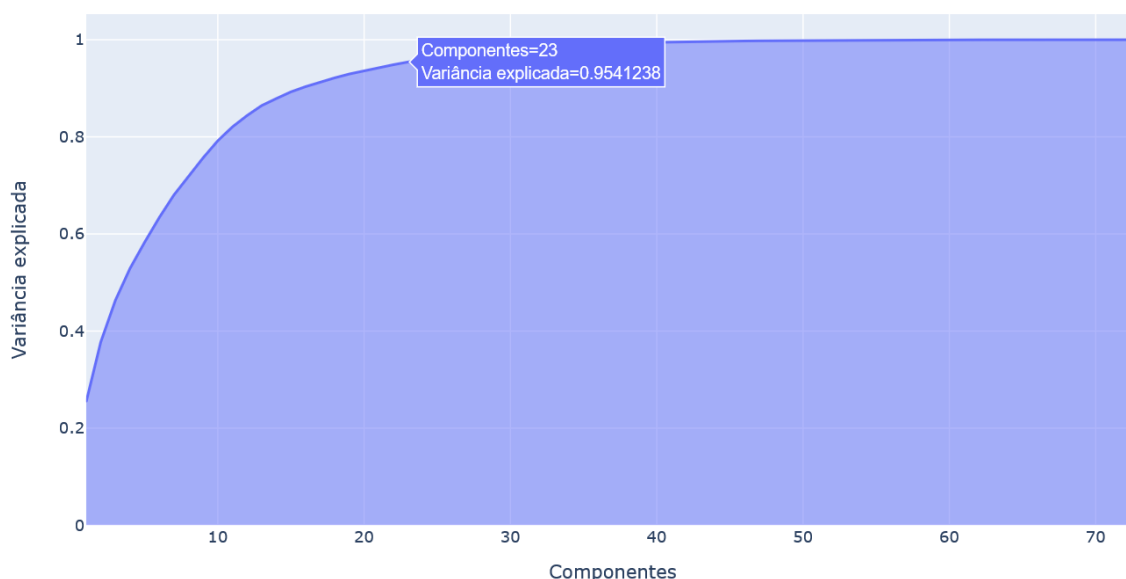
Além da variável da tonelagem bruta, foram incluídas no banco de dados total variáveis climáticas referentes aos trechos da ferrovia. A ferrovia é segmentada em nove microclimas ao longo do seu comprimento e dados históricos climáticos do ponto central de cada um dos microclimas foram consultados no *Weather Channel*, via um contrato que a Vale possui com a empresa. Foram obtidas informações de pluviometria e temperatura com granulometria horária para os pontos centrais e essas informações foram extrapoladas para a extensão do domínio do microclima. Essas variáveis climáticas foram agregadas em 1, 7, 14 e 30 dias utilizando a soma e valor máximo para a pluviometria nesses intervalos e utilizando a máxima, média, mínima e desvio padrão para a temperatura. Assim, para cada registro foram associados dados climáticos referentes a 7, 14 e 30 dias anteriores a medição do carro-controle. Além dessas derivações das variáveis de temperatura e pluviometria, foram mantidas as variáveis de pluviometria e temperatura no dia da medição dos parâmetros da via.

Após a junção de todos os dados e definição da variável alvo, foi realizada uma seleção de variáveis para definir as variáveis de entrada do modelo preditivo. Para a seleção, foram removidas as variáveis que apresentavam mais de 50% de valores nulos em seus registros ou com variação de valores menor que 10%. Após esse processo, o banco de dados total apresentou 73 variáveis, sendo 72 variáveis de entrada e 1 de saída. As variáveis de entrada referentes aos parâmetros geométricos da via permanente podem ser divididas em seis grupos: 1) Dados gerais; 2) Alinhamento; 3) Bitola; 4) Empeno; 5) Nivelamento; e 6) Superelevação. Referente aos dados gerais tem-se as seguintes variáveis: local da via; linha do local; supervisão do local; e identificação de curva/tangente. Referente ao alinhamento: foram usadas as medições do trilho esquerdo com máximo, média, desvio-padrão nas cordas de 5 e 10 metros. Referente à bitola: foram usados o máximo, média, desvio-padrão e mínimo das medições de bitola frente e trás. Referente ao nivelamento: máximo e desvio-padrão para as cordas de 10, 20 e 30 metros. Referente a superelevação: máximo, média e desvio-padrão.

Após limpeza inicial de variáveis, foi utilizada a técnica de PCA para seleção de variáveis para o modelo. Antes de aplicar a técnica, como a ordem de grandeza das

variáveis da base de dados total eram distintas, foi necessário fazer a normalização da base de dados. Essa normalização foi feita por meio do método z-score. Dentre as 72 variáveis de entrada, foi acumulada a variância da base de dados explicada de acordo com o número de componentes utilizadas, como apresentado na Figura 23. Nesse gráfico é possível determinar o número ideal de componentes a serem utilizadas e assim foram selecionadas 23 componentes principais, que explicam aproximadamente 95% da variância da base, como apresentado na

Figura 23 – PCA: variância explicada x componentes



Fonte: Autor (2021).

A partir das componentes principais, foram extraídas as variáveis mais importantes para as componentes principais (Tabela 6).

Tabela 6 – Variáveis mais importantes das componentes principais

| Variáveis mais importantes dos componentes principais |                                                            |
|-------------------------------------------------------|------------------------------------------------------------|
| NLEstd                                                | Desvio padrão do nivelamento no trecho de 100m agrupados   |
| ALEmean                                               | Média do alinhamento no trecho de 100m agrupados           |
| ALEstd                                                | Desvio padrão do alinhamento no trecho de 100m agrupados   |
| SUPrstd                                               | Desvio padrão da superelevação no trecho de 100m agrupados |

|              |                                                                             |
|--------------|-----------------------------------------------------------------------------|
| Emp.1.7mean  | Média do empeno 1.7 no trecho de 100m agrupados                             |
| NLE30mean    | Média do nivelamento corda 30m no trecho de 100m agrupados                  |
| PHAFOTEMPmin | Temperatura mínima do trecho no dia da passagem do carro-controle           |
| PHAFOTEMPstd | Desvio padrão da temperatura do trecho no dia da passagem do carro-controle |
| ALesq20mean  | Média do alinhamento corda 20m no trecho de 100m agrupados                  |
| BitFremean   | Média da bitola frente no trecho de 100m agrupados                          |
| BitFremin    | Mínimo da bitola frente no trecho de 100m agrupados                         |
| MTBT_Acum_60 | Acumulado de tonelada transportada no trecho nos últimos 60 dias            |
| PPPH_Acum_30 | Acumulado de chuva no trecho nos últimos 30 dias                            |
| SUPmmean     | Média da superelevação no trecho de 100m agrupados                          |
| LINHA        | Linha do trecho                                                             |
| SUPmstd      | Desvio padrão da superelevação no trecho de 100m agrupados                  |
| TEMP_Rmax_30 | Máxima temperatura do trecho nos últimos 30 dias                            |
| Emp.1.7max   | Máximo do empeno 1.7 no trecho de 100m agrupados                            |
| NLEmin       | Mínimo do nivelamento no trecho de 100m agrupados                           |
| BitFrestd    | Desvio padrão da bitola frente no trecho de 100m agrupados                  |
| ALesq30min   | Mínimo do alinhamento corda 30m no trecho de 100m agrupados                 |
| ALesq30max   | Máximo do alinhamento corda 30m no trecho de 100m agrupados                 |

---

A base de dados de entrada foi dividida em dois grupos: 80% dos registros foi usado para treinamento e validação dos modelos e 20% para teste. Os registros separados para teste não foram apresentados ao algoritmo no treinamento. A semente aleatória para segmentação dos registros foi a mesma para todas as segmentações, garantindo que os dados de entrada são exatamente iguais para todos os modelos.

O desbalanceamento das variáveis alvo é muito significativo na base de dados. Na partição de treinamento/validação da variável alvo com horizonte de tempo de 30 dias, 99,2% dos registros não apresentam imposição de restrição de velocidade, enquanto na variável alvo com horizonte de tempo de 60 dias esse percentual é de 99%.

Esse desbalanceamento pode ser um problema para os algoritmos de *Machine Learning*, pois eles tendem a focar na classe majoritária, podendo não apresentar

predições para a classe minoritária. Assim, foram testadas duas técnicas de balanceamento: o ADASYN e o SMOTE. Ambas as técnicas de balanceamento foram aplicadas na partição de treinamento dos modelos. Devido ao desbalanceamento, algumas técnicas difundidas para avaliação de modelos, como a acurácia, não são eficazes. Assim, são utilizadas quatro métricas de avaliação dos modelos: a matriz de confusão, taxa de verdadeiro positivo, taxa de falso positivo e a *area under the curve* (AUC).

## 5. ALGORITMO PROPOSTO E RESULTADOS

Dada a complexidade da análise desses dados e a necessidade de prever a ocorrência de restrições, um algoritmo de *machine learning* foi proposto: o *XGBoost*. De acordo com Chen e Guestrin (2016), esse algoritmo é amplamente utilizado por cientistas de dados para alcançar excelentes resultados em muitos desafios de *machine learning*. Assim como outros algoritmos, o *XGBoost* possui hiper parâmetros que não são estimados pelo próprio modelo e que devem ser determinados adequadamente para que sua performance seja a melhor possível. Para a determinação desses hiper parâmetros foi utilizado um algoritmo de otimização Bayesiana.

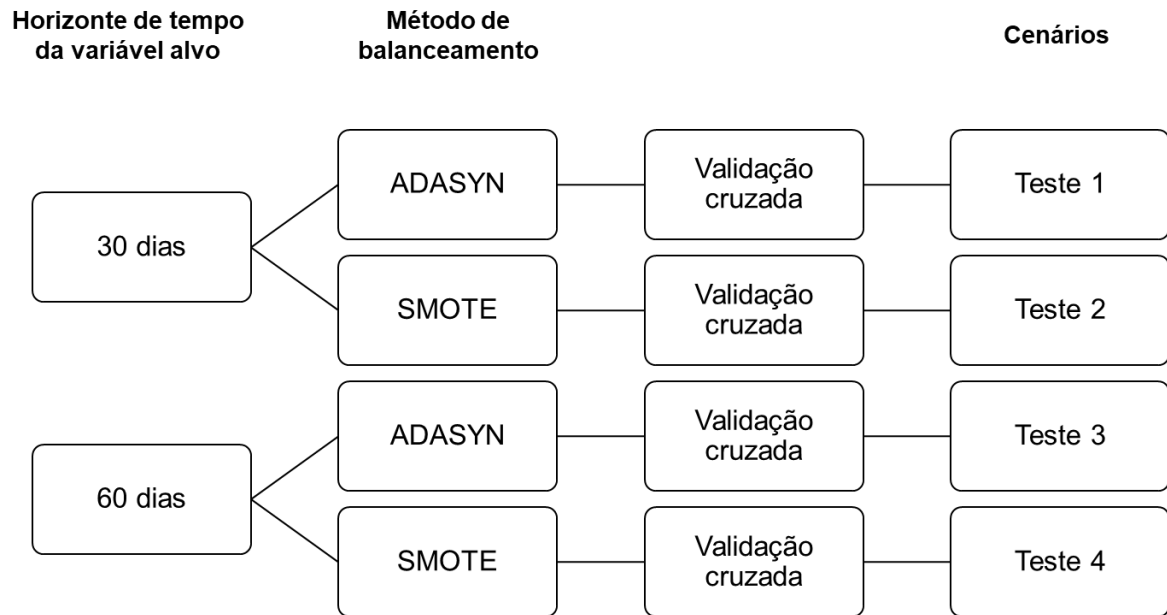
Os hiper parâmetros que tiveram seus valores otimizados foram: máxima profundidade da árvore, variando no intervalo de 3 a 10; o parâmetro gamma, variado de 0 a 1; a taxa de aprendizagem, variando de 0 a 1; e o número de estimadores, variando de 100 a 120. Os parâmetros a serem estimados e seu intervalo de variação foram definidos de acordo com Mohammadi et al. (2019) devido à similaridade das bases de dados e das características dos problemas. Após a otimização Bayesiana, concluiu-se que os parâmetros padrão do algoritmo *XGBoost* (*XGBoost*, 2020) apresentaram os melhores resultados, assim eles foram utilizados para as previsões (Tabela 7).

Tabela 7 - Hiper parâmetros do *XGBoost*

| Parâmetro                     | Valor |
|-------------------------------|-------|
| gamma                         | 0     |
| taxa de aprendizagem          | 1     |
| máxima profundidade da árvore | 10    |
| número de estimadores         | 113   |

Foram testadas duas configurações de balanceamento dos dados: com o balanceamento SMOTE e com o balanceamento ADASYN para os dois horizontes de tempo propostos nas variáveis alvo: 30 e 60 dias futuros. Assim, a Figura 24 ilustra os cenários avaliados. O objetivo era testar se o método de balanceamento poderia influenciar no resultado das previsões. Ao balancear os dados, foram utilizadas as mesmas sementes aleatórias para todos os cenários.

Figura 24 - Cenários avaliados



Fonte: Autor (2021)

Nos cenários propostos, os dados de entrada foram os mesmos e a variável alvo que foi modificada. O tratamento dos dados inicial foi realizado no software IBM SPSS Modeler 18.1 (IBM, 2019) utilizando um computador com dois processadores Intel® Xeon® CPU E5-2640 v3 2.60GHz, 8 núcleos e 64GB de memória RAM e o PCA, os balanceamentos e o algoritmo utilizado para a predição foram da bibliotecas *scikit-learn* (PEDREGOSA et al., 2011), *imbalanced-learn* (LEMAÎTRE; NOGUEIRA e ARIDAS, 2017) e *XGBoost* (XGBOOST, 2020) em Python 3.7, respectivamente. Na Tabela 8 são apresentados os resultados alcançados. A validação cruzada foi realizada com  $k = 10$  *folds* e os resultados de AUC médio das dez iterações em cada cenário e seu desvio padrão são apresentados na terceira coluna da Tabela 8. Na quarta coluna da Tabela 8, são apresentados os valores médios de AUC e seus desvios padrão referentes às predições realizadas com a base de teste. O algoritmo foi treinado trinta vezes para cada cenário e foram realizadas predições com a partição de teste para cada treinamento.

Tabela 8 - Resultados dos modelos preditivos

| Variável alvo                           | Método de balanceamento | AUC Validação cruzada | AUC Partição de testes |
|-----------------------------------------|-------------------------|-----------------------|------------------------|
| Restrições de velocidade em até 30 dias | ADASYN                  | 0,740 ± 0,029         | 0,738 ± 0,010          |
|                                         | SMOTE                   | 0,734 ± 0,038         | 0,743 ± 0,009          |
| Restrições de velocidade em até 60 dias | ADASYN                  | 0,741 ± 0,035         | 0,762 ± 0,007          |
|                                         | SMOTE                   | 0,730 ± 0,035         | 0,761 ± 0,011          |

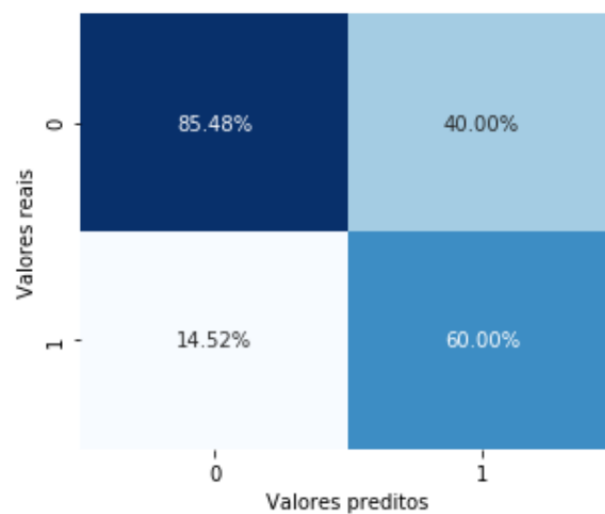
A pequena diferença entre os resultados do AUC na validação cruzada e na partição de teste mostra que o algoritmo não apresenta *overfitting* em nenhum dos cenários avaliados. Assim, ao apresentar aos algoritmos treinados os dados da ferrovia não conhecidos previamente, eles demonstram performances compatíveis com as da validação cruzada. Portanto, pode-se dizer que a capacidade de generalização do algoritmo nesses cenários é boa.

Comparando os resultados de AUC obtidos para os dois métodos de balanceamento testados para cada variável alvo, conclui-se que a diferença entre os métodos não teve forte influência no desempenho dos modelos. A variação no AUC na validação cruzada entre os métodos foi de, em média, 1%. Essa pequena variação mostra que o diferencial do ADASYN sobre o SMOTE, que é a geração de novos registros em quantidade proporcional ao grau de dificuldade de o algoritmo aprender com esses registros, não se destacou na base de dados utilizada. Uma hipótese para esse comportamento é que os registros que representavam futuras restrições de velocidade não eram muito distintos entre si.

Ao observar os desvios padrão dos cenários, conclui-se que a variação dos resultados é quase inexistente. Assim, analisar uma execução de predições do algoritmo de cada cenário é suficiente para representar o cenário. As matrizes de confusão de cada cenário apresentado na Figura 24 são apresentadas nas Figura 25, Figura 26, Figura 27 e Figura 28. De maneira geral, os cenários apresentam elevados valores de verdadeiro negativo e isso representa a situação na qual o algoritmo não prevê uma restrição e ela realmente não ocorre; baixos valores de falsos negativos, que representa a situação do algoritmo não identificar uma restrição futura; porém um número significativo de falsos positivos. Como consequência da última condição, o algoritmo indicaria que mais trechos teriam restrições temporárias de velocidade do

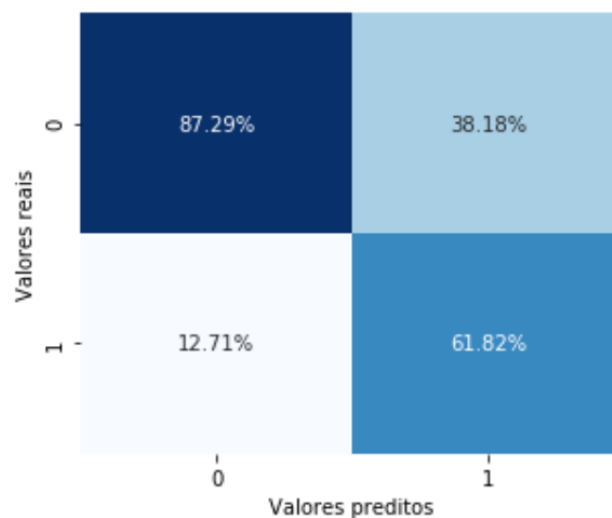
que realmente tiveram. Comparando os resultados entre as variáveis alvo de 30 e 60 dias, conclui-se que os modelos são mais assertivos quando predizemos a necessidade de imposição de restrições temporárias de velocidade em até 60 dias. A proporção de verdadeiros positivos nesse caso é em média 10% maior do que as predições referentes a até 30 dias. Possivelmente isso ocorre devido ao aumento de casos positivos de imposições de restrições temporárias de velocidade nos dados históricos quando considerado um horizonte de até 60 dias.

Figura 25 - Matriz de confusão normalizada - Teste 1



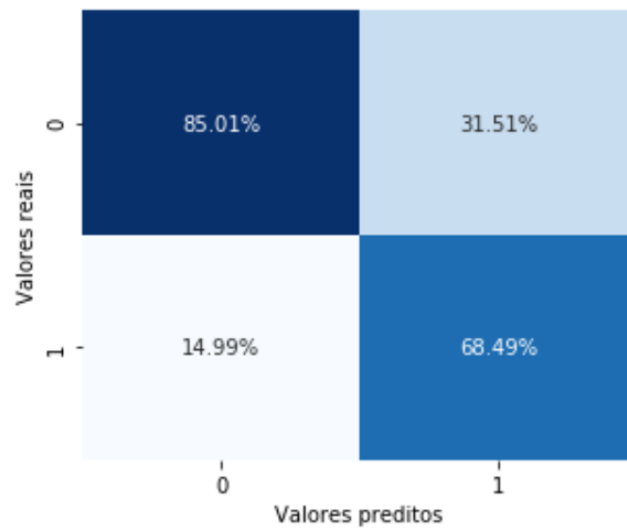
Fonte: Autor (2021)

Figura 26 - Matriz de confusão normalizada - Teste 2



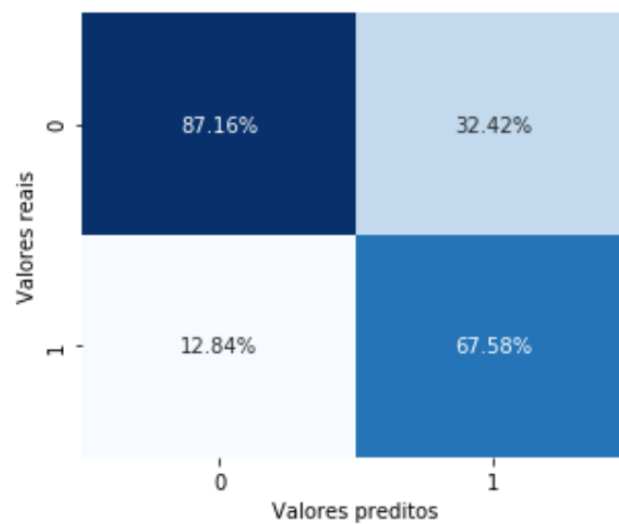
Fonte: Autor (2021)

Figura 27 - Matriz de confusão normalizada - Teste 3



Fonte: Autor (2021)

Figura 28 - Matriz de confusão normalizada - Teste 4



Fonte: Autor (2021)

Para melhor interpretar os resultados das matrizes de confusão, foram calculadas as taxas de Verdadeiro Positivo (sensibilidade) e de Falso Positivo. A sensibilidade, que representa a capacidade do modelo em prever corretamente a ocorrência de

restrições, é fundamental para o direcionamento da equipe de manutenção preventiva da via. Quanto mais próximo de um, melhor é o resultado. A taxa de Verdadeiro Positivo para a variável alvo de 30 dias é de 83% e de 84% para a variável alvo de 60 dias. Dessa maneira, mostra-se que a maior parte das restrições são previstas.

A Taxa de Falso Positivo representa as restrições de velocidade que o modelo prevê, porém que na realidade elas não irão ocorrer em um período de 30 dias ou 60 dias. É essencial que essa taxa seja a menor possível (zero), pois o custo do envio uma equipe de manutenção para realizar uma inspeção na via é elevado. A taxa de Falso Positivo para a variável alvo de 30 dias é de 30% e de 27% para a variável alvo de 60 dias.

Os resultados das taxas de Verdadeiro Positivo e de Falso Positivo mostram que a maior parte das restrições são previstas, porém há chances de a equipe de manutenção ser enviada para um trecho que não haverá restrições. Para a equipe de manutenção, é mais relevante identificar as futuras restrições, mesmo com falsos positivos, do que elas serem impostas. Portanto, os resultados alcançados são compatíveis com as necessidades da equipe da ferrovia.

## 5.1. DISCUSSÕES GERENCIAIS

A seleção de variáveis por meio do PCA apresenta para as equipes da ferrovia as variáveis mais relevantes para a imposição de restrições temporárias de velocidade na via permanente e essas variáveis podem ser divididas em quatro grupos: 1) Variáveis referentes à geometria da via permanente; 2) Variáveis referentes à via permanente; 3) Variáveis climáticas; e 4) Variáveis relacionadas à tonelada bruta transportada na via.

O primeiro grupo referente às variáveis de geometria da via, as variáveis podem ser subdivididas em cinco categorias: 1) Variáveis relacionadas ao alinhamento; 2) Variáveis relacionadas à bitola; 3) Variáveis relacionadas ao empeno; 4) Variáveis relacionadas ao nivelamento; 5) Variáveis relacionadas à superelevação. Essas variáveis representam as condições da via permanente. De maneira geral, conclui-se que todas as categorias mapeadas inicialmente pela equipe da ferrovia são relevantes para caracterizar as restrições temporárias de velocidade.

Na categoria de alinhamento, os seguintes parâmetros se destacaram: a média e o desvio padrão do alinhamento no trecho de 100 metros, a média do alinhamento na corda de 20 metros no trecho de 100 metros, e o máximo e o mínimo do alinhamento na corda de 30 metros no trecho de 100 metros. Portanto, o alinhamento medido nas cordas de 20 e 30 metros também é importante para caracterizar a restrição temporária de velocidade, medições essas que não são usualmente utilizadas pela equipe da via permanente.

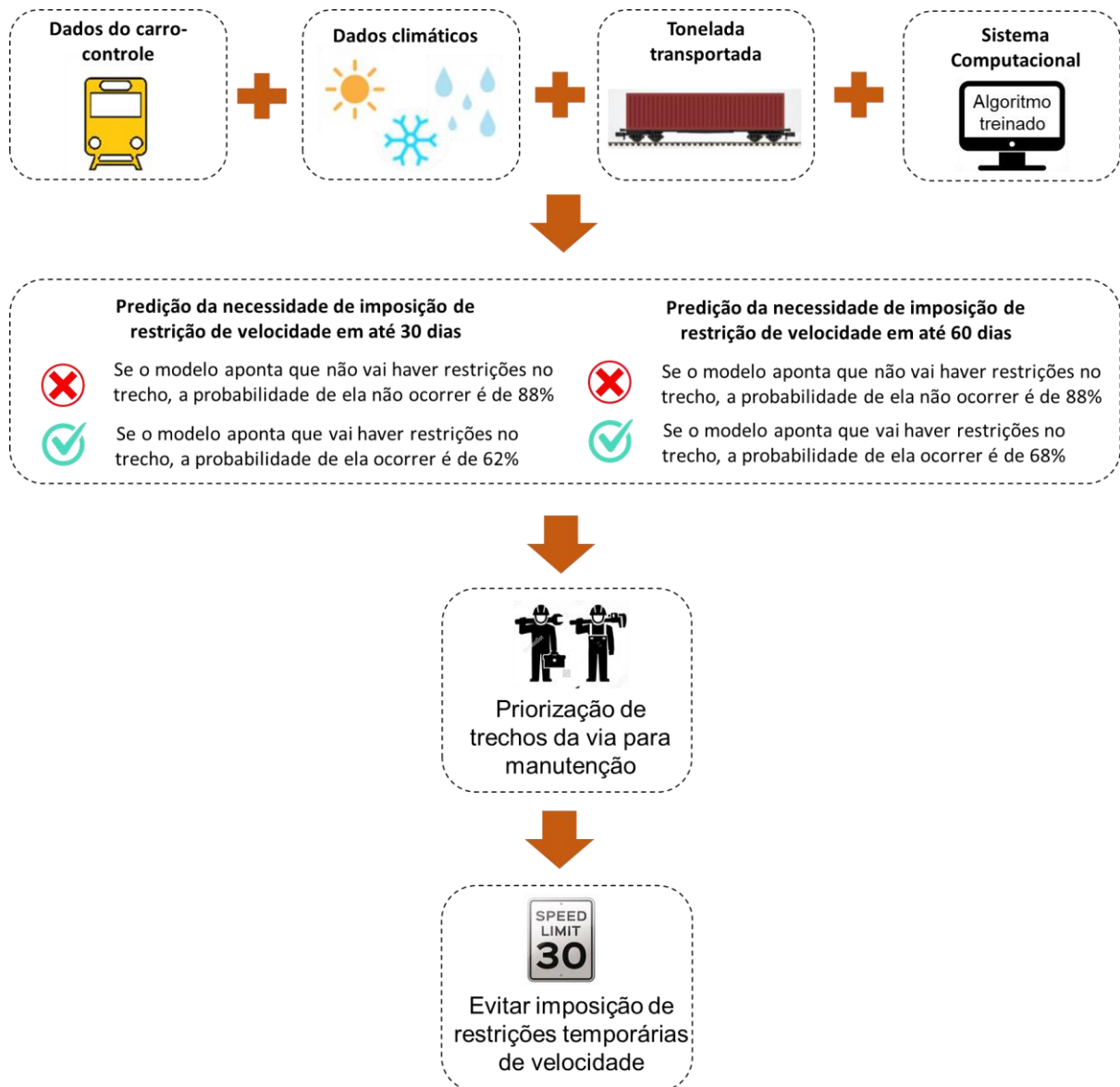
Na categoria de bitola, a média, o mínimo e o desvio padrão da bitola frente no trecho de 100 metros foram mais relevantes. Na categoria empeno, apesar de inicialmente na base de dados haver medições de empeno à 1.7, 3.5 e 5.5 metros, as variáveis mais relevantes foram a média e o máximo do empeno 1.7 no trecho de 100 metros.

Na categoria de nivelamento, as variáveis destacadas foram o mínimo e o desvio padrão no trecho de 100 metros e a média do nivelamento na corda de 30 metros no trecho de 100 metros. Na categoria superelevação, a média e o desvio padrão da superelevação no trecho de 100 metros foram mais importantes.

No segundo grupo, referente às características da via, a linha foi a variável mais relevante. No terceiro grupo, referente às variáveis climáticas, a temperatura mínima e seu desvio-padrão no dia da medição do carro-controle, assim como a temperatura máxima nos últimos 30 dias foram relevantes para o modelo. Além disso, a chuva acumulada nos últimos 30 dias da medição do carro-controle também foi apontada como variável importante. Isso mostra como variáveis externas às condições da via são relevantes para imposição de restrições temporárias de velocidade. No último grupo, referente à tonelada bruta transportada na via, a variável que acumulou a tonelada bruta transportada no trecho nos últimos 60 dias anteriores à medição do carro-controle foi destacada. Isso representa a carga transportada entre, aproximadamente, duas inspeções consecutivas do carro-controle, e demonstra que as imposições de restrições temporárias de velocidade são impactadas pelo transporte de carga na via permanente.

Definidas as variáveis mais relevantes do modelo, o algoritmo de *machine learning* foi treinado para duas variáveis alvo: a necessidade de imposição de restrições de velocidade em até 30 dias e a necessidade de imposição de restrições de velocidade em até 60 dias. Os resultados são apresentados na Figura 29.

Figura 29 - Resultados dos modelos



Fonte: Autor (2021)

Em ambos os modelos, quando a previsão de imposição de restrições é negativa, isso é, eles apontam que não haverá restrições, o acerto do modelo é de quase 90%. Portanto, a probabilidade de os modelos não identificarem uma restrição é baixa. Entretanto, quando os modelos identificam que haverá uma restrição, a probabilidade de ela realmente ocorrer é de 62%, no caso da variável alvo de 30 dias, e de 68% no caso da variável alvo de 60 dias. Assim, a equipe de manutenção pode ser direcionada

para um trecho no qual não haverá restrições. Como para a equipe de manutenção é mais relevante identificar as restrições, mesmo que com alguns falsos indicativos de restrições, os resultados alcançados são compatíveis com as necessidades da equipe da ferrovia.

## 6. CONCLUSÕES

As ferrovias estão constantemente buscando aumentar a quantidade de trens circulando na via. Uma das condições que leva à redução da quantidade de trens em circulação é a qualidade da via permanente. Quando é identificada alguma anomalia na via permanente, a velocidade máxima do trem no trecho que apresenta a anomalia deve ser reduzida para garantir a segurança da operação, caracterizando uma restrição temporária de velocidade. Para prevenir a presença de anomalias na via, a manutenção preventiva é fundamental. Nesse contexto, a priorização de trechos da ferrovia para a manutenção é uma atividade recorrente e que poderia ser facilitada por um modelo que prediz a necessidade da imposição de uma restrição temporária de velocidade no futuro.

Foi desenvolvido um sistema inteligente utilizando o algoritmo *XGBoost* para prever a necessidade de imposição de restrições temporárias de velocidade nos trens a partir dos parâmetros geométricos da via permanente, da carga transportada no período analisado e das condições climáticas da ferrovia. Os resultados desse sistema têm função de auxiliar a tomada de decisão dos especialistas visando evitar a ocorrência de restrições e melhorar a capacidade de circulação da via. Consequentemente, há um potencial ganho financeiro para a ferrovia.

Nesse sistema, foram desenvolvidos dois modelos de *machine learning*: um com a variável alvo de imposição de restrições de velocidade em até 30 dias e o outro com a variável alvo de imposição de restrições de velocidade em até 60 dias. Em ambos os modelos, quando a previsão de imposição de restrições é negativa, isso é, eles apontam que não haverá restrições, o acerto do modelo é de quase 90%. Portanto, a probabilidade de os modelos não identificarem uma restrição é baixa. Entretanto, quando os modelos identificam que haverá uma restrição, a probabilidade de ela realmente ocorrer é de 62%, no caso da variável alvo de 30 dias, e de 68% no caso da variável alvo de 60 dias. Assim, a equipe de manutenção pode ser direcionada para um trecho no qual não haverá restrições. Como para a equipe de manutenção é mais relevante identificar as restrições, mesmo que com alguns falsos indicativos de restrições, os resultados alcançados são compatíveis com as necessidades da equipe da ferrovia.

## 6.1. TRABALHOS FUTUROS

Propõem-se para trabalhos futuros a utilização de métodos combinados de balanceamento para a tentativa de reduzir os casos de falsos positivos, assim como WANG et al. (2019) fizeram. Sugere-se ainda a utilização de novas fontes de dados com uma frequência temporal maior, como os dados diários provenientes do vagão instrumentado que iniciou a operação este ano na Vale. Um outro segmento de pesquisa que pode ser explorado, é tratar o mesmo problema desta dissertação por algoritmos de regressão.

## Referências

ABNT - ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 7590: Trilhos Vignole - Requisitos**. Rio de Janeiro. 2012.

AGÊNCIA BRASIL. **Governo prevê investimento de R\$ 30 bi em ferrovias em cinco anos**. Disponível em: <https://agenciabrasil.ebc.com.br/politica/noticia/2020-02/governo-preve-investimento-de-r30-bi-em-ferrovias-nos-proximos-5-anos>. Acesso em 06 de junho de 2020.

AGÊNCIA NACIONAL DE TRANSPORTE TERRESTRE (ANTT). **Anuário Estatístico 2018**. 2018. Disponível em: <[http://www.antt.gov.br/ferrovias/arquivos/Anuario\\_Estatistico.html](http://www.antt.gov.br/ferrovias/arquivos/Anuario_Estatistico.html)>. Acesso em: 08 de agosto de 2019.

CAI, Jie et al. Feature selection in machine learning: A new perspective. **Neurocomputing**, v. 300, p. 70-79, 2018.

CHAWLA, Nitesh V. et al. SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321-357, 2002.

CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. **Proceedings of the 22nd international conference on knowledge discovery and data mining**. 2016. p. 785-794.

DING, Steven X. **Data-driven design of fault diagnosis and fault-tolerant control systems**. Springer London, 2014.

DNIT- Departamento Nacional de Infraestrutura de Transportes, **PIM Procedimento de Inspeção de Material Trilho**, p. 44. 2015.

DNIT. **Glossário de termos ferroviários**. Disponível em: <https://www.gov.br/dnit/pt-br/ferrovias/glossario-de-termos-ferroviarios/glossario.pdf>. Acesso em 26 de setembro de 2020.

DUDA, Richard O.; HART, Peter E.; STORK, David G.; **Pattern Classification**. John Wiley & Sons, 2012, 738p.

FAWCETT, Tom. ROC graphs: Notes and practical considerations for researchers. **Machine learning**, v. 31, n. 1, p. 1-38, 2004.

FAWCETT, Tom; PROVOST, Foster. **Data Science para Negócios: O que você precisa saber sobre mineração de dados e pensamento analítico de dados**. Alta Books Editora, 2016.

FINK, Olga; ZIO, Enrico; WEIDMANN, Ulrich. Predicting time series of railway speed restrictions with time-dependent machine learning techniques. **Expert Systems with Applications**, v. 40, n. 15, p. 6033-6040, 2013.

FRAZIER, Peter I. Bayesian optimization. **INFORMS**, 2018. p. 255-278.

GOMES, Pedro. **Machine Learning para todos**. Disponível em: <https://www.datageeks.com.br/machine-learning/>. Acesso em 10 de janeiro de 2021

HASANIN, Tawfiq et al. Examining characteristics of predictive models with imbalanced big data. **Journal of Big Data**, v. 6, n. 1, p. 69, 2019.

HE, Haibo et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. **IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)**. IEEE, p. 1322-1328, 2008.

IBM. **IBM SPSS Modeler 18.1**. 2019. Disponível em: <<https://www.ibm.com/br-pt/products/spss-modeler>>. Acesso em: 05 Maio 2019.

KLEIN, Aaron et al. Fast bayesian optimization of machine learning hyperparameters on large datasets. **PMLR**, 2017. p. 528-536.

- KRUMMENACHER, Gabriel et al. Wheel defect detection with machine learning. **IEEE Transactions on Intelligent Transportation Systems**, v. 19, n. 4, p. 1176-1187, 2017.
- KUBAT, Miroslav. **An introduction to machine learning**. Springer International Publishing AG, 2017.
- LASISI, Ahmed; ATTOH-OKINE, Nii. Principal components analysis and track quality index: A machine learning approach. **Transportation Research Part C: Emerging Technologies**, v. 91, p. 230-248, 2018.
- LEE, Jun S. et al. Deterioration prediction of track geometry using periodic measurement data and incremental support vector regression model. **Journal of Transportation Engineering, Part A: Systems**, v. 146, n. 1, p. 04019057, 2020.
- LEE, Wei-Hsun; YEN, Li-Hsien; CHOU, Chien-Ming. A delay root cause discovery and timetable adjustment model for enhancing the punctuality of railway services. **Transportation Research Part C: Emerging Technologies**, v. 73, p. 49-64, 2016.
- LEMAÎTRE, Guillaume; NOGUEIRA, Fernando; ARIDAS, Christos K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. **The Journal of Machine Learning Research**, v. 18, n. 1, p. 559-563, 2017.
- LI, Hongfei 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.
- MARTINEZ-CANTIN, Ruben. BayesOpt: a Bayesian optimization library for nonlinear optimization, experimental design and bandits. **J. Mach. Learn. Res.**, v. 15, n. 1, p. 3735-3739, 2014.
- MITCHELL, Rory; FRANK, Eibe. Accelerating the XGBoost algorithm using GPU computing. **PeerJ Computer Science**, v. 3, p. e127, 2017.
- MOHAMMADI, Reza et al. Exploring the impact of foot-by-foot track geometry on the occurrence of rail defects. **Transportation research part C: emerging technologies**, v. 102, p. 153-172, 2019.
- PEDREGOSA et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011.
- PLASSER & THEURER. Carro Controle Em-80 CVRD - Brasil - **Manual de Operação BA 4992 - Parte 2- Fotografias & Figuras**, Áustria, 2008, 122p.
- Programa de Parcerias de Investimento. **EFVM – Estrada de Ferro Vitória a Minas Gerais**. Disponível em: <https://www.ppi.gov.br/efvm-estrada-de-ferro-vitoria-a-minas> Acessado em 18 de fevereiro de 2020.
- PROVEZZA, Luca et al. Monitoring the Damage Evolution in Rolling Contact Fatigue Tests Using Machine Learning and Vibrations. **Metals**, v. 11, n. 2, p. 283, 2021.
- PUTRA, J C P; SAFRILAH. Application of Artificial Neural Network to Predict the use of Runway at Juanda International Airport. **IOP Conference Series: Materials Science and Engineering**, v. 209, p.012107-012115, jun. 2017.
- RIVES, Fernando Oliveros et al. **Tratado de ferrocarriles**. Rueda, 1977.
- ROSA, Rodrigo de Alvarenga.; RIBEIRO, R. C. H. **Estradas de Ferro: Projeto, Especificação & Construção**. 1. ed. Vitória: Editora da Universidade Federal do Espírito Santo – EDUFE, 2017.
- SILVA, Renato M.; ALMEIDA, Tiago A.; YAMAKAMI, Akebo. Análise de desempenho de redes neurais artificiais para classificação automática de web spam. **Revista Brasileira de Computação Aplicada**, v. 4, n. 2, p. 42-57, 2012.
- VAN DER KOOIJ, Rimmert BK et al. The effect of temporary speed restrictions, analyzed by using real train traffic data. **Transportation research procedia**, v. 22, p. 580-587, 2017.

WANG, Xi et al. Intelligent operation of heavy haul train with data imbalance: A machine learning method. **Knowledge-Based Systems**, v. 163, p. 36-50, 2019.

**XGBoost**. Disponível em: <https://xgboost.readthedocs.io/en/latest/index.html#>. Acesso em 30 de setembro de 2020.

XIA, Yufei et al. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. **Expert Systems with Applications**, v. 78, p. 225-241, 2017.

XIONG, Zheng et al. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. **Computational Materials Science**, v. 171, p. 109203, 2020.

YU, Zhiwen et al. Hybrid k-nearest neighbor classifier. **IEEE transactions on cybernetics**, v. 46, n. 6, p. 1263-1275, 2015.

ZHANG, Shichao et al. Efficient k-nn classification with different numbers of nearest neighbors. **IEEE transactions on neural networks and learning systems**, v. 29, n. 5, p. 1774-1785, 2017.