

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

**Predição de taxa de falha utilizando séries temporais
construídas a partir de dados de garantia**

Mayara Mesiano Carrasco

Monografia - MBA em Inteligência Artificial e Big Data

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Mayara Mesiano Carrasco

Predição de taxa de falha utilizando séries temporais construídas a partir de dados de garantia

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Jó Ueyama

Versão original

São Carlos

2022

Mayara Mesiano Carrasco

**Predição de taxa de falha utilizando séries temporais
construídas a partir de dados de garantia**

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Jó Ueyama

Versão original

São Carlos

2022

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

M311p	Carrasco, Mayara Mesiano Predição de taxa de falha utilizando séries temporais construídas a partir de dados de garantia / Mayara Mesiano Carrasco ; orientador Jó Ueyama. – São Carlos, 2022. 55 p. : il. (algumas color.) ; 30 cm. Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022. 1. Séries Temporais. 2. Modelos de Predição. 3. ARIMA. 4. LSTM. 5. Análise Exploratória. I. Ueyama, Jó, orient. II. Especialização em Inteligência Artificial e Big Data.
-------	--

Mayara Mesiano Carrasco

**Failure rate prediction using time series built from
warranty data**

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Jó Ueyama

Original version

São Carlos

2022

*Dedico este trabalho aos meus pais, Cristiane e Edson,
pelo amor, dedicação e incentivo por toda a vida.*

AGRADECIMENTOS

Dedico meus sinceros agradecimentos:

Aos meus pais, Cristiane e Edson, e a minha avó, Valdira, pela presença, pela torcida em todos os momentos e pelo incentivo na minha formação. A Vocês, meus eternos agradecimentos.

Ao professor Jo Ueyama, docente do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo e meu orientador, pela dedicação e disponibilidade em me auxiliar sempre que solicitado.

Ao corpo docente, coordenação e monitores do curso por todo o suporte e atenção aos alunos durante o curso.

À Laís e Marília pela amizade há tantos anos, pelos momentos de descontração e pela escuta ativa sempre que precisei durante todo esse tempo.

RESUMO

Carrasco, M.M. **Predição de taxa de falha utilizando séries temporais construídas a partir de dados de garantia.** 2022. 55p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

O asseguramento da qualidade e a confiabilidade do produto são questões estratégicas relevantes para fabricantes de veículos comerciais. Durante o período de garantia, o fabricante assegura a assistência técnica ao cliente em eventuais falhas ou defeitos funcionais, utilizando essas informações para amadurecer a performance do produto ao longo da sua vida em série. O monitoramento da confiabilidade de um veículo pode ser feito acompanhando a evolução da taxa de falha do produto e atuando nas resolução das falhas de forma a reduzir este indicador ao longo do tempo. Este trabalho se propõe a construir uma série temporal de taxa de falha de um veículo, analisar as características do conjunto de dados e implementar dois modelos de previsão de séries temporais, a saber, *Autoregressive Integrated Moving Average* (ARIMA) e *Long Short-Term Memory* (LSTM). O modelo de ARIMA apresentou um erro percentual absoluto médio de 3,12%, enquanto o modelo LSTM apresentou um erro de 5,05%. Tanto o modelo de ARIMA, quanto o de LSTM apresentaram a métrica de erro bastante equivalentes. Entretanto, os resultados apresentados pelos modelos ARIMA e LSTM apresentam uma característica de tendência contrárias entre si, sendo que o modelo ARIMA está aderente ao valor observado no conjunto de dados real.

Palavras-chave: Séries Temporais. Modelos de Previsão. ARIMA. LSTM. Dados de Garantia.

ABSTRACT

Carrasco, M.M. **Failure rate prediction using time series built from warranty data**. 2022. 55p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

Quality assurance and product reliability are strategic issues relevant to commercial vehicle manufacturers. During the warranty period, the manufacturer ensures the technical assistance to the customer in case of any failures or functional defects, using this information to improve the performance of the product throughout its life in series. Monitoring the reliability of a vehicle can be done by following the evolution of the failure rate of the product and acting on the resolution of failures in order to reduce this indicator over the time. This work proposes to build a time series of a vehicle failure rate, analyze the characteristics of the data set and implement two time series prediction models, namely Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM). The ARIMA model presented an average absolute percentage error of 3.12%, while the LSTM model presented an error of 5.05%. Both the ARIMA model and the LSTM model presented a very equivalent error metric. However, the results presented by the ARIMA and LSTM models present a characteristic of contrary trends, and the ARIMA model is adherent to the value observed in the real data set.

Keywords: Time Series. Forecasting Models. ARIMA. LSTM. Warranty Data.

LISTA DE FIGURAS

Figura 1 – Fluxo da retro-alimentação de falhas de garantia (Fonte: Autora) . . .	25
Figura 2 – Curva teórica de taxa de falha ao longo do tempo (MAISONNIER, 2018)	26
Figura 3 – Modelo de Rede Neural de camada única (HAYKIN, 2008)	29
Figura 4 – Modelo de Rede Neural de múltiplas camadas (HAYKIN, 2008)	30
Figura 5 – Modelo de Rede Neural recorrente (HAYKIN, 2008)	30
Figura 6 – Arquitetura de um RNN LSTM (CORREA, 2021)	31
Figura 7 – Etapas do Projeto de Ciência de Dados (Fonte: Autora)	37
Figura 8 – Etapas da Coleta e Tratamento de Dados (Fonte: Autora)	37
Figura 9 – Etapas da Análise Exploratória (Fonte: Autora)	39
Figura 10 – Etapas da Modelagem (Fonte: Autora)	40
Figura 11 – Etapa da Avaliação de Desempenho (Fonte: Autora)	41
Figura 12 – Série Temporal de Taxa de Falha (Fonte: Autora)	43
Figura 13 – Boxplot (Fonte: Autora)	44
Figura 14 – Histograma do Conjunto de Dados (Fonte: Autora)	44
Figura 15 – Decomposição das Características da Série Temporal (Fonte: Autora) .	46
Figura 16 – Primeira diferenciação do conjunto de treinamento (Fonte: Autora) . .	47
Figura 17 – Segunda diferenciação do conjunto de treinamento (Fonte: Autora) . .	47
Figura 18 – Gráfico de Predição de Série Temporal a partir do Modelo ARIMA (Fonte: Autora)	47
Figura 19 – Arquitetura LSTM (Fonte: Autora)	48
Figura 20 – Gráfico de Predição de Série Temporal a partir do Modelo LSTM (Fonte: Autora)	48
Figura 21 – Comparação dos Dados Preditos versus Dados Observados (Fonte: Autora)	49

LISTA DE TABELAS

Tabela 1 – Comparativo dos Trabalhos Relacionados (Fonte: Autora)	34
Tabela 2 – Tabela de Dados de Produção (Fonte: Autora)	38
Tabela 3 – Tabela de Dados de Falha (Fonte: Autora)	38
Tabela 4 – Tabela de Taxa de Falha (Fonte: Autora)	39
Tabela 5 – Divisão da Série Temporal de Referência (Fonte: Autora)	39

SUMÁRIO

1	INTRODUÇÃO	21
1.1	Motivação	22
1.2	Objetivos de Pesquisa	22
2	FUNDAMENTAÇÃO TEÓRICA	25
2.1	Conceitos relativos à falha de um produto em garantia	25
2.2	Fundamentos de séries temporais	27
2.2.1	Modelos de erro ou de regressão	27
2.2.2	Modelos <i>ARIMA</i>	28
2.2.3	Modelos estruturais	28
2.2.4	Modelos não lineares	28
2.3	Fundamentos de Redes Neurais	28
2.3.1	Rede Neural com Camada Única	29
2.3.2	Rede Neural com Múltiplas Camadas	29
2.3.3	Rede Neural Recorrente	30
2.4	Trabalhos relacionados à análise de dados de vida	31
3	DESENVOLVIMENTO	37
3.1	Coleta e Tratamento dos Dados	37
3.2	Análise Exploratória	39
3.3	Modelagem	40
3.4	Avaliação de Desempenho	41
4	RESULTADOS	43
4.1	Análise Exploratória	43
4.2	Predição de Séries Temporais a partir do Modelo ARIMA	46
4.3	Predição de Séries Temporais a partir do Modelo LSTM	48
4.4	Comparação dos Resultados dos Modelos de Predição	49
5	CONCLUSÃO	51
	REFERÊNCIAS	53

1 INTRODUÇÃO

A garantia sobre um produto é um serviço oferecido pelos fabricantes a fim de assegurar a assistência técnica ao cliente em eventos de falhas prematuras ou defeitos funcionais durante o período que compreende a garantia legal, estabelecida pelo código de defesa do consumidor, e a garantia complementar comercial, oferecida pelo fabricante em contrato (WU, 2012).

Durante este período de garantia, a empresa deve monitorar a confiabilidade do produto no campo e atuar nas falhas recebidas, conforme a sua criticidade e priorização. O conjunto de dados de vida, ou dados de garantia, são constituídos de dois tipos distintos de informação, a saber, os dados de falha e os dados de suspensão (WU, 2013). Os dados de falha são os dados de reclamações reportadas pelos clientes e suas tratativas internas de resolução. Já os dados de suspensão, são dados referentes aos produtos que foram vendidos ao cliente final, mas que ainda não apresentaram defeitos de qualidade no campo.

Utilizando os dados de garantia, é possível estimar a confiabilidade de um produto ao longo de todo o seu ciclo de vida. O propósito de avaliar a confiabilidade de um produto é confrontar seu desempenho em campo com a performance esperada durante a etapa de projeto (KAILASH; PECHT, 2014). Quando o desempenho do produto em campo está abaixo do esperado, ações corretivas são necessárias e provisões financeiras adicionais são empenhadas para adequar o orçamento ao cenário mais desafiador.

A análise de dados de garantia visa extrair informações sobre a performance do produto no campo, viabilizar a priorização dos defeitos e fornecer dados para a análise de causa raiz (HRYCEJ; GRABERT, 2007). Este projeto se propõe a, a partir de uma série histórica de taxa de falha, uma das métricas utilizadas para estimar a confiabilidade de um produto, aplicar técnicas de regressão para, em um primeiro momento, avaliar a aderência do uso da técnica no problema de estimativa de confiabilidade e, em um segundo momento, avaliar o comportamento futuro desta série.

Neste projeto serão aplicadas técnicas de Ciência de Dados e Redes Neurais no desenvolvimento de um modelo de predição de defeitos, com o objetivo de aprimorar a predição de defeitos realizada a partir de distribuições estatísticas, como a distribuição de Weibull. Os dados de falha e dados de suspensão serão estruturados e utilizados como base de dados de entrada em um algoritmo de aprendizado com o objetivo de detectar novos defeitos. Com este trabalho, espera-se desenvolver um modelo confiável e capaz de fornecer informações que serão utilizadas para tomadas de decisão.

1.1 Motivação

O asseguramento da qualidade e a confiabilidade do produto são questões estratégicas relevantes para fabricantes de veículos comerciais (HIPP; LINDNER, 1999). Ao adquirir um equipamento como este, o cliente tem a expectativa de receber um produto com confiabilidade elevada e baixo custo operacional. Além disso, o fabricante é o agente responsável técnico pela segurança e conformidade do produto e, durante o seu ciclo de vida, é responsável por orientar e oferecer suporte ao cliente quanto a necessidade de um plano de manutenções preventivas e corretivas.

Para o fabricante, subestimar as não conformidades do produto pode resultar em perdas financeiras significativas. Ao passo que superestimar essas não conformidades pode reduzir a competitividade do seu produto no mercado. Por outro lado, para o cliente, o custo de paradas não planejadas, como reparos corretivos, afeta diretamente o resultado e eficiência do seu negócio, além do risco de multas e processos jurídicos por parte dos órgãos reguladores (TEO, 2010).

Nesse contexto, a área de qualidade, como responsável por coordenar a análise de causa raiz é a chave do negócio para retroalimentar as áreas de desenvolvimento, produção e toda a cadeia de fornecedores quanto à performance do produto em campo.

Trabalhos recentes sobre qualidade e engenharia de confiabilidade descrevem técnicas baseadas em análise de dados de garantia para retroalimentar o fabricante em problemas como a detecção precoce de não conformidades (LEE *et al.*, 2007a; LEE *et al.*, 2007b), a sugestão de modificações de conceito do produto, além de estimativas e previsões de confiabilidade em termos de quantidade de reclamações e em termos financeiros (WU, 2012).

No contexto do problema de estimativa de confiabilidade, baseado na previsão de séries temporais, técnicas de REDES NEURAIS ARTIFICIAIS (KUTYLOWSKA, 2015), *SUPPORT VECTOR MACHINES* (SAPANKEVYCH; SANKAR, 2009), ou ainda, modelos de redes neurais combinados com a distribuição de Weibull (SINHA, 2017) estão entre as principais técnicas empregadas.

A proposta deste projeto trata da aplicação de técnicas de Análise de Dados e de Redes Neurais para a construção de um sistema de detecção de falhas a partir dos dados de garantia com o objetivo final de estimar a confiabilidade do produto em campo.

1.2 Objetivos de Pesquisa

Neste projeto, espera-se que a partir da base de dados de garantia e de produção de veículos, o algoritmo seja capaz de estimar a confiabilidade do produto em campo. Visando atender às expectativas de um projeto deste porte, são propostos os seguintes objetivos gerais:

- Revisar a literatura recente sobre estimativa de confiabilidade e técnicas de regressão;
- Identificar quais técnicas de ciência de dados e redes neurais são adequadas ao problema de regressão;
- Avaliar técnicas de ciências de dados para prever falhas durante o período de garantia.

Para o problema de estimativa de confiabilidade, são propostos os seguintes objetivos específicos:

- Coletar e manipular dados de produção e de falhas de campo;
- Avaliar as técnicas e algoritmos empregados em problemas de estimativa de confiabilidade a partir de dados de campo;
- Aplicar e avaliar o desempenho dos algoritmos construídos.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão descritos os principais conceitos que embasam as técnicas que serão utilizadas ao longo deste estudo. A fundamentação teórica está dividida em 4 partes, abordando (i) conceitos relativos ao tema do projeto, (ii) revisão de modelos já aplicados ao problema disponíveis na literatura, (iii) breve descrição da modelagem de séries temporais e (iv) conceitos de Redes Neurais Artificiais.

2.1 Conceitos relativos à falha de um produto em garantia

Segundo (WULPI, 2013) a falha pode ser descrita como a incapacidade de um componente funcionar conforme previsto em projeto. As falhas podem ocorrer em qualquer momento do ciclo de vida de um produto, seja durante os testes de engenharia, seja durante a montagem ou ainda, no caso mais crítico, durante o uso pelo consumidor final.

No contexto de uma fabricante de veículos comerciais, esta constrói em toda a extensão do seu mercado de interesse uma rede de concessionárias, que são responsáveis por suportar as atividades de vendas e pós-vendas aos seus clientes. No evento de uma falha em campo, ou seja, no consumidor final, este se apresenta na concessionária para reportar o problema e ser atendido conforme os termos descritos no contrato de prestação de serviço de garantia.

Após o reparo, as peças que apresentaram defeito, seja como causa do problema, seja como consequência de outras peças defeituosas, são encaminhadas para a fábrica de forma a retroalimentá-la com os defeitos para que sejam sanados conforme a criticidade além de munir informações para a análise de risco estratégico.

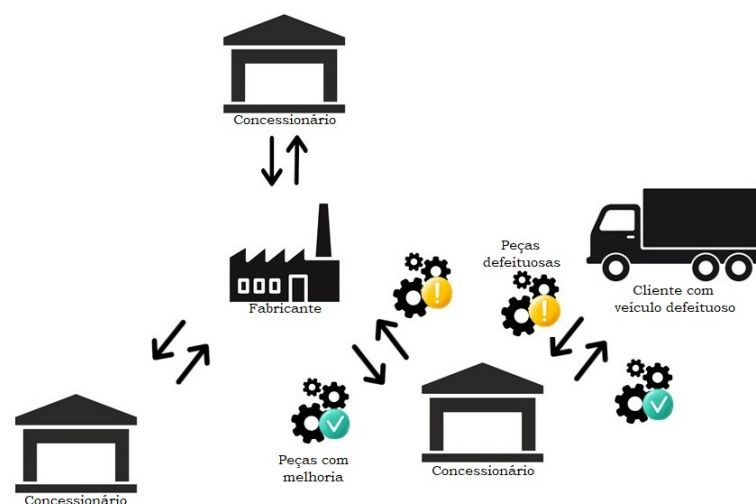


Figura 1 – Fluxo da retro-alimentação de falhas de garantia (Fonte: Autora)

Em termos de dados, todo o fluxo de informação entre a reclamação do cliente e o retorno do fabricante com as melhorias implementadas é documentado nos bancos de dados internos do fabricante.

A Figura 1 apresenta o fluxograma deste processo de atendimento de uma falha de um produto em campo. As setas indicam o fluxo físico da peça ao longo do processo. Com o veículo reparado, o cliente disponibiliza a peça defeituosa ao concessionário, que por sua vez, envia a peça para o fabricante. O fabricante avalia o defeito e, conforme criticidade, retorna ao cliente a peça com as devidas melhorias implementadas, de acordo com as políticas de garantia contratuais.

A tabela que descreve os dados de garantia, ou dados de vida, possui atributos do veículo reclamado, como o número de chassi e a composição dos grandes sistemas do veículo, atributos da falha como o modo de falha reclamado, o texto da reclamação descrita pelo concessionário e o texto do técnico da fábrica que analisou o processo e quilometragem da falha, além de métricas como custo total do processo incluindo custo de material/peça e custo de mão de obra do serviço prestado.

Os modos de falha apresentados pelo veículo são modelados a partir da análise de dados de vida (LI; MOBIN, 2018).

A taxa de falha ou defeito por unidade produzida (DPU) representa a ocorrência de não conformidades em uma amostra, descrita em um intervalo de tempo de produção. Ela não pode ser considerada uma probabilidade, pois pode exceder o limite máximo de 1.

$$DPU = \frac{\text{quantidade de reclamações}}{\text{quantidade de veículos produzidos}} \quad (2.1)$$

(MAISONNIER, 2018), (THOMAS, 2006) e (STOLFI, 2017) apresentam a “curva da banheira” na Figura 2. Trata-se de uma representação teórica taxa de falha ao longo do tempo, relacionando as causas das falhas em cada momento-chave.

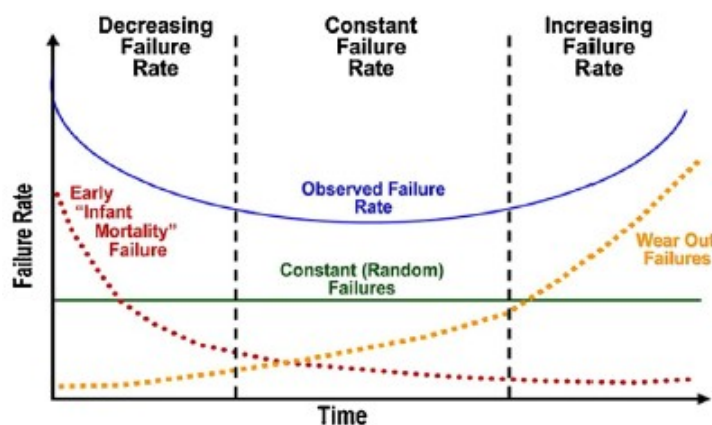


Figura 2 – Curva teórica de taxa de falha ao longo do tempo (MAISONNIER, 2018)

Falhas prematuras, ou falhas de mortalidade infantil, são falhas que ocorrem no início da vida do produto em campo. Em geral, são causadas por falhas no processo de fabricação do produto ou da matéria-prima utilizada além do controle de qualidade inadequado.

Falhas aleatórias, ou falhas na operação normal, são falhas que ocorrem durante boa parte do ciclo de vida, em decorrência de limitações de projeto, defeitos latentes nos componentes e agentes externos.

Finalmente a falhas de desgaste são falhas que ocorrem ao fim da vida útil do produto, sendo relacionadas a fissura, fadiga, rupturas mecânicas, além de contaminação, oxidação e corrosão do material.

2.2 Fundamentos de séries temporais

Para este estudo, interessam as ocorrências mensais de reclamações de garantia em veículos comerciais em função da quantidade de veículos produzidos neste determinado mês, sendo que o objetivo da análise de série temporal será prever a taxa de falha futura. Essa tendência no comportamento da taxa de falha será utilizada para avaliar destinação de recursos para aprimorar a confiabilidade do produto. (MORETTIN; TOLOI, 2006) descreve uma série temporal como um conjunto de observações ordenadas no tempo. (EHLERS, 2007) decompõe uma série temporal X_t na componente de tendência T_t , na componente cíclica ou sazonal C_t e uma componente aleatória ou ruído R_t

$$X_t = T_t + C_t + R_t \quad (2.2)$$

Segundo (MORETTIN; TOLOI, 2006), os modelos que descrevem séries temporais estão relacionados a processos estocásticos, isto é, cujo processo é controlado por leis probabilísticas.

A análise e modelagem dinâmica de séries temporais é um campo bastante explorado na literatura com aplicações não somente restrita ao domínio das ciências da computação, mas também à áreas de negócios, economia, saúde e indústria. Dada a relevância do tema, diferentes técnicas foram desenvolvidas com o objetivo de melhorar a acurácia dos modelos e ampliar a aplicabilidade das técnicas a diferentes características de séries temporais (SIAMI-NAMINI; TAVAKOLI; Siami Namin, 2019).

2.2.1 Modelos de erro ou de regressão

Neste modelo, a função de tempo $f(t)$ tem um fator de natureza determinística e não correlacionada à sequência aleatória a_t . (MORETTIN; TOLOI, 2006) avalia que o modelo de erro é clássico para análise de séries temporais em economia, onde:

$$f(t) = T_t(t) + S_t(t) + a_t(t) \quad (2.3)$$

Dessa forma, $T_t(t)$ é um polinômio que representa a componente sistemática, ou determinística. A componente $S_t(t)$ é um polinômio harmônio representando as flutuações cíclicas e as variações sazonais.

2.2.2 Modelos *ARIMA*

O Modelo de Média Móvel Integrada e Autoregressiva, da sigla em inglês *ARIMA*, é um modelo que combina processos autoregressivos e de médias móveis para construir uma série temporal (MORETTIN; TOLOI, 2006). Proposto a partir do método de George Box e Gwilym Jenkins em 1970, o modelo é definido com 3 parâmetros-chave, a saber:

- AR (Autoregressão) - avalia a ordem de defasagem do conjunto de dados ao longo do tempo;
- I (Integrada) - indica o processo de diferenciação dos valores observados para alcançar a série estacionária;
- MA (Média Móvel) - avalia a dependência entre os valores observados e o erro residual da média móvel frente a defasagem dos valores observados.

2.2.3 Modelos estruturais

Os modelos estruturais, ou modelos de espaço de estados, permitem quantificar o componente sistemático em função do modelo de erro. Em geral, são divididos em modelo de nível local ou no modelo de extendido de tendência linear local com uma inclinação estocástica representando um comportamento sazonal.

2.2.4 Modelos não lineares

Segundo (MORETTIN; TOLOI, 2006), existem situações em que o comportamento que se deseja medir não é aderente ao modelo linear. Os comportamentos incluem mudanças repentinas, alteração da variância com o tempo e irreversibilidade no tempo.

2.3 Fundamentos de Redes Neurais

Segundo (HAYKIN, 2008), uma Rede Neural é um sistema de processamento distribuído com o objetivo de armazenar conhecimento e disponibilizar esse conhecimento para uso. Matematicamente, uma Rede Neural é um grafo orientado por nós e conectado por interligações sinápticas, que possui 4 propriedades:

- Cada neurônio é composto por um conjunto de elos sinápticos lineares, um bias externo e uma energia de ativação de caráter não linear;
- Os elos sinápticos de um neurônio ponderam seus respectivos sinais de entrada;
- A soma ponderada dos sinais de entrada define o campo local induzido do neurônio;
- O elo de ativação limita o campo local induzido do neurônio para produzir uma saída.

2.3.1 Rede Neural com Camada Única

Uma Rede Neural de camada única é composta por uma camada de entrada de neurônios, que não é contabilizada como camada pela inexistência de operações computacionais nesta região, e uma camada de saída de neurônios (HAYKIN, 2008). A Figura 3 apresenta um esquema desta classe de Rede Neural.

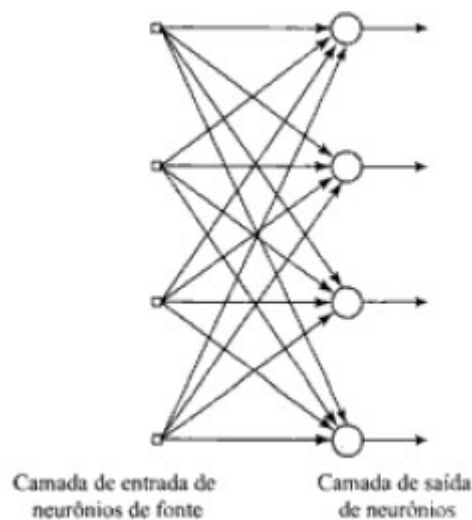


Figura 3 – Modelo de Rede Neural de camada única (HAYKIN, 2008)

2.3.2 Rede Neural com Múltiplas Camadas

A Rede Neural com múltiplas camadas possui pelo menos uma camada de neurônios ocultos, cuja função é intervir entre a entrada e a saída de neurônios de forma útil (HAYKIN, 2008). A Figura 4 apresenta um esquema desta classe de Rede Neural.

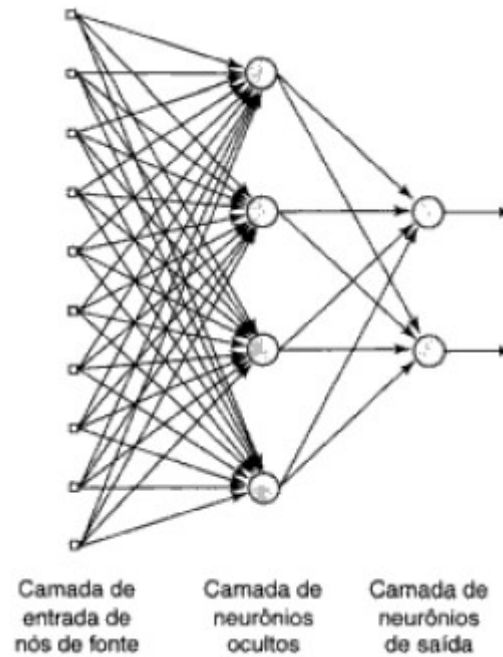


Figura 4 – Modelo de Rede Neural de múltiplas camadas (HAYKIN, 2008)

2.3.3 Rede Neural Recorrente

Rede Neural Recorrente difere das Redes Neurais de camada única ou múltipla em decorrência de ter pelo menos um laço de retroalimentação. O laço de realimentação favorece a capacidade de aprendizagem da rede, permitindo comportamentos não lineares (HAYKIN, 2008) e (BRAGA; CARVALHO; LUDERMIR, 2007). A Figura 5 apresenta um esquema desta classe de Rede Neural.

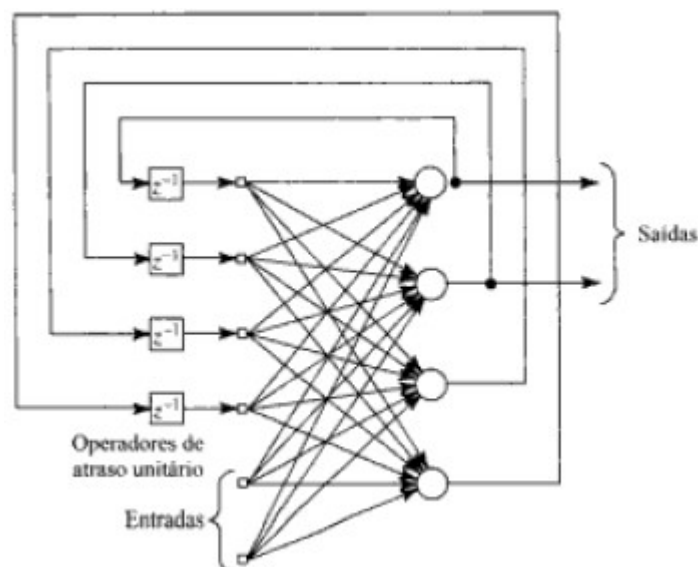


Figura 5 – Modelo de Rede Neural recorrente (HAYKIN, 2008)

O modelo de Memória Longa de Curto Prazo, na sigla em inglês LSTM, é um tipo de

Rede Neural Recorrente com a capacidade de memorizar uma sequência de dados proposto por Hochreiter e Schmidhuber em 1997. Cada LSTM é uma célula de memória contendo uma porta de entrada e uma de saída de dados, além de uma porta de esquecimento. (CORREA, 2021) apresenta na Figura 6 a arquitetura de uma LSTM, afirmando que essas portas são geralmente compostas por uma função de ativação sigmoide, que gera um valor entre 0 e 1, descrevendo o nível de ativação do porta, onde 0 está completamente fechado e 1 está completamente aberto.

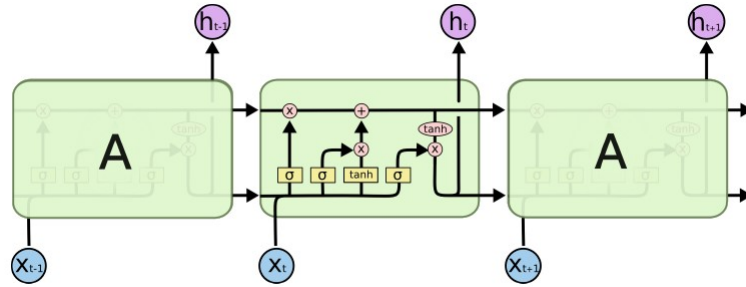


Figura 6 – Arquitetura de um RNN LSTM (CORREA, 2021)

A configuração do modelo LSTM utiliza os seguintes parâmetros (TAVARES; QUADRELLI, 2022):

- input: o número de entradas anteriores a serem usadas como entrada para o modelo;
- nodes: o número de nós a serem usados na camada oculta;
- epochs: o número de períodos de treinamento;
- batch: o número de amostras a serem incluídas em cada mini batch.

O trabalho de (BROWNLEE, 2018) afirma que o modelo LSTM mapeia um conjunto de dados observados na porta de entrada para avaliar uma estimativa na porta de saída. Esse conjunto de dados é transformado em várias amostras utilizadas no aprendizado supervisionado do modelo.

2.4 Trabalhos relacionados à análise de dados de vida

A análise de dados de vida com foco na predição de falha em garantia satisfaz diversas demandas estratégicas das empresas. (ELERATH; PECHT, 2012) discute que as predições de confiabilidade são utilizadas entre outros motivos para atender necessidades regulatórias de segurança, ajustar reservas de recursos para atendimento de produção em série e de campo e ainda aprimorar alternativas conceituais do projeto de um produto.

Dessa forma, diferentes ramos de negócios buscam utilizar técnicas de Machine Learning para melhorar a predição de confiabilidade de seus produtos.

O estudo de (TORGUNOV *et al.*, 2020) apresenta um método para predição de reparos em garantia utilizando os dados históricos de reparo em garantia de veículos e informações de telemetria diagnóstica. As diferentes técnicas de Machine Learning utilizadas, a saber, *Support Vector Machines*, *Random Forests* e Árvore de Decisão, em um problema de classificação binária, demonstraram performance equivalente embora a Árvore de Decisão apresente um resultado mais consistente. O estudo demonstra que o problema de classificação binária pode ser uma modelagem adequada para a predição de falhas futuras, embora a acurácia seja limitada em função da disponibilidade do conjunto de dados. Como trabalhos futuros, serão aprofundados os estudos sobre estas técnicas de forma a delimitar o impacto do tamanho do conjunto de dados na acurácia destes modelos.

Ainda na indústria automotiva, (KHOSHKANGINI *et al.*, 2020) estudou modelos de predição de defeitos de qualidade baseado em Machine Learning, combinando informações de aplicação dos veículos e histórico de falhas anteriores em problemas de classificação e de regressão. Em ambos os problemas, os modelos propostos foram capazes de prever os defeitos atendendo aos critérios estabelecidos. Entretanto, o modelo de regressão se mostrou mais aderente a predição de falhas em veículos com menos de 10 meses de uso. Já o modelo de classificação apresentou melhor performance com um histórico maior de dados de campo. O estudo aponta ainda que os modelos de predição estão restritos a influência e dependência dos componentes do produto, sendo que o mapeamento dessa rede de dependências será o objeto de estudo em trabalhos futuros.

Se por um lado, (KHOSHKANGINI *et al.*, 2020) propôs um modelo alimentado por dados de aplicação em campo, por outro lado, (SABRI-LAGHAIE *et al.*, 2021), estudou um modelo alimentado com parâmetros de produção para detectar antecipadamente defeitos na produção de motores veiculares. O modelo de Redes Neurais proposto foi capaz de prever o nível de confiabilidade dos motores. Em trabalhos futuros, um algoritmo de clusterização será incorporado ao modelo com o objetivo de identificar o modo de falha que o produto pode apresentar.

No ramo de geração de energia hidrelétrica, (ASSIS *et al.*, 2020) propôs uma estimativa de parâmetros de confiabilidade utilizando um modelo baseado em Redes Neurais Artificiais combinado com distribuições estatísticas compatíveis com a análise de problemas complexos, como distribuição de Weibull e exponencial com uma função de predição. Os resultados compararam o modelo baseado em Redes Neurais com as distribuições estatísticas comumente utilizadas neste tipo de análise. A predição proposta pelo modelo apresentou um resultado superior ao modelo estatístico tradicional, superando em 89% as análises nos critérios de coeficiente de determinação (R^2) e erro quadrático médio (MSE). O autor ressalta ainda que o modelo baseado em Redes Neurais Artificiais

requer tratativas adicionais para se adequar ao modelo probabilístico.

Para aplicações em usinas eólicas, (KADHEM *et al.*, 2017) propõe um modelo de predição da velocidade do vento em séries temporais baseado na combinação da distribuição de Weibull com Redes Neurais Artificiais. A dependência entre a velocidade do vento e a sazonalidade climática alimenta este modelo híbrido, sendo capaz de identificar flutuações na velocidade dos ventos ao longo ano e em diferentes localidades. Além disso, o erro do modelo de predição foi menor do que o erro gerado pelo modelo que utiliza puramente a distribuição de Weibull. Futuramente, este modelo será empregado na análise de confiabilidade de unidades de potência em geradores eólicos.

Os dados operacionais de uma concessionária polonesa de fornecimento de água foram utilizados para prever a taxa de falha do sistema. (KUTYLOWSKA, 2015) propôs em seu trabalho um modelo de Redes Neurais Artificiais treinado por Quase-Newton que apresentou, do ponto de vista de engenharia, um resultado aceitável para avaliação do nível de falhas no sistema, de forma a suportar estratégias de renovação e manutenção do sistema. O autor destaca que o modelo alimentado com informações adicionais como a distribuição da rede de tubulações, a temperatura local e o tipo de solo deve apresentar resultados superiores aos atuais.

Na área de mineração, por sua vez, (SINHA, 2017) propôs um modelo utilizando Redes Neurais Artificiais para avaliar a taxa de falha de equipamentos trituradores. Os resultados mostraram que este modelo apresenta valores de R^2 tão elevados quanto um modelo de simulação de Monte Carlo, evidenciando que os resultados são fortemente correlacionados.

(STAMENKOVIC, 2015) propõe em seu estudo um modelo de Redes Neurais Artificiais com simulação de Monte Carlo para prever a confiabilidade de lâmpadas utilizando dados de testes em diferentes condições de operação. Os resultados mostraram que o modelo proposto possui maior acurácia para esta predição do que modelos de regressão linear comumente utilizados na literatura para atividades similares. O estudo ressalta ainda que trabalhos futuros podem avaliar a aplicabilidade do modelo para problemas de análise de custos de garantia e adaptado para outros produtos que não sejam modelados por uma distribuição de Weibull.

A versatilidade do modelo híbrido de Redes Neurais Artificiais alimentado por parâmetros da distribuição de Weibull foi reforçada no estudo de (YANG *et al.*, 2015) para avaliação do nível de confiabilidade de máquinas de usinagem CNC. Os resultados mostraram que o modelo de Redes Neurais com Weibull apresenta desempenho superior ao modelo tradicional de Weibull biparamétrico.

Finalmente, (LEE *et al.*, 2021) propôs um comparativo de 3 modelos de aprendizagem profunda - Rede Neural 1D Convolutiva, Rede Neural Recorrente e Rede Neural

Sequence to Sequence para avaliar a predição de confiabilidade de peças automotivas baseada em dados de vida do produto. Os modelos foram aplicados em séries temporais segundo o modelo de *ARIMA*, modelos de *Support Vector Machine* e *Random Forest*. Dentre todos os modelos testados, o modelo de Rede Neural Recorrente apresentou performance superior na predição de falhas e de confiabilidade. Em trabalhos futuros, os modelos serão alimentados com informações das peças de forma a replicar os recursos em peças com padrões semelhantes.

A Tabela 1 dispõe de um comparativo dos trabalhos discutidos anteriormente e sua aplicabilidade ao problema proposto neste estudo de acordo com 5 critérios:

Estudos	Uso de Machine Learning	Avaliação do Modelo	Uso em Série Temporal	Uso de Dados de Vida	Aplicação em Veículos
(TORGUNOV <i>et al.</i> , 2020)	sim	sim	não	sim	sim
(ASSIS <i>et al.</i> , 2020)	sim	sim	não	sim	não
(KHOSHKANGINI <i>et al.</i> , 2020)	sim	sim	sim	sim	sim
(SABRI-LAGHAIE <i>et al.</i> , 2021)	sim	sim	não	sim	sim
(KADHEM <i>et al.</i> , 2017)	sim	sim	não	não	não
(KUTYLOWSKA, 2015)	sim	sim	sim	sim	não
(SINHA, 2017)	sim	sim	sim	sim	não
(STAMENKOVIC, 2015)	sim	sim	sim	sim	não
(YANG <i>et al.</i> , 2015)	sim	sim	não	sim	não
(LEE <i>et al.</i> , 2021)	sim	sim	sim	sim	sim
Trabalho Proposto	sim	sim	sim	sim	sim

Tabela 1 – Comparativo dos Trabalhos Relacionados (Fonte: Autora)

Os trabalhos relacionados validaram a complexidade embutida no estudo de predição de confiabilidade em função das diferentes abordagens possíveis para alimentar o modelo preditivo e reforçou a constante busca de diferentes ramos de negócios por aplicar técnicas de *Machine Learning* para prever a confiabilidade de seus ativos. Considerando a relevância deste problema, este estudo visa aplicar modelos de Redes Neurais Artificiais para prever a tendência da taxa de falha para veículos em campo a partir de dados de vida dos veículos. Para comparação com os trabalhos relacionados, o estudo proposto por este projeto foi avaliado pelos mesmos critérios na Tabela 1.

Dentro da literatura relacionada à dados de vida acima citada, o artigo de (LEE *et al.*, 2021) é o que mais se assemelha à proposta deste trabalho, sendo avaliadas diversas técnicas de aprendizado de máquinas e aprendizado profundo e comparando com métodos paramétricos. Entretanto, a medida de taxa de falha utilizada é diferente da proposta neste trabalho, considerando o volume absoluto de falhas entre dois períodos de tempo

subsequentes, ao passo que neste trabalho é considerado o volume relativo de falhas sempre a partir da data de produção do veículo, delimitando apenas o limite superior do tempo de medição. Com o ganho de escala obtido ao replicar este mesmo estudo em diferentes limites superiores do tempo de medição, espera-se visualizar a evolução da maturidade do produto. Adicionalmente, o trabalho de (KHOSHKANGINI *et al.*, 2020) também apresenta diversas similaridades com este trabalho proposto, utilizando conjuntos de dados de histórico de garantia e de telemetria veicular para detectar falhas em garantia. O que difere este trabalho do proposto é a técnica de predição utilizada na modelagem, sendo implementada a técnica de *Gradient Boosting* ao passo que este trabalho discutirá a modelagem de ARIMA e LSTM.

3 DESENVOLVIMENTO

O escopo deste trabalho trata da predição de séries temporais a partir de uma base de dados de produção e de falhas recebidas em garantia por um fabricante de veículo comerciais. Neste capítulo será apresentada a metodologia do estudo em desenvolvimento.

A Figura 7 descreve as etapas do estudo conforme o escopo do projeto.

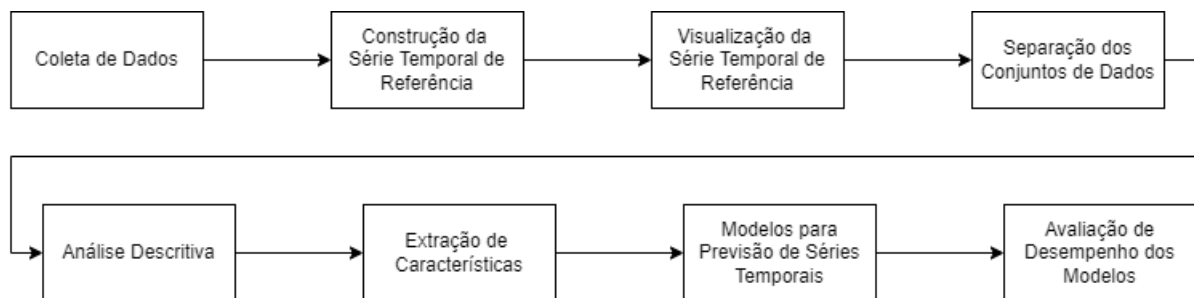


Figura 7 – Etapas do Projeto de Ciência de Dados (Fonte: Autora)

3.1 Coleta e Tratamento dos Dados

A etapa de coleta e tratamento dos dados compreende as 3 etapas em destaque na Figura 8.

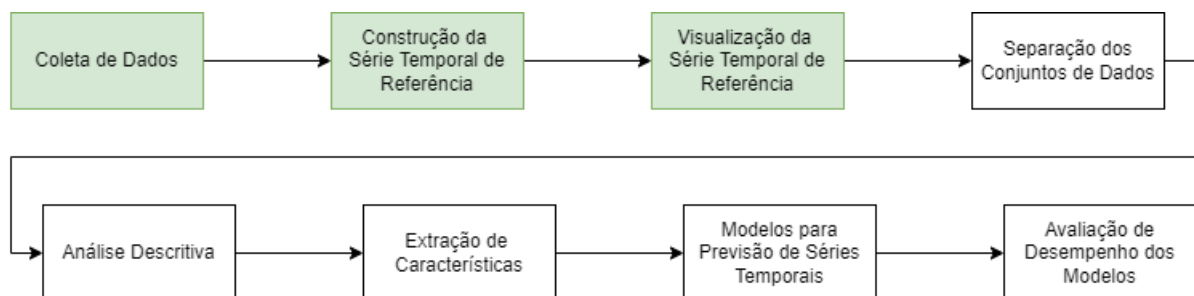


Figura 8 – Etapas da Coleta e Tratamento de Dados (Fonte: Autora)

Os dados utilizados para a construção da série temporal foram obtidos a partir de bases de dados hipotéticas contendo dados de produção e de falhas de veículos comerciais desde janeiro de 2012. A base de dados de produção possui os atributos de chassi, modelo de veículo e data de produção, sendo que cada registro da tabela indica um veículo produzido, conforme Tabela 2. Já a base de dados de falha, possui os atributos de chassi, modelo de veículo, sistema defeituoso, data de produção, data de reparo e número da

ordem de serviço, sendo que cada registro da tabela indica um reparo realizado, conforme Tabela 3.

Chassis	Veículo	Data Produção
VIN-1	Veículo A	Data
VIN-2	Veículo A	Data
VIN-3	Veículo B	Data

Tabela 2 – Tabela de Dados de Produção (Fonte: Autora)

Chassis	Veículo	Sistema Defeituoso	Data Produção	Data Reparo	Ordem Serviço
VIN-1	Veículo A	Sistema X	Data Prod	Data Rep	OS-1
VIN-2	Veículo B	Sistema Y	Data Prod	Data Rep	OS-2
VIN-3	Veículo C	Sistema Z	Data Prod	Data Rep	OS-3

Tabela 3 – Tabela de Dados de Falha (Fonte: Autora)

As duas bases de dados serão processadas em um software de gerenciamento de banco de dados, de forma a anonimizar o nome modelo do veículo e o nome do sistema defeituoso. Além disso, ambas as bases de dados serão apresentadas em novas tabelas com amostragem mensal. Finalmente, para a base de dados de falha, será calculada o tempo de operação até a falha em meses (MIS), a partir da seguinte fórmula:

$$\text{Tempo de Operação (MIS)} = \frac{\text{Data de Reparo} - \text{Data de Produção}}{30} \quad (3.1)$$

Com o relacionamento entre as duas bases de dados estabelecidos pela chave primária composta (Modelo de Veículo e Mês de Produção), será criada uma nova tabela que apresentará a taxa de falha referente as falhas que ocorreram em até 12 meses da data de produção. Esse intervalo-chave é característicos para representar a política de garantia em vigor. A taxa de falha será calculada conforme a equação abaixo. Dessa forma, a série temporal será apresentada conforme a Tabela 4.

$$\text{Taxa de Falha} = \frac{\text{Quantidade de Falhas}}{\text{Quantidade de Veículos Produzidos}} \quad (3.2)$$

Mês de Produção	Taxa de Falha
Data Prod	Taxa i
Data Prod	Taxa j
Data Prod	Taxa k

Tabela 4 – Tabela de Taxa de Falha (Fonte: Autora)

Na presença de dados ausentes, ou seja, meses sem volume de produção ou sem registro de falha, a taxa de falha que será registrada na mês de referência será 0%.

3.2 Análise Exploratória

A etapa de análise exploratória compreende as 3 etapas em destaque na Figura 9.

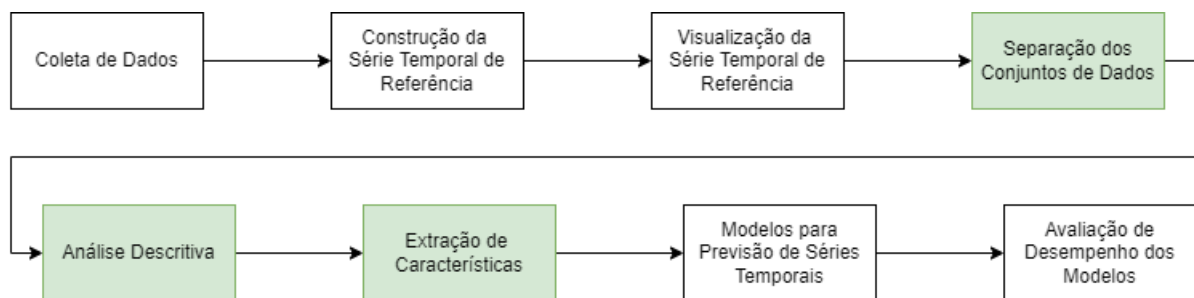


Figura 9 – Etapas da Análise Exploratória (Fonte: Autora)

Antes de iniciar a análise exploratória dos dados, a série temporal será dividida em duas partes, de forma a constituir uma base de dados de treinamento e uma base de teste/dados observados. A base de dados observados será composta pelos últimos 24 meses de produção. A divisão da base de dados pode ser verificada na tabela 5.

Tipo da Série	Período de Tempo
Série de Treinamento	de Jan/2012 a Jun/2020
Série de Teste	de Jul/2020 a Jun/2022

Tabela 5 – Divisão da Série Temporal de Referência (Fonte: Autora)

A análise exploratória será feita a partir da série temporal de treinamento. Sobre essa série serão avaliadas as métricas básicas de estatística descritiva (média, quartis,

variância, desvio padrão, obliquidade e curtose), além de avaliar a distribuição de falhas por sistema e por ano de produção.

Para concluir a análise exploratória, serão extraídas as características da série temporal a partir de testes estatísticos. Será considerado o intervalo de confiança de 95% para todos os testes realizados neste estudo. Para a extração de características, os seguintes testes serão realizados.

- Normalidade: Teste de Kolmogorov-Smirnov
- Sazonalidade: Teste de Kruskal-Wallis
- Tendência: Teste de Mann-Kendall
- Estacionariedade: Teste de Dickey-Fuller

3.3 Modelagem

A etapa de coleta e tratamento dos dados compreende a etapa em destaque na Figura 10.

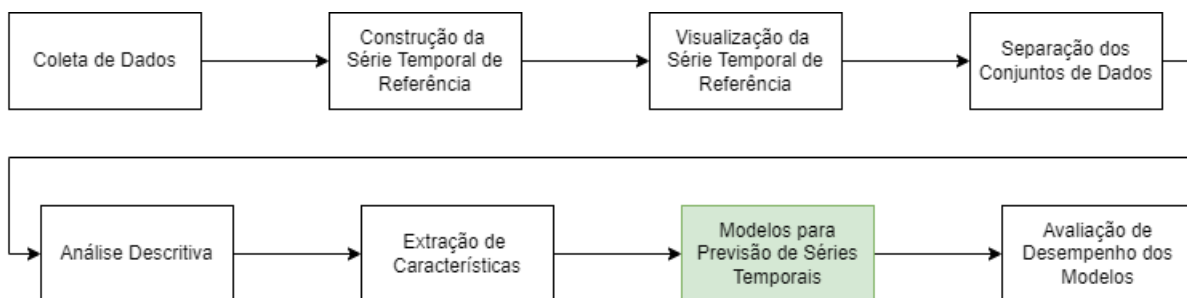


Figura 10 – Etapas da Modelagem (Fonte: Autora)

Para comparação do desempenho da predição da série temporal, serão construídos 2 modelos preditivos:

- Modelo ARIMA
- Modelo LSTM (RNN)

Os parâmetros de cada modelo, bem como a arquitetura de cada modelo serão discutidos na seção de resultados.

Outras técnicas envolvendo redes neurais artificiais convolucionais e um modelo híbrido envolvendo redes neurais artificiais e a distribuição de Weibull foram estudados para inclusão no escopo deste trabalho, mas não foram implementados em função da

execução do cronograma proposto na etapa de projeto. Essas e outras técnicas bastante utilizadas na literatura podem integrar um trabalho futuro envolvendo predição de séries temporais.

Para todos os modelos, é esperado como saída do processo de previsão uma tabela contendo os valores estimados de taxa de falha por mês para os últimos 24 meses de produção. As séries estimadas também serão apresentadas de forma gráfica, em conjunto à série temporal de referência.

O algoritmo desenvolvido para este trabalho foi desenvolvido na plataforma Jupyter Notebook, em linguagem Python. Os experimentos foram conduzidos em computador de configuração Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz - 2.90GHz, memória RAM 8GB e Windows 10 64 bits. O algoritmo e a base de dados hipotética utilizada estão armazenados no repositório do GitHub disponível no link: <https://github.com/mayaracarrasco/Time-Series-Forecasting>.

3.4 Avaliação de Desempenho

A etapa de coleta e tratamento dos dados compreende a etapa em destaque na Figura 11.

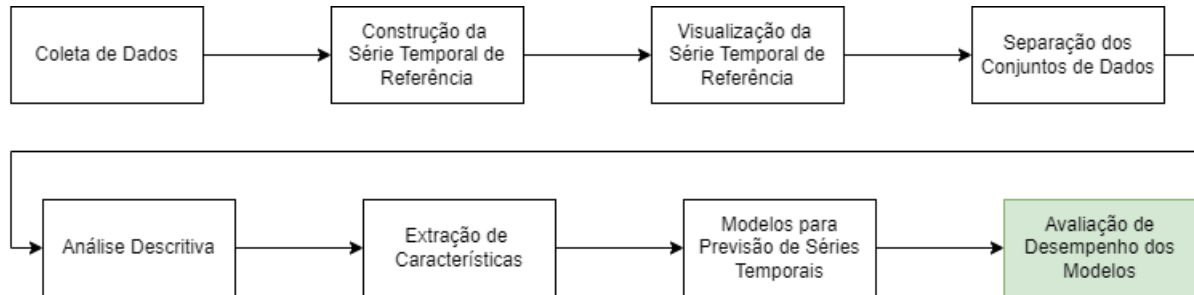


Figura 11 – Etapa da Avaliação de Desempenho (Fonte: Autora)

Para avaliar o desempenho dos modelos, serão utilizadas métricas que calcularão o erro da estimativa gerada pelos modelos, comparando a estimativa F_t com o dados observados Y_t na série temporal de teste.

Erro Percentual Absoluto Médio (MAPE)

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{F_t - Y_t}{Y_t} \right| \quad (3.3)$$

Raiz Quadrada do Erro Médio (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (F_t - Y_t)^2} \quad (3.4)$$

Sendo que o modelo mais adequado para a estimativa será o que apresentar as menores métricas de erro.

Além de avaliar a performance dos modelos, também será avaliada a aderência do conjunto de dados predito frente ao conjunto de dados observados a partir do coeficiente de correlação de Pearson (r), considerando $\bar{F}$ e $\bar{Y}$ as médias aritméticas dos dados preditos e observados, respectivamente.

$$r = \frac{\sum_{t=1}^n (F_t - \bar{F}) (Y_t - \bar{Y})}{\sqrt{\left[\sum_{t=1}^n (F_t - \bar{F})^2 \right] \left[\sum_{t=1}^n (Y_t - \bar{Y})^2 \right]}} \quad (3.5)$$

Em função do valor de r , podemos avaliar a qualidade da correlação entre os conjuntos de dados:

- $r > 0,9$ positivo ou negativo indica uma correlação muito forte
- $0,7 < r < 0,9$ positivo ou negativo indica uma correlação forte
- $0,5 < r < 0,7$ positivo ou negativo indica uma correlação moderada
- $0,3 < r < 0,5$ positivo ou negativo indica uma correlação fraca
- $0 < r < 0,3$ positivo ou negativo indica uma correlação desprezível

4 RESULTADOS

Neste capítulo são apresentados e discutidos os resultados da análise e previsão de séries temporais mensais de taxa de falha de um produto em garantia. A análise foi executada conforme o planejamento descrito no capítulo anterior e foram implementados os modelos ARIMA e LSTM para a etapa de previsão da série temporal.

4.1 Análise Exploratória

A série temporal foi construída a partir dos dados de produção e o histórico de falhas em garantia, conforme descrito na metodologia. A Figura 12 apresenta a série temporal compreendendo o período de Jan/2012 a Jun/2022, com granularidade mensal.

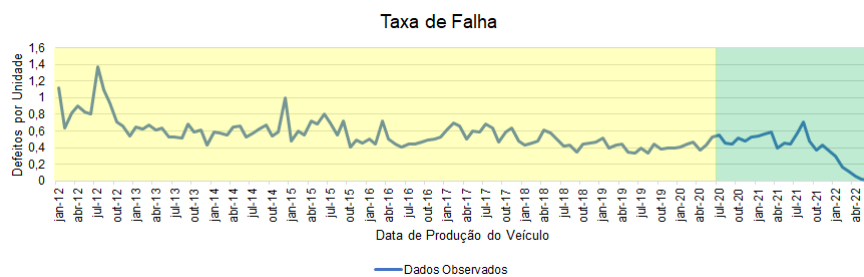


Figura 12 – Série Temporal de Taxa de Falha (Fonte: Autora)

Para viabilizar os dados de treinamento e teste, a série temporal foi dividida em 2 seções, indicadas na Figura 12. Na região amarela está delimitada a base de treinamento e na região verde está delimitada a base de teste.

Em ambas as bases, não havia presença de dados nulos ou faltantes que necessitassem de pré-tratamento. A taxa de falha, medida em defeitos por unidade, tem na série construída os seguintes parâmetros descritivos:

- Valor Médio: 0,57
- Valor Mínimo: 0,34
- Primeiro Quartil (Quartil Inferior) Q_1 : 0,44
- Segundo Quartil (Mediana) Q_2 : 0,54
- Terceiro Quartil (Quartil Superior) Q_3 : 0,64
- Valor Máximo: 1,37

A Figura 13 apresenta o boxplot da distribuição de dados, mostrando as métricas de posição calculadas acima. É possível identificar a presença de *outliers* acima do limite superior do processo, indicando anormalidades eventuais no histórico da distribuição. Por ora, neste estudo, os *outliers* serão mantidos no conjunto de dados.

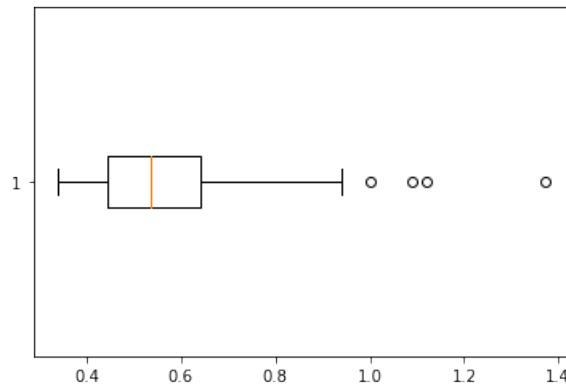


Figura 13 – Boxplot (Fonte: Autora)

A dispersão do conjunto de dados foi medido por meio das métricas de momento central, descritas abaixo:

- Variância: 0,03
- Desvio Padrão: 0,17
- Obliquidade: 1,78 (distribuição assimétrica à esquerda)
- Kurtosis: 4,85 (distribuição mais alongada do que a gaussiana)

A Figura 14 apresenta o histograma da distribuição de dados, mostrando as características observadas anteriormente.

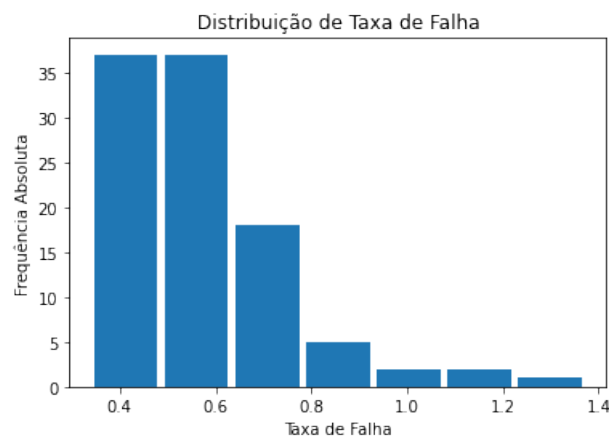


Figura 14 – Histograma do Conjunto de Dados (Fonte: Autora)

Afim de identificar as características da série temporal de estudo, foram propostos os seguintes testes estatísticos.

Teste de Normalidade de Kolmogorov-Smirnov

- H_0 (Hipótese Nula): A distribuição é normal
- H_1 (Hipótese Alternativa): A distribuição não é normal

Para um nível de significância estatística de 5%, o valor-p resultante do teste foi de $6,25 \times 10^{-11}$. Portanto, a hipótese nula foi rejeitada e os dados não são normais.

Teste de Sazonalidade de Kruskal-Wallis

- H_0 (Hipótese Nula): A mediana em todos os agrupamentos é a mesma, ou seja, conjunto apresenta sazonalidade
- H_1 (Hipótese Alternativa): A mediana dos agrupamentos não é igual, ou seja, conjunto não apresenta características de sazonalidade

Para um nível de significância estatística de 5%, o valor-p resultante do teste foi de 0,0024. Portanto, a hipótese nula foi rejeitada e os dados não apresentam sazonalidade em nenhum agrupamento de dados.

Teste de Tendência de Mann-Kendall

- H_0 (Hipótese Nula): O conjunto de dados não apresenta tendência
- H_1 (Hipótese Alternativa): O conjunto de dados apresenta alguma tendência

Para um nível de significância estatística de 5%, o valor-p resultante do teste foi de $2,99 \times 10^{-14}$. Portanto, a hipótese nula foi rejeitada e os dados apresentam uma tendência, no caso caracterizada como decrescente.

Teste de Estacionariedade de Dickey-Fuller

- H_0 (Hipótese Nula): A série de dados não é estacionária
- H_1 (Hipótese Alternativa): A série de dados é estacionária

Para um nível de significância estatística de 5%, o valor-p resultante do teste foi de 0,025. Portanto, a hipótese nula foi rejeitada e os dados apresentam a característica de estacionaridade.

Em resumo, conjunto de dados que é objeto de estudo desta análise apresenta como principais características uma distribuição de dados não normal, sem sazonalidade, com tendência decrescente e estacionária.

Dentro das ferramentas disponíveis para análise de séries temporais na linguagem Python, existe uma funcionalidade que decompõe as características avaliadas acima em um conjunto de dados de série temporal. A Figura 15 apresenta a decomposição para o conjunto de dados de estudo. Visualmente é possível identificar a tendência de decrescimento da curva de taxa de falha, a ausência de sazonalidade e, adicionalmente ao estudo, a presença de um resíduo com características normais.

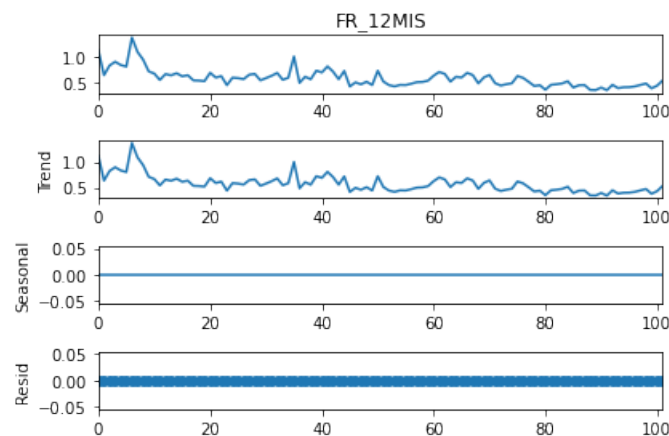


Figura 15 – Decomposição das Características da Série Temporal (Fonte: Autora)

4.2 Predição de Séries Temporais a partir do Modelo ARIMA

O modelo de ARIMA foi construído com base nas características de autoregressão, diferenciação e média móvel avaliados para o conjunto de dados. Dado que durante a análise exploratória foi identificado que o conjunto de dados de treinamento é uma série estacionária, o parâmetro de autoregressão d será nulo.

Para avaliar a diferenciação, o conjunto de dados foi analisado quanto a autocorrelação parcial para a primeira (Figura 16 e segunda diferenciação 17. Comparando as duas análises, foi definido que o parâmetro de integração será 1.

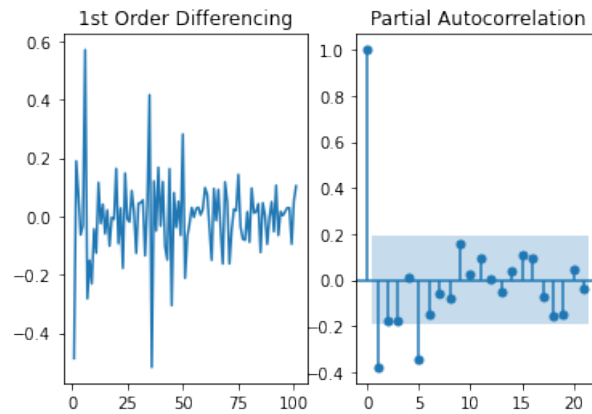


Figura 16 – Primeira diferenciação do conjunto de treinamento (Fonte: Autora)

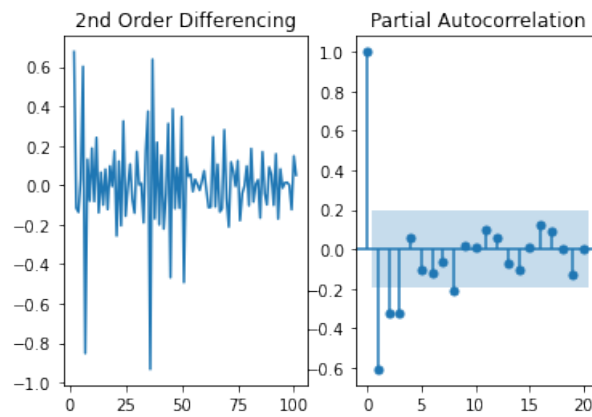


Figura 17 – Segunda diferenciação do conjunto de treinamento (Fonte: Autora)

Finalmente, a média móvel foi estabelecida no intervalo 2 medidas observadas.

A Figura 18 apresenta o conjunto de dados original e a conjunto de dados da previsão da série temporal para os últimos 24 meses.

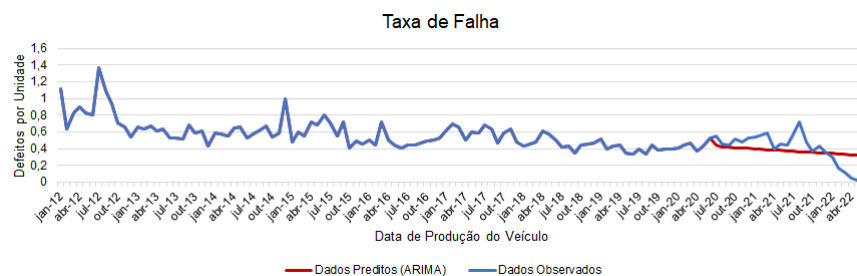


Figura 18 – Gráfico de Predição de Série Temporal a partir do Modelo ARIMA (Fonte: Autora)

Em termos de performance, para o modelo ARIMA, foi identificado MAPE de

3,12%, o RMSE de 0,17 e o coeficiente de correlação de Pearson de 0,74, fortemente positivamente correlacionado com o conjunto de dados observados.

4.3 Predição de Séries Temporais a partir do Modelo LSTM

A Figura 19 apresenta a arquitetura da rede neural LSTM implementada.

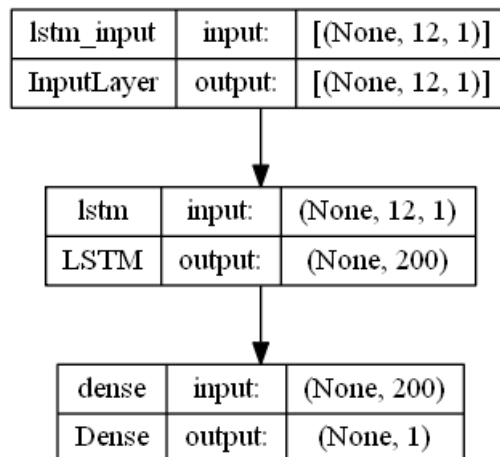


Figura 19 – Arquitetura LSTM (Fonte: Autora)

O algoritmo ajusta e treina o modelo LSTM com o conjunto de dados de treinamento 20 épocas em 1 batch. A performance de cada época é avaliada para minimizar a média do erro quadrático frente ao conjunto de dados. O modelo treinado é exposto ao conjunto de dados de teste gerando o conjunto de dados predito. A Figura 20 apresenta o conjunto de dados original e o conjunto de dados da previsão da série temporal para os últimos 24 meses.

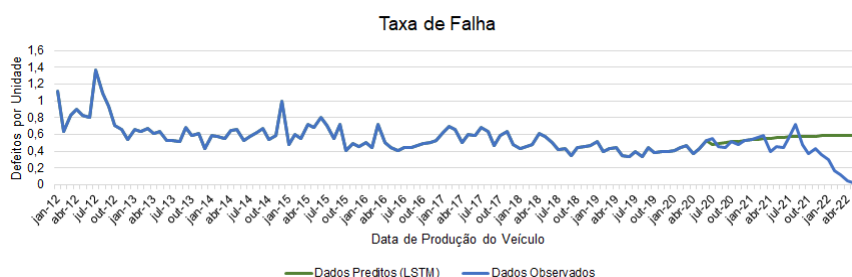


Figura 20 – Gráfico de Predição de Série Temporal a partir do Modelo LSTM (Fonte: Autora)

Em termos de performance, para o modelo LSTM, foi identificado o MAPE de 5,05%, o RMSE de 0,22 e o coeficiente de correlação de Pearson de $-0,69$, fortemente negativamente correlacionado com o conjunto de dados original.

4.4 Comparação dos Resultados dos Modelos de Predição

A Figura 21 apresenta o resultado gráfico comparado o desempenho das 2 soluções propostas.

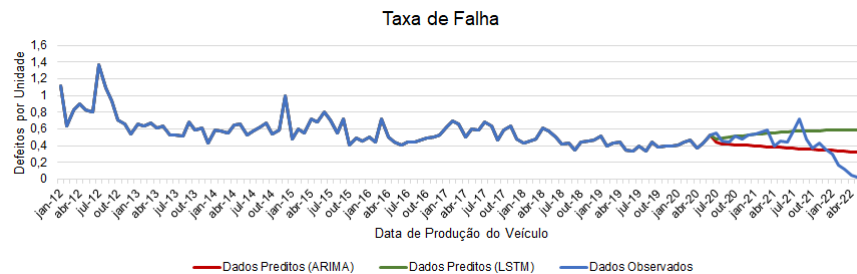


Figura 21 – Comparação dos Dados Preditos versus Dados Observados (Fonte: Autora)

Tanto o modelo de ARIMA, quanto o de LSTM apresentaram a métrica de erro bastante equivalente quando comparamos o desempenho dos valores preditos frente aos valores observados. Entretanto, identificamos visualmente que os resultados apresentados pelos modelos ARIMA e LSTM apresentam uma característica de tendência contrárias entre si, sendo que o modelo ARIMA está aderente ao valor observado no conjunto de dados real. A avaliação de performance dos modelos concluiu que o conjunto de dados predito pelo modelo ARIMA apresenta uma correlação positiva com o dados observado com um erro quadrático médio menor do que o resultado apresentado pelo modelo LSTM. Dessa forma, a solução apresentada pelo modelo ARIMA proposto está mais adequada ao problema analisado neste trabalho.

5 CONCLUSÃO

Ao longo deste trabalho, foram apresentadas e discutidas as principais técnicas de ciências de dados e redes neurais que são utilizadas em um problema de análise de séries temporais. Uma série temporal foi construída a partir de duas base de dados, sendo uma base de dados de produção e outra de história de falhas em garantia, utilizando os conceitos de gerenciamento de bancos de dados. Essa série temporal foi analisada de forma a descrever o conjunto de dados gerado e identificar as principais características da série. Finalmente, dois modelos amplamente utilizados na literatura deste problema, ARIMA e LSTM, foram implementados, tendo a sua performance avaliada e comparada. Concluiu-se que para o presente conjunto de dados, o resultado gerado pelo modelo ARIMA se adequa melhor ao problema de previsão de séries temporais.

A principal contribuição deste trabalho foi conduzir um estudo sistemático reunindo ferramentas de ciência de dados e engenharia de dados para investigar o problema de previsão de séries temporais. Para o problema proposto, novos trabalhos podem ser desenvolvidos no âmbito das técnicas implementadas, de forma a refinar os parâmetros adotados para os modelos e aplicar outras técnicas também utilizadas em problemas envolvendo séries temporais, como por exemplo, Random Forest, Redes Neurais Convolucionais e Modelos Híbridos. Com o objetivo de aprofundar o estudo da série temporal baseada em dados de garantia, novos trabalhos podem ser desenvolvidos de forma a escalar a implementação da metodologia para outros modelos de veículos e a identificar quais sistemas ou itens são potenciais causadores de falhas em garantia.

REFERÊNCIAS

- ASSIS, E. M. D. *et al.* Machine learning and q-weibull applied to reliability analysis in hydropower sector. **IEEE Access**, v. 8, p. 203331–203346, 2020.
- BRAGA, A.; CARVALHO, A. P. d. L. D.; LUDERMIR, T. B. **Redes Neurais Artificiais: teoria e aplicações**. 1. ed. [S.l.: s.n.]: LTC, 2007. 1–262 p.
- BROWNLEE, J. **Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python**. [S.l.: s.n.]: Machine Learning Mastery, 2018. 1–575 p.
- CORREA, J. Aplicação de Redes Neurais LSTM para Previsão de Séries Temporais Financeiras Aplicação de Redes Neurais LSTM para Previsão de Séries Temporais Financeiras. 2021.
- EHLERS, R. S. Análise de séries temporais. 2007.
- ELERATH, J. G.; PECHT, M. IEEE 1413: A standard for reliability predictions. **IEEE Transactions on Reliability**, v. 61, n. 1, p. 125–129, 2012. ISSN 00189529.
- HAYKIN, S. **Redes Neurais: Princípios e Práticas**. 3. ed. [S.l.: s.n.]: Bookman, 2008. 1–889 p. ISBN 9788521203896.
- HIPP, J.; LINDNER, G. Analysing warranty claims of automobiles an application description following the crisp-dm data mining process. **Lecture Notes in Computer Science**, v. 1749, p. 31–41, 1999. ISSN 16113349.
- HRYCEJ, T.; GRABERT, M. Warranty cost forecast based on car failure data. **IEEE International Conference on Neural Networks - Conference Proceedings**, n. m, p. 108–113, 2007. ISSN 10987576.
- KADHEM, A. A. *et al.* Advanced wind speed prediction model based on a combination of weibull distribution and an artificial neural network. **Energies**, v. 10, n. 11, 2017. ISSN 19961073.
- KAILASH, K.; PECHT, M. Reliability estimation techniques. *In*: _____. **Reliability Engineering**. [S.l.: s.n.]: John Wiley & Sons, Ltd, 2014. cap. 13, p. 247–293. ISBN 9781118841716.
- KHOSHKANGINI, R. *et al.* Early prediction of quality issues in automotive modern industry. **Information (Switzerland)**, v. 11, n. 7, p. 15, 2020. ISSN 20782489.
- KUTYLOWSKA, M. Neural network approach for failure rate prediction. **Engineering Failure Analysis**, v. 47, p. 41–48, 2015. ISSN 13506307.
- LEE, J. G. *et al.* Automobile parts reliability prediction based on claim data: The comparison of predictive effects with deep learning. **Engineering Failure Analysis**, Elsevier Ltd, v. 129, n. October 2020, p. 105657, 2021. ISSN 13506307.

LEE, S. *et al.* A study on warning / detection degree of warranty claims data using neural network learning. *In: Conference on Advanced Language Processing and Web Information Technology*. [S.l.: s.n.], 2007. p. 492–497. ISBN 0769529305.

LEE, S. H. *et al.* A study on development of time series warning module in warranty claims database. **Proceedings - ALPIT 2007 6th International Conference on Advanced Language Processing and Web Information Technology**, p. 553–558, 2007.

LI, Z.; MOBIN, M. A dfmea-based reliability prediction approach in early product design. *In: Annual Reliability and Maintainability Symposium (RAMS)*. [S.l.: s.n.], 2018. p. 1–7.

MAISONNIER, D. Rami: The main challenge of fusion nuclear technologies. **Fusion Engineering and Design**, Elsevier, v. 136, p. 1202–1208, 2018. ISSN 09203796.

MORETTIN, P.; TOLOI, C. **Análise de Séries Temporais**. 2. ed. [S.l.: s.n.]: Blucher, 2006. 1–538 p. ISBN 9788521203896.

SABRI-LAGHAIE, K. *et al.* Early detection of product reliability based on the parameters of the production line and warranty data. **International Journal of Reliability, Quality and Safety Engineering**, v. 28, n. 5, p. 1–19, 2021. ISSN 02185393.

SAPANKEVYCH, N.; SANKAR, R. Time series prediction using support vector machines: A survey. **IEEE Computational Intelligence Magazine**, v. 4, n. 2, p. 24–38, 2009. ISSN 1556603X.

SIAMI-NAMINI, S.; TAVAKOLI, N.; Siami Namin, A. A Comparison of ARIMA and LSTM in Forecasting Time Series. **Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018**, IEEE, p. 1394–1401, 2019.

SINHA, R. S. e. a. Failure rate analysis of jaw crusher using weibull model. **Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering**, v. 231, n. 4, p. 760–772, 2017. ISSN 09544089.

STAMENKOVIC, D. Warranty optimisation based on the prediction of costs to the manufacturer using neural network model and monte carlo simulation. **International Journal of Systems Science**, v. 46, n. 3, p. 535–545, 2015. ISSN 14645319.

STOLFI, G. Ptc2527 confiabilidade. 2017.

TAVARES, U.; QUADRELLI, G. Análise Comparativa entre os Modelos ARIMA e LSTM na Previsão de Curto Prazo da Demanda de Potência Ativa. v. 14, n. 1, p. 60–70, 2022.

TEO, H. G. **Data Mining in Automotive Warranty Analysis**. 2010. 1–159 p. Tese (Doutorado) — Victoria University of Wellington, 2010.

THOMAS, M. U. **Reliability and warranties: Methods for product development and quality improvement**. [S.l.: s.n.], 2006. 1–191 p. ISBN 9781420006575.

TORGUNOV, D. *et al.* Vehicle warranty claim prediction from diagnostic data using classification. *In: Advances in Computational Intelligence Systems*. [S.l.: s.n.], 2020. p. 483–492. ISBN 9783030299330.

WU, S. Warranty data analysis: A review. **Quality and Reliability Engineering International**, v. 28, n. 8, p. 795–805, 2012. ISSN 07488017.

WU, S. A review on coarse warranty data and analysis. **Reliability Engineering and System Safety**, Elsevier, v. 114, n. 1, p. 1–11, 2013. ISSN 09518320.

WULPI, D. J. **Understanding How Components Fail**. [*S.l.: s.n.*], 2013. 1–344 p. ISBN 9781627080149.

YANG, Z. *et al.* Reliability assessment of cnc machining center based on weibull neural network. **Mathematical Problems in Engineering**, v. 2015, 2015. ISSN 15635147.