Previsão de duração de paradas de linha em uma indústria automotiva

Eduardo Viana, Gabriel Lenon, Leonardo Guerra, Wendel Vanderley

1 INTRODUÇÃO

1.1 Contextualização

Processos de Manufatura Avançada (P.M.A.) necessitam de rigoroso controle da cadeia produtiva e de qualidade dos itens a serem fabricados [1]. Ferramentas de gestão de processos foram concebidas para auxiliar na análise de problemas, bem como na busca de possíveis causa-raízes e motivos que levaram a ocorrência de falha na cadeia de produção[2], [3].

Com o advento da Manufatura 4.0, sistemas informatizados que atuam em tempo real com o processo de produção, chamados de *Manufacturing Execution System* (M.E.S., ou Sistema de Execução da Manufatura, tradução-livre do inglês) surgiram, a fim de garantir o supervisão do fluxo fabril através de sistemas computacionais[1]–[4].

Sendo assim, ao longo do processo de manufatura automotiva, surgem uma grande quantidade de dados e registros das operações, peças e consumo de itens utilizados no processo de fabricação, recursos esses que se tratados de maneira adequada e disponibilizados para tomada de decisão do gestor do processo, podem facilitar qual estratégia de negócio e de produção deve ser tomada[4]–[8].

As ferramentas de gestão de processo possuem certas limitações com relação a análise de grande volume de dados, o que torna as tomadas de decisão restritas à análises de maneira empírica, dado experiências pontuais com os problemas e falhas que ocorrem na cadeia de produção, onde pode se considerar ou não a repetibilidade das ocorrências[9]. Em contrapartida, um aspecto positivo da utilização de sistemas integrados da Manufatura 4.0 é a possibilidade da coleta dessa informação, pois ela gera padrões que podem ser encontrados, analisados e assim serem tratados de certa forma com técnicas computacionais que tem por base a mineração dos dados[2], [4], [5], [7], [10].

1.2 Descrição do Problema

Interrupções em processos de manufatura ocorrem por diversos motivos, e cada um possui o que é chamado de *modo de falha* [2], ou seja, como a falha veio a ocorrer. Em uma linha de produção contínua, esses modos de falha são motivo para registro de paradas de linha, e assim perda financeira.

Em uma linha de produção de uma indústria do setor automotivo, o processo contínuo de fabricação de veículos automotores pode sofrer de paradas de linha por registro dos colaboradores de cada estação de trabalho, por parada de ferramenta, por falta de qualidade das peças, por falta de entrega de peças no tempo correto [8]. Registros são coletados do tempo inicial, final, de qual tipo de parada são coletados, porém quando

aplicadas as ferramentas de gestão de processos, muitas delas ficam restritas a análises pontuais do problema, onde se considera ou não sua recorrência, o que depende de cada caso.

1.3 Objetivos

1.3.1 Objetivo Geral:

Previsão do perfil e do tipo de parada de linha (quanto tempo vai durarão as paradas de linha no instante em que são geradas, se causada por operador ou máquina), dada entrada de um evento de parada de linha que ocorre em tempo real.

1.3.2 Objetivos Específicos:

1.3.2.1 Tendência de parada de linha de acordo com as operações executadas nas estações de trabalho dos colaboradores;

1.3.2.2 Projeção das possíveis paradas de linha que possam impactar, com maior ou menor tempo, futuras ocorrências;

1.3.2.3 Modelo de previsão retornar valores com a maior precisão possível de estimativa de tempo para longas paradas de linha, pois assim será possível mobilizar o gestor do negócio e o gestor da manutenção nas melhores estratégias para retorno das condições de base da linha de produção.

1.4 Justificativa

A linha de produção da indústria a ser analisada possui capacidade produtiva de 1.040 carros/dia sendo que, por conta de paradas de linha, a mesma atinge 920 carros/dia de produção real. É de interesse do operador do negócio o estudo, dos últimos 4 anos (de 2017 até 2020), de todas as paradas de linha, a fim de que se encontre oportunidades de ganho de produção e tempo na execução das operações em todo o processo, bem como reduzir o tempo ocioso de parada de linha, ou seja, procurar entender os modo de falha, para reduzir sua ocorrência. Estima-se um ganho de 10%, com desvio percentual de 2%, no valor final de carros/dia no que se refere a produção real.

1.5 Escopo Negativo

1.5.1 Não estamos contemplando neste trabalho a justificativa do modo de falha, que originou a parada de linha, pois os dados coletados da justificativa do modo de falha não possuem a qualidade e quantidade esperada de informação (registros incompletos, registro somente do ano de 2019 e 2020 de produção);

1.5.2 A tomada de decisão será realizada exclusivamente pelo dono do negócio, não cabendo ao trabalho desenvolvido neste artigo essa atribuição;

1.5.3 Os dados aqui apresentados, bem como os gráficos oriundos do estudo de caso, não vão expor informações sensíveis (dados do comprador do veículo, dados do colaborador que executa atividade na estação de trabalho).

2 FUNDAMENTAÇÃO TEÓRICA

Para abordagem das informações que são geradas a todo momento no ambiente industrial, a metodologia CRISP-DM, que é a abreviação de *Cross Industry Standard Process for Data Mining* que, trazendo para o português, pode ser entendida como processo padrão da indústria cruzada para mineração de dados, foi elaborada. Essa é uma metodologia capaz de transformar os dados da empresa em conhecimento e informações de gerenciamento.

Criada na década de 1990, a CRISP-DM surgiu da necessidade dos profissionais de Data Mining de padronização do processo de análise e mineração de grande quantidade de dados. Apesar de existir uma série de ferramentas capazes de nortear esses profissionais, quando o assunto é Big Data e o seu grande volume de dados, elas deixam a desejar. O CRISP-DM surgiu justamente para atender aos projetos que estão diretamente envolvidos com o processamento e a análise de um grande volume de dados. [11]

O DM faz parte de Data Science, que utiliza estatística e matemática como base para cruzamento de dados, por meio de técnicas de indução para propor hipóteses e solucionar questões empresariais. Simplificando, é a mineração de dados que vai conseguir transformar todo o volume de dados em informações úteis para o gerenciamento e a tomada de decisões.

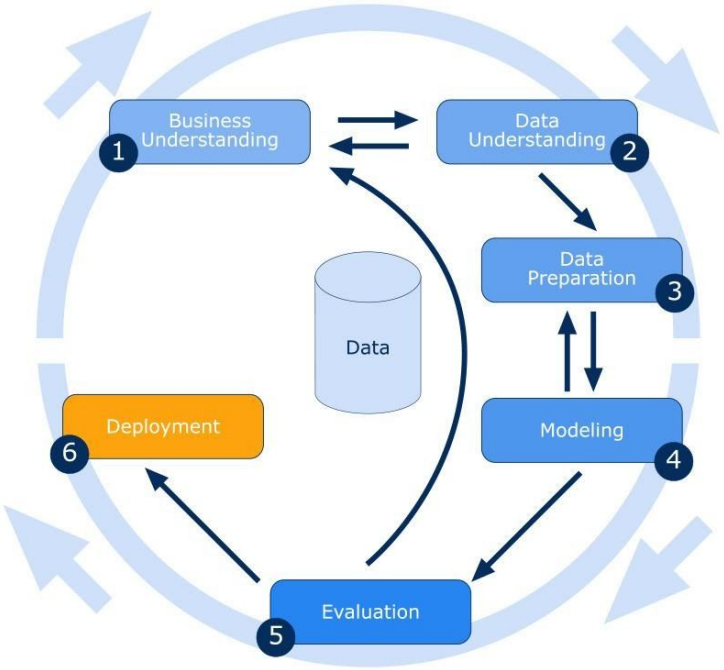

Figura 6: Ciclo de gerenciamento e implementação do CRISP-DM. Fonte: https://semantix.com.br/como-explorar-e-gerenciar-dados-com-o-crisp-dm/

O CRISP-DM reúne as melhores práticas para que a mineração de dados seja a mais produtiva e eficiente possível, analisando por exemplo dados financeiros, de recursos humanos, produção, hábitos dos clientes e outros, para propor modelos de melhoria ou solução de problemas [11]. Para melhor entendimento do processo, foram criadas etapas organizacionais, das quais:

1) Entendimento do problema: A primeira coisa a ser feita é entender de fato qual o problema a ser resolvido, buscando todos os detalhes sobre o impacto dele na empresa e quais os objetivos em relação ao trabalho;

2) Análise dos dados: Essa etapa consiste em organizar e documentar todos os dados que se encontram disponíveis. É aqui que começa de fato o trabalho de mineração de dados, pois o profissional deve ser capaz de identificar quais são os dados importantes para a resolução do problema. Nesse momento, o lado investigativo deve entrar em campo, para que os dados revelam problemas, soluções e tendências dos negócios;

3) Preparação dos dados: Agora que os dados já foram identificados, documentados e analisados, é hora de aplicar a parte técnica de análise deles. Serão preparadas as databases e definidos os formatos e questões técnicas da análise. Nessa etapa, é feita a escolha dos dados que serão trabalhados e de como eles serão cruzados para resolver o problema da empresa;

4) Modelagem: É nesta fase que são aplicadas de fato as técnicas de Data Mining, com base nos objetivos identificados no primeiro momento. A partir de agora, a mineração de dados pode ser associada a análises preditivas, para que a empresa prepare-se para o futuro, resolvendo a questão principal. Como? Os dados minerados podem ser usados para alimentar algoritmos que preveem as tendências dos negócios;

5) Avaliação: Trata-se de um momento muito importante, pois é quando serão acompanhados os resultados em relação aos objetivos e também à aplicação dos conhecimentos obtidos com a mineração de dados realizada, e para apreciação do gestor do negócio realizar a tomada de decisão final;

6) Implementação dos modelos na empresa: Nesta última fase é onde tudo que foi obtido de conhecimento dos dados são entregues de forma a ser aplicada. A partir disso, podem ser mudados os processos dentro da organização e criados novos produtos — tudo com base em dados, garantindo, assim, o sucesso dos negócios.

2.1 Área do Negócio

A busca por competitividade entre grandes marcas no mercado, bem como a constante melhoria dos processos de fabricação, faz com que novos modelos e estruturas de organização da manufatura venham a surgir como resultado final desses fatores citados. Modelos de produção em grande escala, com definições de tarefas a serem executadas por mão-de-obra, seja ela qualificada ou não, foram conceituados à luz da revolução industrial ocorrida no século XIX e XX [12].

Dos modelos de organização industrial existentes, três exemplos se destacam, dos quais: Taylorismo, Fordismo e Toyotismo. Do primeiro citado, o Taylorismo, temos a criação do conceito de Administração Científica de um processo de produção [12], que se caracteriza por:

1. Priorização do métodos científicos, em face dos empíricos, para atribuição e execução de atividades humanas;
2. Recrutamento do melhor colaborador para cada parte do processo de manufatura, ou seja, um colaborador/um local de trabalho/uma tarefa;
3. Capacitação constante do colaborador;
4. Cooperação entre líderes e liderados, onde líderes planejam e verificam a maneira como é exercida a função atribuída para o colaborador;
5. Estudos de tempos e movimentos dos colaboradores, a fim de termos a eficiência e eficácia garantida para o processo a ser executado.

Seguindo as definições, o Fordismo foi conceituado por Henry Ford como uma adaptação do Taylorismo: com a adição de um processo contínuo e ininterrupto, a exemplo uma “esteira rolante”, ditando assim um novo ritmo de trabalho. Com isso, Henry Ford conseguiu reduzir ainda mais o tempo no processo de manufatura, além de permitir a introdução de maquinário trabalhando em conjunto com seres humanos, como também fazendo com que fornecedores de peças e itens a serem montados no produto ficassem localizados cada vez mais próximos do processo de montagem final [12].

Com a mudança no perfil de consumo das pessoas nos séculos XX e XXI, um sistema surgido no Japão pós-segunda guerra mundial, chamado Toyotismo, emergiu com o conceito de fabricação simples e enxuta [12]. Nesse modelo de produção, os lotes de produção são baseados na demanda dos clientes, não possuindo assim estoque como no Taylorismo e Fordismo, o que já evita desperdício de materiais e mão-de-obra. Não existe estoque de produção, o controle de tempo de execução das atividades e qualidade é feito desde o começo do processo de produção até o produto final. Além disso, o colaborador executa várias tarefas em escala (multitarefas) e existe uma integração maior com o espaço de trabalho. Também é levado em conta aspectos ergonômicos, de segurança e de qualidade de execução da tarefa no Toyotismo.

Para o setor automotivo, o modelo Toyotista ganhou cada vez mais espaço, também com o advento da manufatura enxuta [13], que pelo seu conceito visa cada vez mais a melhoria do processo com o controle e diminuição dos custos e tempo de produção, mantendo a qualidade de todo o processo, em conjunto com a manufatura 4.0, recebendo esse nome por ser contemporânea a 4ª revolução industrial – uso de tecnologias de ponta e processos computacionais [2], [14]. Sendo assim, possui como meta a integração dos processos de manufatura a sistemas computacionais de controle de produção, qualidade e tempo de operação na linha contínua. Adicionalmente, o conceito de Manufatura de Classe Mundial (do inglês *World Class Manufacturing* – W.C.M.) surge já na década de 1980 para integrar as diversas definições de melhoria de processo de produção da manufatura, visando sempre o controle do tempo de execução, qualidade, segurança do colaborador e entrega eficaz e efetiva do produto final para o cliente [14].

Portanto, a proposta de melhoria de processo por paradas de linha de produção, bem como a diminuição desse tempo à luz do que ocorre em toda a cadeia de produção de uma empresa do setor automotivo, surge como proposta para implementação a mineração de dados coletados pelos sistemas de controle do processo de manufatura, a fim de que sejam encontrados padrões de perdas produtivas, com proposta através da análise realizada de diminuir os impactos na produção, em conjunto com o uso de ferramentas empíricas de gestão de processos, como o W.C.M.

2.2 Mineração de Dados

Mineração de Dados (Data Mining, em tradução livre) é o nome dado a um conjunto de técnicas e procedimentos que tenta extrair informações de nível semântico mais alto a partir de dados brutos, em outras palavras, permitindo a análise de grandes volumes de dados para extração de conhecimento [1], [3].

Este conhecimento pode ser na forma de regras descritivas dos dados, modelos que permitem a classificação de dados desconhecidos a partir de análise de dados já conhecidos, previsões, detecção de anomalias, visualização anotada ou dirigida, entre outros. Embora muitas destas abordagens tratem com dados tabulares, é possível extrair informações tabulares de dados estruturados de forma diferente (como encontrados na Web) ou mesmo usar algoritmos específicos para minerar dados da Web como conjuntos de links entre documentos, textos e outras aplicações [11].

Para identificarmos quando uma base de dados é adequada para uma mineração de dados ou não, o conjunto de dados deve representar, de forma confiável, o universo (mundo real) a ser investigado, possibilitando assim inferir a situação problema como um todo, seja pela perspectiva de completeza ou complexidade do problema. Um dos “mitos” criados, a partir da motivação inicial das discussões sobre a mineração de dados, era ser uma alternativa para “grandes” bases de dados [4], [9].

Este fato decorreu da própria dificuldade de processamento inerente à descoberta e identificação de informações oportunas ao processo decisório em grandes conjuntos. Mas, dispor de um grande conjunto não constitui requisito obrigatório desde que este represente o universo, por amostragem ou não. Portanto, quando a mineração de dados resultar na

descoberta de elementos potencialmente úteis para o apoio a tomada de decisão, pode-se afirmar que a base de dados atendeu ao esperado para solução proposta a ser resolvida.

Como muitas bases de dados sofrem interferências de diversas formas, a elas atribuimos o nome de ruído. “Ruído” representa conteúdo nas bases de dados que pode prejudicar a qualidade da informação extraída, a partir de qualquer método, seja ele tradicional ou baseado em estratégias mais elaboradas. Destacam-se como ruídos: valores fora do domínio, ausência de valores, inconsistências, dentre outras [9], [11].

É importante lembrar que o mundo real é ruidoso, ou seja, se uma base de dados representa uma abstração deste mundo real, esta será ruidosa a despeito dos esforços despendidos para a sua modelagem e respectiva população. Cabe aos profissionais da área de tecnologia da informação minimizar o impacto negativo que estes ruídos possam representar nas informações extraídas e disponibilizadas aos gestores. Por exemplo, todas as vezes que, ao informar os dados cadastrais se omite ou não se informa corretamente a renda, gera-se um ruído no conjunto de dados [11].

2.3 Trabalhos Relacionados

Processos de Manufatura Inteligente demandam aplicações de alta complexidade, no que se refere a estudos de caso que indiquem soluções para, por exemplo, problemas de falta de qualidade no item manufaturado, ou até mesmo paradas de produção provocadas por falha em equipamentos. A integração de sistemas de automação com soluções de Internet das Coisas (do inglês *Internet of Things, IoT*) permite a aquisição de grande volume de informações, o que muitas vezes deixam de ser analisados por falta de conhecimento para tal [15]. Para esses e outros casos ligados a otimização da produtividade em indústria e processos de manufatura, a mineração de dados surgiu com o propósito de extração de informações, que eram desconhecidas até então, e assim torná-las válidas, bem como compreensíveis, de grandes bases de dados com o intuito de melhorar e otimizar as tomadas de decisão que devem ser feitas pelo gerente do negócio [9].

Técnicas diversas para extração de informação de grande quantidade de dados, tais como *Random Forest*, redes *Bayseanas*, algoritmos de classificação, redes neurais artificiais (da sigla RNA), árvores de decisão, sistemas de lógica *Fuzzy* e algoritmos genéticos, são utilizadas como propostas de solução, dependendo do cenário que exige solução de problema proposto [4], [6]–[8], [16].

Em se tratando da abordagem *Random Forest*, este método consiste nas correlações que são estimadas de acordo com os resultados de saída de várias árvores de decisão, onde esses diversos resultados das saídas são combinados, para assim termos uma saída única. Cada resultado de saída corresponde a uma “escolha” que é feita pela árvore de decisão, em sua unidade, sendo que cada árvore trabalha com um trecho de amostras da base de dados, onde esses trechos são aleatórios e de mesmo tamanho, para que não haja desbalanceamento das sub-bases. A combinação de todas as escolhas denota na escolha global, onde cada árvore de

decisão que participa do processo possuem igual poder de decisão no valor de saída final, sendo esse do *Random Forest*, correspondente então a base de dados completa [17].

Aplicações do Random Forest para detecção de defeitos em soldas feitas por robô [17], bem como detecção de falhas em camadas de dispositivos semicondutores em sua fase de fabricação [16] são exemplos de situações onde o *Random Forest* foi aplicado para previsão de situações que estavam fora do padrão esperado, para controle do processo de manufatura avançada e de qualidade exigido em cada cenário citado.

3 MATERIAIS E MÉTODOS

3.1 Descrição da Base de Dados

A Base de dados analisada possui, em tamanho total, 1.2GB de informações, contendo na mesma 12 colunas com 579.237 linhas de informações acerca dos mais variados tipos de operações que resultaram em paradas de linha. Foram levantados dados de 2018 até 2020 de histórico de paradas de linha, com extração realizada em um banco de dados SQL. Foi montado, assim, um dicionário de dados que segue:

DICIONÁRIO DE DADOS - Análise de paradas de linha de produção em uma indústria automotiva

NOME DO CAMPO	TIPO DO DADO	DESCRIÇÃO DO CAMPO
ProductionOrderId	**Long**	*Número de batismo do veículo.*
CallStartTime	**Time**	*Tempo que marca o início do alerta de da parada de linha.*
CallEndTime	**Time**	*Tempo que marca o fim de parada de linha.*
Time_DIFF	**Time**	*Diferença de tempo entre CallEndTime - CallStartTime*
StopLineStartTime	**Time**	*Tempo que marca o início de parada de linha.*
StopLineEndTime	**Time**	*Tempo que marca o fim de parada de linha.*
StopButtonStartTime	**Time**	*Tempo que marca o início da solicitação de parada de linha*
StopButtonEndTime	**Time**	*Tempo que marca o fim da solicitação de parada de linha*
Workplace	**String**	*Código referente a estação de trabalho onde houve parada de linha.*
Type	**String**	*Tipo de parada: se causada pelo operador, se a parada da linha foi causada por falha na máquina.*
Type_BIN	**Long**	*0 = Operator; 1 = Fail*
Operation	**String**	*Tipo de operação que houve parada.*

Tabela 1: Dicionário de dados da tabela utilizada para análise de paradas de linha de produção de uma indústria do setor automotivo.

3.2 Análise Descritiva dos Dados

Temos, de parte dos atributos numéricos, os tipos de dados sinalizados como *Time*, que correspondem a todos os tempos de início e fim de parada de linha na produção. Já os que estão indicados como sendo de atributo nominal, que são do tipo *String*, representam o local onde houve a parada de linha, o modo de parada – se parada executada por ação humana, ou máquina – e qual tipo de operação foi impactada com a parada de linha.

Em relação a este último – nome do campo *Operation* – temos a classificação em: parada por conta de ferramentas (índice SCR, GEN, F), paradas por falta de qualidade na operação (tipos QCE), parada por falta de finalização do processo (tipos PCE) e paradas por falta de leitura de peças montadas (TRA, TRP, OGG, CERT, CEL). O tipo de dado que ficou sinalizado como *Long* representa o número de batismo da carroceria, o mesmo que vai na documentação do carro e fica atrelado ao chassis do mesmo. Toda vez que existe uma parada de linha na produção, o valor de ínicio, fim, qual estação de trabalho e qual operação sofreu impacto é registrado no sistema de manufatura.

Através da tabela de informações sobre as paradas de linha, foram escolhidos como atributos numéricos a diferença de tempo entre o início e fim da parada (chamado de *CallStartTime* e *CalllEndTime*, respectivamente) e como atributo de classe se a parada de linha foi gerada por operador (nomeada de *Operator*), ou por falha no processo (Nomeada de *Fail*). Com relação ao atributo numérico do tempo de parada, as seguintes médias foram extraídas:

FALHA DE OPERADOR	
Médias (atributo numérico)	
MÉDIA	426.3555138
MEDIANA	42
AMPLITUDE	86391
MODA	0
VARIÂNCIA	10912639.59
DESVIO-PADRÃO	3303.428459
DESVIO ABSOLUTO	657.6382508
VALOR MÁXIMO	86391
VALOR MÍNIMO	0

Tabela 2: Valores das médias calculadas, com relação ao atributo numérico escolhido, neste caso a diferença de tempos entre a parada final e inicial, para modo de falha por Operador.

FALHA DE MÁQUINA	
Médias (atributo numérico)	
MÉDIA	62.45367921
MEDIANA	6
AMPLITUDE	85897
MODA	1
VARIÂNCIA	1756192.681
DESVIO-PADRÃO	1325.214202
DESVIO ABSOLUTO	98.33370661
VALOR MÁXIMO	85897
VALOR MÍNIMO	0

Tabela 3: Valores das médias calculadas, com relação ao atributo numérico escolhido, neste caso a diferença de tempos entre a parada final e inicial, para modo de falha por Máquina.

Extrações de gráficos também foram realizadas, de tal forma que foram expostos na primeira análise de tempo, todos os valores de tempo encontrados, em segundos.

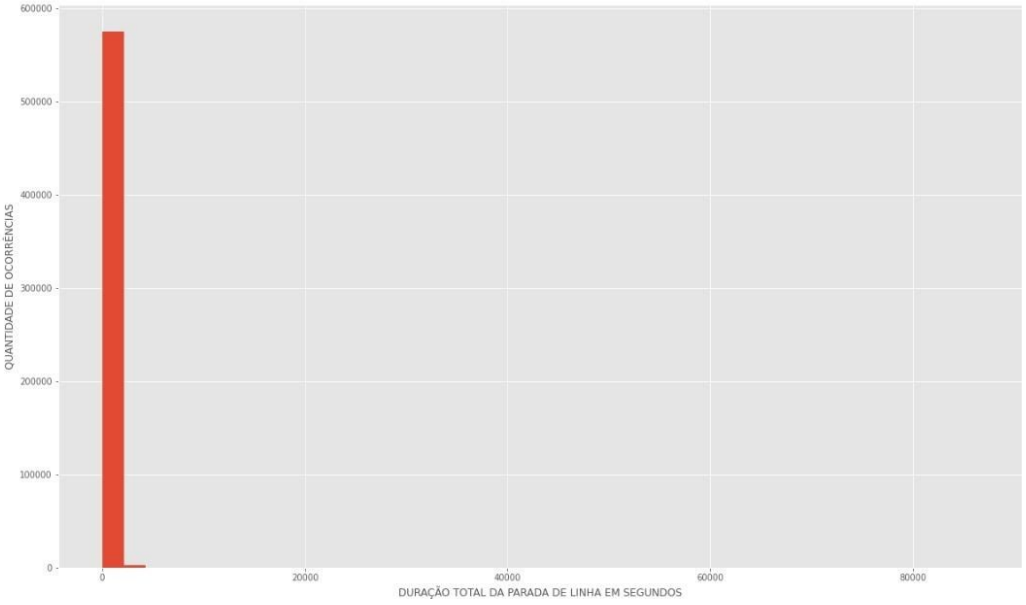

Figura 1: Análise do total de ocorrências de paradas de linha, em segundos.

Diante de melhor análise a se tornar evidente, foi realizada nova extração, essa com os primeiros 1000 segundos de paradas, pois foi percebido que as paradas de linha estavam mais concentradas nesse intervalo de tempo.

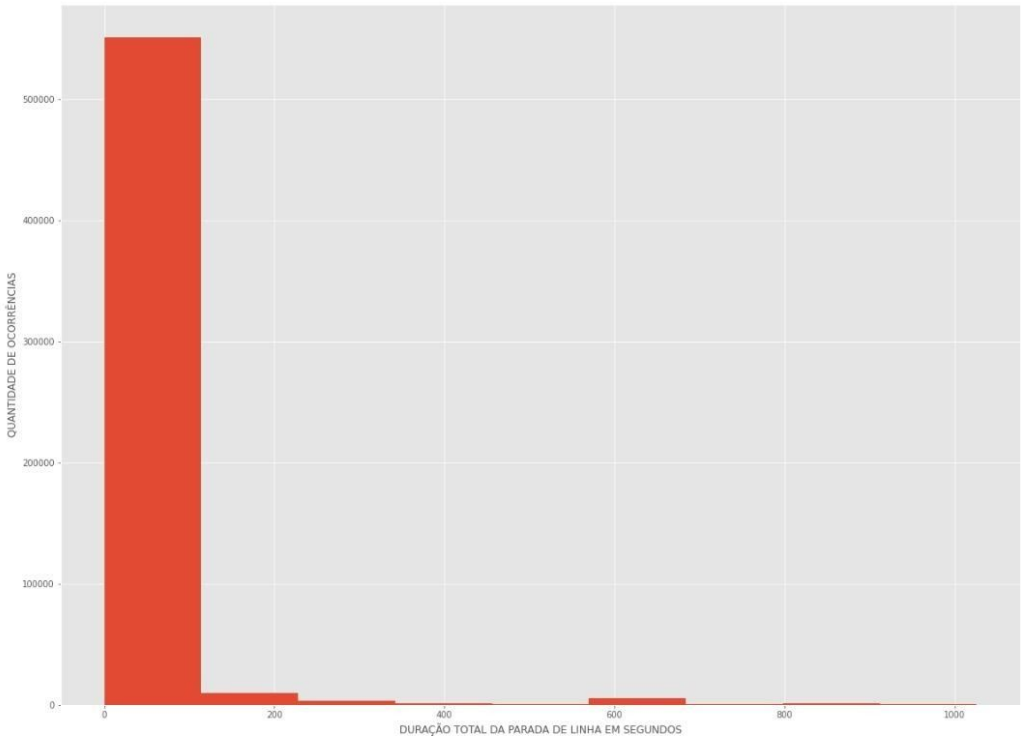

Figura 2: Análise dos primeiros 1000 segundos de parada de linha.

Logo em seguida, para refinamento ainda mais detalhado da informação, foram eviidenciados os 50 primeiros valores medidos de tempo, também em segundos, da quantidade de ocorrências de parada de linha de produção (ou seja, quantas vezes uma parada de 1 segundo influenciou no total de paradas).

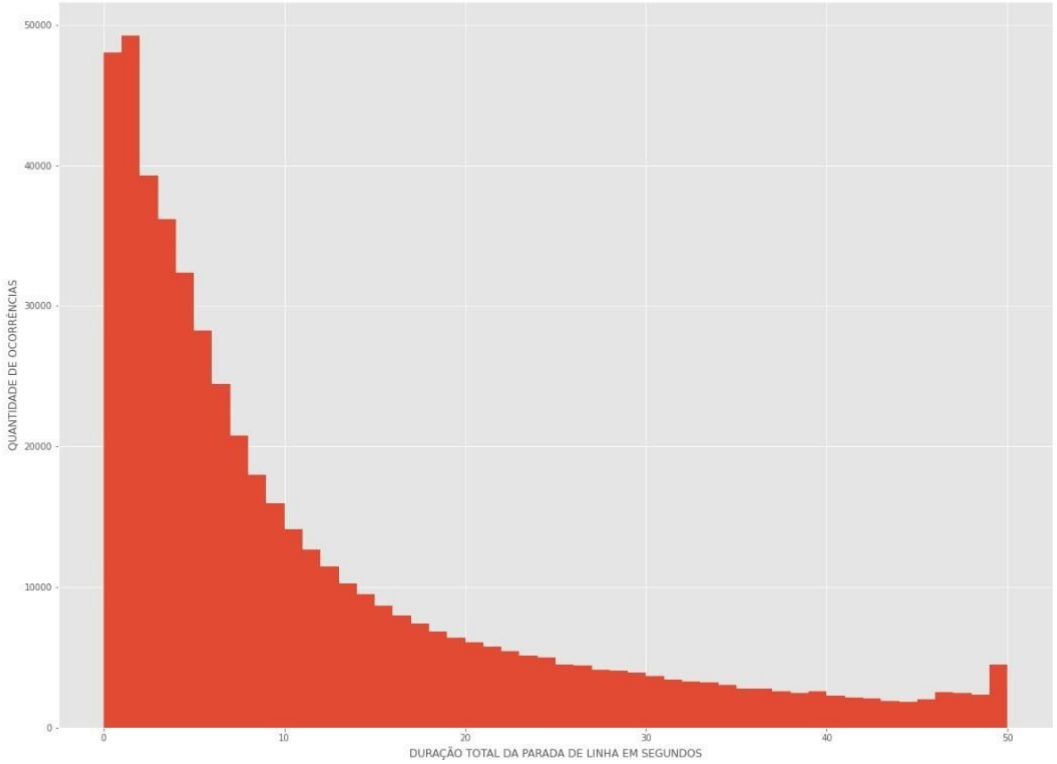

Figura 3: Análise dos 50 primeiros valores de tempo, em segundos, correspondente as ocorrências de paradas de linha.

Temos, como análise desses 45 primeiros segundos, a seguinte quantidade de frequência de valores (onde o primeiro remonta ao tempo, em segundos, e o segundo valor a frequência de ocorrência desse tempo de parada de linha):

[(1, 49192), (0, 47993), (2, 39289), (3, 36201), (4, 32371), (5, 28230), (6, 24470), (7, 20761), (8, 17998), (9, 15984), (10, 14139), (11, 12687), (12, 11472), (13, 10245), (14, 9506), (15, 8650), (16, 7997), (17, 7388), (18, 6832), (19, 6370), (20, 6098), (21, 5782), (22, 5433), (23, 5133), (24, 4984), (25, 4499), (26, 4441), (27, 4094), (28, 4032), (29, 3902), (30, 3663), (31, 3390), (32, 3300), (33, 3209), (34, 3051), (35, 2803), (36, 2765), (37, 2595), (39, 2567), (46, 2544), (38, 2495), (47, 2464), (48, 2328), (49, 2277), (40, 2266), (50, 2201), (41, 2155), (42, 2089), (51, 2030), (45, 2001)]

Foram também amostradas informações dos valores obtidos na forma de box plot, para análise dos grupos de quartis mais recorrentes.

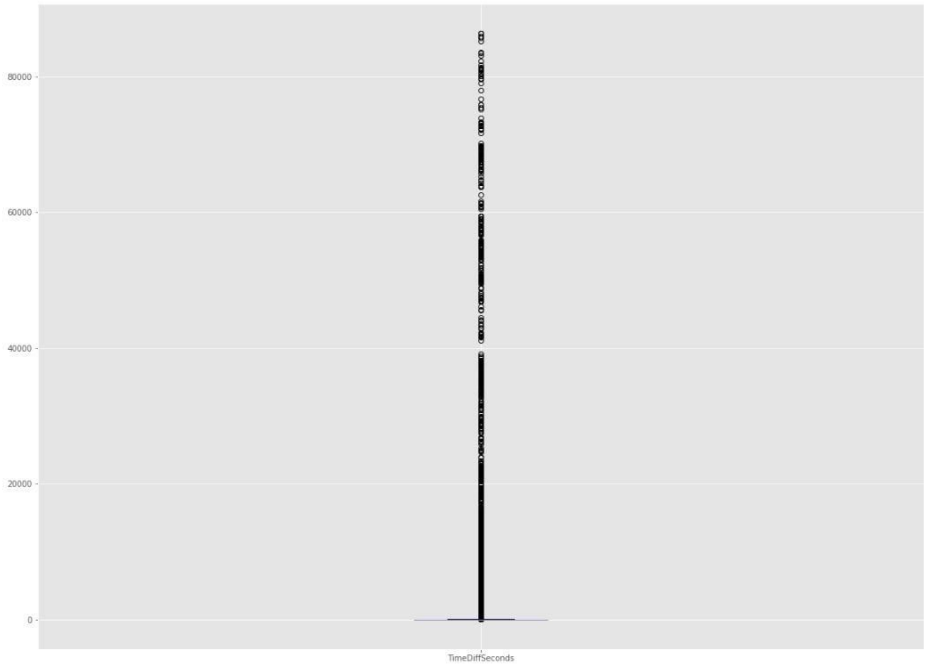

Figura 4: Análise, em box plot, de todo o conjunto de dados coletados, com *outliers*.

Com a concentração dos valores nos primeiros 1000 segundos, uma extração mais detalhada, em box plot, foi feita para exclusão dos valores que são outliers.

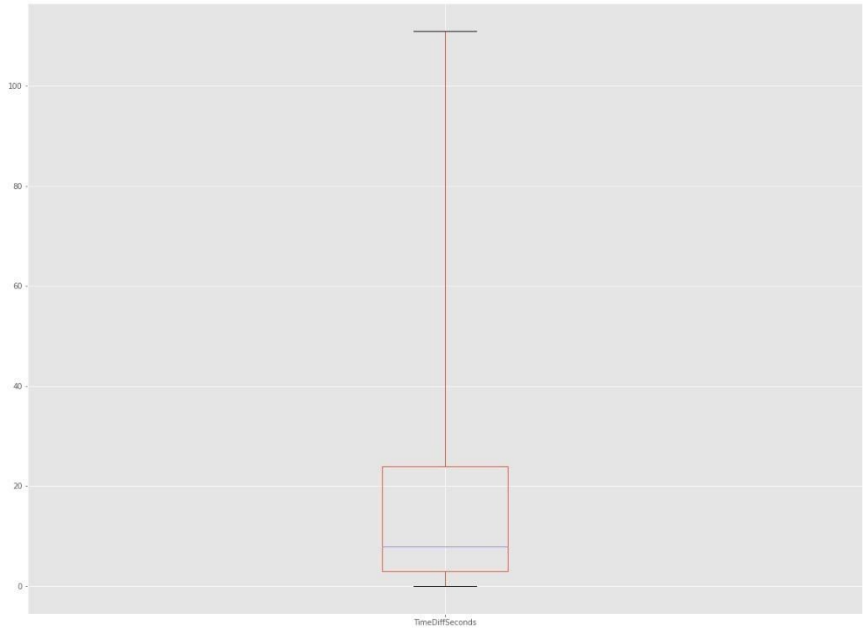

Figura 5: Análise, em box plot, de todo o conjunto de dados coletados, sem *outliers*.

3.3 Pré-processamento dos Dados

Por critério de seleção, foram escolhidas as colunas *CallStartTime* e *CallEndTime* para análise a diferença de tempo de parada de linha na produção, e assim criada uma nova coluna chamada *Time_Diff.* Também foi realizada a criação de uma coluna entitulada *Type_BIN*, que corresponde a informação binária da coluna Type, onde o índice 0 corresponde a todas as paradas relacionadas a *operator* e o índice 1 a todas as paradas relacionadas a *fail* das operações, constando sempre o tipo de operação que provocou a parada, na coluna *Operation.*

Para também melhorar a execução de nosso algoritmo, como também termos uma categorização das classes de maneira mais adequada, foram criados labels das colunas trabalhadas nesta base de dados, dos quais:

1) Workplace_Category: label encoder de workplace, que significa a estação de trabalho do colaborador/máquina;

2) Operation_Category: label encoder de operation, ou seja, a operação que é executada na estação de trabalho;

3) Criação de StopLineShift (período do dia da parada): com label enconder, categorizada para StopLineShift_Category, se no primeiro, segundo ou terceiro turno de produção;

4) Criação de StopLineCriticality (criticidade da parada): com label encoder, categorizada para StopLineCriticality_Category, criticidade da parada de linha, em que organizamos de acordo com as classes a serem trabalhadas;

5) Chassis_Category - label enconder de ProductionOrderId, que corresponde a um componente em parte da numeração do chassis da carroceria que foi impactada com a parada de linha.

Também foram necessárias algumas transformações de colunas e variáveis, para correto ajuste da base de dados disponibilizada ao modelo de previsão, que foram as seguintes:

1) TimeDiffSeconds: transformação de TIME_DIFF em segundos;

2) Criação da coluna Time_diff_seconds_bins: distribuição da classe TimeDiffSeconds em ranges de tempo, para melhor análise da abordagem de *Random Forest* na base de dados;

3) Colunas da base de dados utilizadas para treinamento do modelo: *["Chassis_Category", "Type_BIN", "Workplace_Category", "Operation_Category", "Time_diff_seconds_bins", "StopLineShift_Category", "StopLineCriticality_Category"]*

Finalmente, temos a organização tanto da base de dados, como dos gráficos gerados e do treinamento e validação do modelo no formato de arquivos de processamento, debug e geração de dados no formato de plataforma colaborativa, como exemplo a que foi utilizada, o *Google Colab*. Na plataforma em questão, foram elencados coluna para as seguintes distribuições:

1) Turno de ocorrência da parada de linha, se primeiro, segundo ou terceiro turno de produção;

2) Criticidade da parada de linha, de acordo com a quantidade, em segundos da parada de linha. Com essa classificação, conseguimos mapear as ocorrências pelo tempo de duração. Os indicadores para criação dessa nova coluna foram sinalizados como:
    a. Se tempo de parada entre 0 e 540 segundos, criticidade baixa;
    b. Se tempo de parada entre 541 e 1140 segundos, criticidade média;
    c. Se tempo de parada entre 1141 e 1740 segundos, criticidade alta;
    d. Se tempo de parada entre 1741 e 2340 segundos, criticidade altíssima;
    e. Se tempo de parada maior que 2341 segundos, criticidade desastre;

3.4 Metodologia Experimental

Primeiramente, a coleta de informações foi realizada através de um arquivo .csv disponibilizado pelo *stakeholder*, para fins de análise e tratamento das linhas, quando necessário, com características já citadas em 3.1. Logo após, foram definidas quais transformações seria necessárias para adaptação da base de dados, e assim ter a possibilidade de manipulação o mais adequada possível dos dados, tanto para a aquisição de gráficos relativos a paradas de linha de produção, e sua quantidade de paradas de acordo com a estação de trabalho, por exemplo, como para o modelo de previsão proposto, para nosso cenário o *Random Forest,* conforme mencionado em 3.3.

Em seguida, foi realizado o processamento da base, de acordo com os atributos de classe escolhidos e assim categorizados, para melhor análise e performance do modelo utilizado. Com isso, foi possível a extração total da quantidade de paradas de linha existentes, do tempo que cada uma dessas paradas de linha influenciou na linha de produção, a tendência de maior ocorrência de paradas de linha, por exemplo, se quando com tempo menor que 60 segundos ou maior que 60 segundos.

Finalmente, já para o modelo de previsão utilizado, as colunas criadas em 3.3 suportaram a estimativa dos valores encontrados, tanto para a parte da base de dados que foi utilizada para treinamento do modelo, como para validação do mesmo. Foi feita uma distribuição de classes, onde os períodos de tempo que houveram maiores ocorrências de parada de linha recebem o menor intervalo de classes, para assim a previsão ocorrer o mais próxima do esperado, em se tratando da realidade de paradas de linha da produção, e um balanceamento da quantidade de amostras de cada classe, de tal forma que para a base de treinamento ter a mesma quantidade de amostras da classe com menor quantidade de amostras, que no caso foi a classe 10 com 2410 amostras, independente do intervalo de classes, fazer o treinamento da previsão o mais balanceado possível.

Adicionalmente, das classes com mais amostras foi realizado um permuta de índices, para validação do treinamento. Com o proposto, conseguimos uma estratificação de 12 intervalos de classe, que nos permite assim identificar melhor os padrões de criticidade de parada de linha. Os intervalos de classes e de registros foram:

CLASSE	INTERVALO	REGISTROS	CLASSE	INTERVALO	REGISTROS
0	0s - 10s	326.628	6	91s - 120s	8.070
1	11s - 20s	87.245	7	121s - 180s	6451
2	21s - 30s	45.963	8	181s - 360s	5.903
3	31s - 45s	38.414	9	361s - 720s	7.832

4	46s - 60s	26.615	10	721s - 1440s	2.410
5	61s - 90s	19.298	11	+1440s	4.408

Tabela 4: Valores das classes, intervalo e de registros para a base de dados.

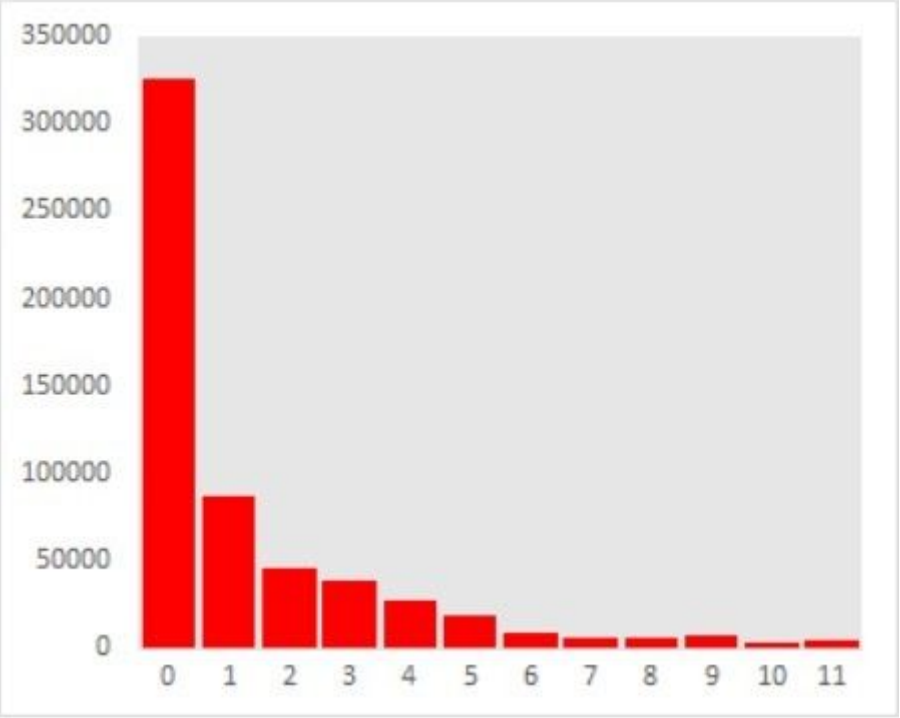

Figura 6: Gráfico contendo a estratificação da quantidade de amostras, de acordo com as classes separadas para previsão das paradas de linha.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

4.1 Resultados

Antes do balanceamento das amostras nas classes ser realizado, conforme comentado em 3.4, o modelo de previsão fez o processamento das informações contidas na base de dados, de acordo com a quantidade de amostras presentes em cada intervalo de classes, desconsiderando assim o mínimo valor de amostras apontado na classe 10 utilizado para, posteriormente, balancear a quantidade de amostras nas demais classes, de 2410 registros. Eis que os resultados obtidos pela precisão do modelo *Random Forest* seguem abaixo:

	precision	recall	f1-score	support
0	0.75	0.71	0.73	68834
1	0.19	0.21	0.20	16129
2	0.13	0.14	0.14	8374
3	0.14	0.15	0.14	7173
4	0.21	0.23	0.22	5074
5	0.11	0.12	0.12	3599
6	0.09	0.10	0.09	1432
7	0.08	0.10	0.09	1147
8	0.20	0.20	0.20	1192
9	0.68	0.67	0.67	1585
10	0.45	0.48	0.46	439
11	0.98	0.99	0.98	870
accuracy			0.51	115848
macro avg	0.33	0.34	0.34	115848
weighted avg	0.52	0.51	0.52	115848

Tabela 5: Valores de acurácia do modelo de previsão, sem o balanceamento da quantidade de amostras entre as classes.

Com a acurácia em 51%, foi necessário reanalisar o pré-processamento, para entendimento dos valores, e foi constatado que seria necessário uma nova distribuição de amostras entre as classes era necessária para uma nova validação do modelo de previsão, mantendo como meta os intervalos de tempo que possuem maiores ocorrências de parada de linha, como o menor intervalo intra-classe.

Sendo assim, a validação do modelo fica mais detalhada, de tal forma a saber se várias paradas de criticidade baixa influenciam com maior ou menor frequência do que uma parada longa que causa um valor alto de parada de linha de produção. Depois do balanceamento das classes de acordo com o menor valor de amostras registrado, que foi o da classe 10 com 2410 registros, nos intervalos de tempo propostos em 3.4, temos o seguinte resultado para a previsão:

	precision	recall	f1-score	support
0	0.40	0.32	0.35	300
1	0.18	0.17	0.18	260
2	0.15	0.18	0.17	204
3	0.19	0.19	0.19	241
4	0.25	0.24	0.25	246
5	0.15	0.16	0.15	215
6	0.16	0.17	0.17	225
7	0.24	0.23	0.23	245
8	0.35	0.33	0.34	256
9	0.51	0.56	0.53	223
10	0.71	0.72	0.71	237
11	0.99	0.99	0.99	240
accuracy			0.36	2892
macro avg	0.36	0.36	0.35	2892
weighted avg	0.36	0.36	0.36	2892

Tabela 6: Valores de acurácia do modelo de previsão, com o balanceamento da quantidade de amostras entre as classes, baseado na classe com menor quantidade de índices, no caso a classe 11, com 2410 registros.

	precision	recall	f1-score	support
0	0.51	0.31	0.39	1173
1	0.21	0.20	0.20	781
2	0.14	0.16	0.15	629
3	0.17	0.19	0.18	636
4	0.26	0.28	0.27	668
5	0.13	0.17	0.15	544
6	0.17	0.20	0.19	603
7	0.18	0.20	0.19	656
8	0.41	0.31	0.35	941
9	0.49	0.71	0.58	503
10	0.82	0.71	0.76	838
11	0.97	0.99	0.98	753
accuracy			0.38	8725
macro avg	0.37	0.37	0.37	8725
weighted avg	0.40	0.38	0.38	8725

Tabela 7: Valores de acurácia do modelo de previsão, com o balanceamento da quantidade de amostras entre as classes, baseado na classe com menor quantidade de índices, no caso a classe 11, com 2410 registros, variando em termos de amostras, conforme as já utilizadas em validação mostrada na Tabela 5.

A avaliação que é feita também para treinamento do modelo, considerou a substituição de valores onde tivemos classes com mais do que 2410 registros, e não houve mudança brusca nos valores de precisão atingidos, conforme tabelas 5 e 6 acima, que registraram 36% e 38% de previsão final, respectivamente. Houve considerável melhora nas previsões das classes 5 até a 11, que para o gerente do negócio torna-se mais interessante saber, por exemplo, com mais precisão quando paradas registradas na base de dados que são acima de 1 minuto podem impactar o processo de produção. Com isso, fica viável o planejamento de novas estratégias para mitigar os possíveis problemas registrados, sejam eles por falha de operador, equipamento ou processo de certificação de peças.

Em se tratando das paradas com menos de 1 minuto, por mais que ocorram com maior frequência, é de se saber que são paradas que ocorrem, pelo presente estudo, por falta de adequação do tempo-ciclo de operação dos colaboradores nas estações de trabalho, sendo assim necessitando de ajustes para melhorar a saturação dos postos de trabalho ao longo do processo produtivo, pois cada colaborador possui um rendimento individual a ser considerado.

4.2 Discussão

Com os resultados encontrados, o time de alta gestão da indústria estudada pode, então, ter tomadas de decisão de tal forma que as paradas de linha que mais impactam o processo são aquelas em que:

1. O templo-ciclo de execução da operação não está conforme o previsto para o ritmo da linha de produção, ou seja, curto espaço de tempo para várias atividades na estação de trabalho serem exercidas;
2. Falta ou reciclagem de treinamento e capacitação do colaborador em realizar as atividades, de acordo com a estação de trabalho;
3. O número de carros por hora que são fabricados na linha não está de acordo com a quantidade de colaboradores exigidos para execução da quantidade de atividades nas estações de trabalho (chamado também de saturação da linha de produção);
4. Quebras ocasionadas por ferramentas do processo precisam de maior monitoramento do time da Manutenção, com relação a ações de execução das atividades necessárias de manutenção preventiva/preditiva, para assim evitar cada vez mais ações corretivas nos dispositivos e equipamentos do processo produtivo, para o registro de longas paradas de linha. Portanto, podemos afirmar que maior velocidade da linha não significa mais carros produzidos, se temos mais quebras dos equipamentos na linha de produção;
5. O perfil de perda das paradas de linha, que antes era determinado por ferramentas de puro conhecimento empírico, agora pode ser encontrado através de um modelo de previsão, proposto neste artigo.

5 CONCLUSÕES E TRABALHOS FUTUROS

O perfil de perda a ser previsto auxiliará na tomada de decisão do gerente do negócio quando, por exemplo, eventos de paradas de programação para reparação dos equipamentos serão necessários de acontecer, para prevenir grandes paradas de linha. Como também um esforço maior em estudar os tempos e movimentos dos colaboradores nas estações de trabalho, a fim de melhorar a relação com o tempo-ciclo de operação.

É esperado, também, com os valores encontrados pela previsão, aumento na eficiência e eficácia em cada estação de trabalho, considerando que correções no tempo-ciclo de operação já citados sejam realizados, para assim atingir o objetivo da previsão mais próxima da realidade de modos de falha para as paradas de linha que ocorrem na manufatura.

Como proposta de trabalho futuro, a correlação dessa base de dados com outras bases de dados que coletam informações distintas de perdas produtivas, como perdas por atraso logístico externa a cadeia de produção, perdas por falha de outras oficinas ligadas a entrega de peças e carrocerias para manufatura, a saber da oficina de Prensas de chapas e do parque de fornecedores.

Finalmente, também segue a proposta da aplicação de outras técnicas de previsão, a saber séries temporais, algoritmos genéticos, e quaisquer outras técnicas de algoritmos de aprendizado de máquina ou de inteligência artificial, que possa retornar valores adequados para auxílio do gerente do negócio na tomada de decisão.

6 Referências Bibliográficas

[1] M. G. B, S. Chren, B. Rossi, and T. Pitner, *for Smart Grid Systems*, vol. 1. Springer International Publishing, 2019.

[2] M. Syafrudin, G. Alfian, N. L. Fitriyani, and J. Rhee, "Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing," *Sensors (Switzerland)*, vol. 18, no. 9, 2018.

[3] L. Rokach and O. Maimon, "Data mining for improving the quality of manufacturing: A feature set decomposition approach," *J. Intell. Manuf.*, vol. 17, no. 3, pp. 285–299, 2006.

[4] C. Gröger, F. Niedermann, and B. Mitschang, "Data mining-driven manufacturing process optimization," *Lect. Notes Eng. Comput. Sci.*, vol. 3, pp. 1475–1481, 2012.

[5] G. Filios, "An Agnostic Data-Driven Approach to Predict Stoppages of Industrial Packing Machine in Near Future," pp. 236–243, 2020.

[6] C. Dudas, M. Frantzén, and A. H. C. Ng, "A synergy of multi-objective optimization and data mining for the analysis of a flexible flow shop," *Robot. Comput. Integr. Manuf.*, vol. 27, no. 4, pp. 687–695, 2011.

[7] C. Dudas, A. H. C. Ng, L. Pehrsson, and H. Boström, "Integration of data mining and multi-objective optimisation for decision support in production systems development," *Int. J. Comput. Integr. Manuf.*, vol. 27, no. 9, pp. 824–839, 2014.

[8] A. Luckow *et al.*, "Artificial Intelligence and Deep Learning Applications for Automotive Manufacturing," *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018*, pp. 3144–3152, 2019.

[9] K. Wang, "Applying data mining to manufacturing: The nature and implications," *J. Intell. Manuf.*, vol. 18, no. 4, pp. 487–495, 2007.

[10] A. Majeed, J. Lv, and T. Peng, "A framework for big data driven process analysis and optimization for additive manufacturing," *Rapid Prototyp. J.*, vol. 25, no. 2, pp. 308–321, 2019.

[11] P. C. Ncr *et al.*, "Step-by-step data mining guide," *SPSS inc*, vol. 78, pp. 1–78, 2000.

[12] C. J. Muller, "A evolução dos sistemas de manufatura e a necessidade de mudança nos sistemas de controle e custeio," p. 222, 1996.

[13] L. C. Maia, a. C. Alves, and C. L. Leão, "Metodologias Para Implementar Lean Production: Uma Revisão Critica De Literatura," *ClLME´2011*, p. 0915A, 2011.

[14] F. Pereira, L. Carelli, C. Manuel, T. Rodriguez, and L. M. Rôa, "Proposta de adequação do processo de inspeção com base nos conceitos do lean manufacturing : estudo de caso em um fabricante de equipamentos agrícolas.," vol. 1, p. 86, 2016.

[15] D. Wu, C. Jennings, J. Terpenny, R. X. Gao, and S. Kumara, "A Comparative Study on Machine Learning Algorithms for Smart Manufacturing: Tool Wear Prediction Using Random Forests," *J. Manuf. Sci. Eng. Trans. ASME*, vol. 139, no. 7, 2017.

[16] L. Puggini, J. Doyle, and S. McLoone, "Fault detection using random forest similarity

distance," *IFAC-PapersOnLine*, vol. 28, no. 21, pp. 583–588, 2015.

[17] Z. Zhang, Z. Yang, W. Ren, and G. Wen, "Random forest-based real-time defect detection of Al alloy in robotic arc welding using optical spectrum," *J. Manuf. Process.*, vol. 42, no. April, pp. 51–59, 2019.