UNIVERSIDADE FEDERAL DE JUIZ DE FORA
CURSO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

FELIPE CARRULLO PUMAR

Aplicação da mineração de processos para a previsão do tempo de conclusão de chamados de TI

JUIZ DE FORA
2022

FELIPE CARRULLO PUMAR

Aplicação da mineração de processos para a previsão do tempo de conclusão de chamados de TI

Trabalho de Conclusão de Curso apresentado a Faculdade de Engenharia da Universidade Federal de Juiz de Fora, como requisito parcial para a obtenção do título de Engenheiro de Produção.

Orientador: DSc. Antônio Ângelo Missiaggia Picorone

JUIZ DE FORA

2022

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Pumar, Felipe Carrullo.
Aplicação da mineração de processos para a previsão do tempo de conclusão de chamados de TI / Felipe Carrullo Pumar. -- 2022.
58 p.

Orientador: Antônio Ângelo Missiaggia Picorone
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, ICE/Engenharia, 2022.

1. Acordo de Nível de Serviço. 2. Mineração de Processos. 3. Regressão Não-paramétrica. I. Picorone, Antônio Ângelo Missiaggia, orient. II. Título.

AGRADECIMENTOS

Este trabalho marca um novo ciclo que se inicia. Espero que esse novo momento venha cheio de novos desafios e me proporcione grande crescimento assim como o meu período na universidade fez.

Agradeço em primeiro lugar a Deus, por todas as oportunidades que tive desde que entrei na faculdade. Agradeço pelos momentos bons e pelos momentos difíceis que passei que testaram meu caráter e possibilitaram que eu aprendesse com os meus erros.

Agradeço a minha mãe que com muito amor sempre foi a primeira pessoa a vibrar com as minhas conquistas e agradeço o meu pai pelos conselhos e conversas que me moldaram profissionalmente e pessoalmente. As minhas irmãs Clara e Louise, agradeço demais o companheirismo de sempre, vocês são incríveis.

Não posso deixar de agradecer todos os amigos que fiz durante esses anos de faculdade. Aos que já começaram na minha sala e aos que fiz através de tantos trabalhos, jantares no RU e horas despendidas na Mais Consultoria, GET Produção e Liga das Empresas Juniores, muito obrigado por tantos momentos felizes e todo o apoio nessa jornada. Giulia, muito obrigado por todo o companheirismo e carinho em diversos momentos.

Por fim, não posso deixar de agradecer ao meu orientador Antônio Picorone cuja parceria e suporte foram imprescindíveis para a realização desse trabalho. No mais, agradeço a todos os professores e profissionais do curso de Engenharia de Produção, vocês me inspiram!

FELIPE CARRULLO PUMAR

Aplicação da mineração de processos para a previsão do tempo de conclusão de chamados de TI

Trabalho de Conclusão de Curso apresentado a Faculdade de Engenharia da Universidade Federal de Juiz de Fora, como requisito parcial para a obtenção do título de Engenheiro de Produção.

Aprovada em 14 de fevereiro de 2022.

BANCA EXAMINADORA

DSc. Antônio Ângelo Missiaggia Picorone (Orientador)

Universidade Federal de Juiz de Fora

p/ DSc. Marcos Vinícius Rodrigues

Universidade Federal de Juiz de Fora

p/ Thais de Andrade Maestro

Ambev

RESUMO

Os acordos de nível de serviço (SLA) são importantes para garantir os deveres e responsabilidades entre os solucionadores e os usuários. Entretanto, sua gestão é comumente feita de forma reativa. Este trabalho tem como objetivo principal investigar a aplicação das ferramentas de mineração de processo (MP) e regressões não-paramétricas (NPR) para auxiliar a gerenciar proativamente o SLA. No presente trabalho o SLA e suas implicações nas organizações foram estudados, foram apresentadas diferentes técnicas de MP e de NPR e os modelos construídos foram aplicados em um estudo de caso com dados reais. Os resultados obtidos no estudo de caso sugerem que o modelo da regressão de Kernel Gaussiana, utilizando as métricas de tempo decorrido e da ocorrência e duração das atividades, obteve os melhores resultados em prever o tempo remanescente de chamados de TI em andamento. Os resultados obtidos nesse trabalho também sugerem que a MP e os modelos de NPR podem ser considerados como uma opção para prever o tempo remanescente de chamados de TI e auxiliar no gerenciamento proativo do SLA.

Palavras-chave: Acordo de Nível de Serviço, Mineração de Processos, Regressão Não-paramétrica

ABSTRACT

Service level agreements (SLA) are important to ensure the duties and responsibilities between the service providers and users. However, SLA management is commonly done reactively. This study has the objective to investigate the application of process mining (PM) and non-parametric regressions (NPR) to help to proactively manage the SLA and analyze different forecast models. In the present work, SLA and its implications in organizations were studied, different PM and NPR techniques were presented, and the models built were implemented in a case study with real data. The results obtained in the case study suggest that the Gaussian Kernel regression utilizing metrics of elapsed time and the occurrence and duration of activities, obtained the best results in predicting the remaining time of ongoing IT tickets. The results obtained in this work also suggest that PM and NPR models can be considered as an option for predicting remaining time of ongoing IT ticket and help to proactively manage SLA.

Keywords: Service Level Agreement, Process Mining, Non-parametric Regression

LISTA DE FIGURAS

Figura 1- Metodologia do estudo 17
Figura 2 - Entradas e saídas dos três tipos básicos de mineração de processos a) descoberta b) conformidade e c) aprimoramento 24
Figura 3 – Exemplo de um registro de eventos 24
Figura 4 - Comparação entre diferentes regressões paramétricas 28
Figura 5 – Comparação entre uma regressão linear e uma regressão não paramétrica 28
Figura 6 – Regressões não paramétricas considerando parâmetros de suavização diferentes 29
Figura 7 – Classe de funções de regressão de Kernel Gaussiana e de Epanechnikov 36
Figura 8 –Atividades dos chamados ao longo do ano 40
Figura 9 – Atividades dos chamados ao longo de sua duração 40
Figura 10 – Fluxograma do processo minerado do RE 41
Figura 11 – Histogramas da duração dos chamados por *status* 42
Figura 12 – Histogramas da duração em horas úteis dos chamados por *status* 43
Figura 13 – Tempo remanescente médio por *status* 46
Figura 14 – MAE por status do modelo de referência 47
Figura 15 – Erros médios por status para a regressão de Kernel Gaussiana 49
Figura 16 – Percentual de erro entre a regressão de Kernel Gaussiana e o modelo de referência 50
Figura 17 – Erros médios por status para a Regressão de Kernel de Epanechnikov.. 51
Figura 18 – Percentual de erro entre a regressão de Kernel de Epanechnikov e o modelo de referência 52
Figura 19 – Percentual de chamados classificados corretamente 53

LISTA DE TABELAS

Tabela 1 – Exemplo de um registro de eventos 23
Tabela 2 – Exemplo da medida de ocorrências das atividades para cada caso 26
Tabela 3 – Exemplo da medida de duração média (em dias) das atividades para cada caso 27
Tabela 4 – Funções de Kernel 30
Tabela 5 – Chamado do RE utilizado no estudo de caso 39
Tabela 6 – Métrica do tempo decorrido (ID R4906956) 44
Tabela 7 – Métrica de ocorrência de atividades (ID R4906956) 44
Tabela 8 – Métrica de duração das atividades em horas (ID R4906956) 45
Tabela 9 – Conjunto de todas as métricas (ID R4906956) 45
Tabela 10 – Melhores resultados das regressões de Kernel 48

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

CS – Centrais de Serviço

ID – Identificador único

MAE – *Mean Absolute Error* ou erro absoluto médio

MP – Mineração de processos

NPR - *Non-parametric Regression* ou regressão não paramétrica

PAIS - *Process aware information systems* ou sistemas de informação cientes do processo

RE – Registro de eventos

SLA – *Service Level Agreement* ou acordo de nível de serviço

TI – Tecnologia da Informação

SUMÁRIO

1. INTRODUÇÃO 13

1.1 JUSTIFICATIVA 14
1.2 ESCOPO 15
1.3 OBJETIVOS 16
1.4 METODOLOGIA 17
1.5 ESTRUTURA DO TRABALHO 18
1.6 SUMÁRIO DO CAPÍTULO 18

2. REVISÃO DE LITERATURA 20

2.1 ACORDO DE NÍVEL DE SERVIÇO (SLA) 20
2.2 MINERAÇÃO DE PROCESSOS (MP) 22
2.3 PREVISÃO DE CASOS EM ANDAMENTO 25
2.4 REGRESSÃO NÃO-PARAMÉTRICA (NPR) 27
2.5 NORMALIZAÇÃO DOS DADOS 31
2.6 DADOS DE TREINO E TESTE 31
2.7 SUMÁRIO DO CAPÍTULO 32

3. METODOLOGIA DE IMPLEMENTAÇÃO DOS MODELOS 33

3.1 ANÁLISE EXPLORATÓRIA E PRÉ-PROCESSAMENTO DOS DADOS 33
3.2 MINERAÇÃO DAS MÉTRICAS DO RE 34
3.3 NORMALIZAÇÃO DOS DADOS 34
3.4 SEPARAÇÃO DOS DADOS DE TREINO E TESTE 34
3.5 IMPLEMENTAÇÃO DO MODELO DE REGRESSÃO 35
3.6 COMPARAÇÃO DOS RESULTADOS OBTIDOS 36
3.7 SUMÁRIO DO CAPÍTULO 37

4. RESULTADOS OBTIDOS NO ESTUDO DE CASO 38

4.1 ANÁLISE EXPLORATÓRIA E PRÉ-PROCESSAMENTO DOS DADOS 38

4.2 MINERAÇÃO DAS MÉTRICAS DO RE 43
4.3 NORMALIZAÇÃO DOS DADOS 45
4.4 SEPARAÇÃO DOS DADOS DE TREINO E TESTE 45
4.5 IMPLEMENTAÇÃO DO MODELO DE REGRESSÃO 46
4.5.1 MODELO DE REFERÊNCIA 46
4.5.2 REGRESSÕES DE KERNEL 47
4.6 SUMÁRIO DO CAPÍTULO 54

5. CONCLUSÕES 55

REFERÊNCIAS 56

1. INTRODUÇÃO

Atualmente, a Tecnologia da Informação (TI) desempenha um papel fundamental na gestão empresarial, uma vez que a maioria das atividades empresariais são baseadas em sistemas ou tecnologias de informação (NETO, 2009).

Para que esses sistemas se mantenham funcionando, além de atender outras demandas que usuários de uma organização possuam relacionados a área de TI, existem as plataformas chamadas de Centrais de Serviço (CS). Na CS, os usuários podem registrar suas demandas, conhecidas como chamados, que serão então atendidas pelo time de TI. Dessa forma, as CS geram e armazenam um grande volume de dados referentes aos processos realizados por esses times (FREITAS, 2011).

A grande maioria dos sistemas de informações usados nas CS de TI são classificados como sistemas de informação cientes do processo (PAIS – *process aware information systems*). Um sistema PAIS armazena informações sobre eventos executados em determinado processo. Alguns exemplos de PAIS são os Sistemas de Gerenciamento do Fluxo de Trabalho (WMS – *Workflow Management System)*, Gestão de Relacionamento com o Cliente (CRM – *Customer Relationship Management)* e Planejamento dos Recursos Empresariais (ERP – *Enterprise Resources Planning*) (JUNIOR, 2017).

Dados referentes ao processo de TI são armazenadas em registros de eventos (RE). Um RE contém informações sobre as atividades que foram tomadas para a resolução de uma demanda dos usuários. No RE também são incluídos dados referentes à duração de cada atividade realizada, além de outras informações que auxiliam o enriquecimento do RE como detalhes sobre os recursos ou responsáveis que participaram em cada atividade (VAN DER AALST; *et al.*, 2004).

Deveres e responsabilidades entre os solucionadores e os usuários são definidos pelos Acordos de Nível de Serviço (SLA - *Service Level Agreement)* (MAGER, 2014). Um exemplo de acordo, é o tempo de resposta máximo que o time de solucionadores de TI tem para atender à um chamado. Esse tempo de resposta serve como uma garantia para o usuário de que toda vez que ele tiver esse problema, este será resolvido no prazo acordado. Por outro lado, o SLA também auxilia os solucionadores a dimensionar seu time e negociar uma remuneração de acordo com o nível de serviço demandado (RAMAKRISHNAN; KAUR, 2020). Dessa forma, monitorar o cumprimento desses acordos é relevante nas organizações para evitar os custos de não atendimento e manter o nível de qualidade da organização.

No entanto, o monitoramento do SLA das empresas é comumente feito de forma reativa, em que violações que já ocorreram são avaliadas e medidas preventivas são traçadas para o futuro. Enquanto essa abordagem pode ser útil, uma vez que pode alertar a organização para possíveis problemas de qualidade em seu processo, ela não impede os erros de acontecerem. Para isso, um gerenciamento antes da ocorrência é preferível, uma vez que permite a previsão de possíveis violações de SLA (LEITNER; *et al.*, 2009). Um primeiro passo para realizar uma gestão proativa das violações, é analisar os processos abordados pelo acordo de forma proativa.

A mineração de processos (MP) é uma disciplina de pesquisa relativamente nova que oferece métodos poderosos para extrair (minerar) conhecimento de REs de processos já executados. Desse modo, a principal vantagem da MP é sua independência de plataforma e formato. Registros de eventos de todo tipo de sistema de informação podem ser usados e transformados por um processo de mineração (VAN DER AALST; *et al.*, 2012).

Portanto, através da mineração de processos, é possível extrair informações dos REs de chamados de TI. Essas informações, por sua vez, podem ser utilizadas em modelos de previsão para prever possíveis violações do SLA, assim, gerenciando o SLA de maneira proativa. O presente estudo investiga os diferentes modelos que podem ser utilizados para gerenciar proativamente o cumprimento do SLA, além de implementá-los em um estudo de caso.

1.1 JUSTIFICATIVA

SLA’s são traçados entre ofertantes e usuários de um serviço para que se tenham regras bem definidas dos deveres e responsabilidades de ambas as partes. Dessa forma, violações dos SLA’s causam diversos prejuízos para ambas as partes, como atrasos nas entregas, multas e quaisquer outras penalidades explicitadas no acordo (MAGER, 2014).

No entanto, o gerenciamento após o ocorrido é ineficiente, pois não consegue evitar o prejuízo que já foi causado. Para evitar que o SLA seja violado, em especial, os acordos referentes ao tempo máximo de resposta do atendimento de uma demanda, é preciso prever o infringimento antes de sua ocorrência.

Dado os recentes avanços e novas técnicas de mineração de processos, se torna relevante um estudo das diferentes abordagens disponíveis na MP para a predição do tempo remanescente de casos em andamento. Portanto, o presente trabalho visa explorar tais técnicas,

e aplicá-las em casos reais de uma empresa brasileira de grande porte através de um estudo de caso.

1.2 ESCOPO

Acordos de nível de serviço podem metrificar uma grande variedade de indicadores como: disponibilidade, precisão, atualidade, qualidade, capacidade e desempenho de um sistema (EDDINGTON, 2010). O presente estudo, se limita à análise de desempenho, através do tempo de resposta de casos em andamento. A utilização de técnicas de MP para esse fim se utiliza da previsão desses tempos para auxiliar na gestão proativa do SLA. Dessa forma, não será proposto um modelo de gestão do SLA, uma vez que esse tema é muito abrangente e fica fora do escopo desse estudo, mas os modelos apresentados e aplicados nesse estudo auxiliarão na tomada de decisão preventiva.

Dentro da mineração de processos há diversas técnicas para extrair conhecimento de registro de eventos. Os objetivos dessas técnicas incluem: o descobrimento de modelos de processo, verificação de conformidade, aprimoramento de modelos de processo, previsão de casos dos processos, entre outros (VAN DER AALST; *et al.*, 2010). Esse estudo é voltado para as técnicas de previsão do tempo remanescente de processos em andamento, para que se possa prever violações nos prazos acordados no SLA. Dessa forma, a mineração de processos pode ser utilizada para extrair do RE métricas que alimentam um modelo preditivo, conforme proposto por Van Dongen *et al* (2008).

A fim de prever os tempos de duração de casos em um processo não especificado, não se pode assumir uma forma específica para a distribuição dos tempos dos casos desse processo. Dessa forma, para realizar previsões é vantajoso a utilização de modelos de regressão não paramétrica (NPR – *Non-parametric Regression*) em oposição a outros métodos de previsão, principalmente porque a NPR é mais adequada em situações com poucos ou nenhum precedente disponível. A abordagem não paramétrica apenas assume que há alguma relação entre as variáveis preditoras e a variável alvo, a forma dessa relação não precisa ser especificada (VAN DONGEN; CROOY; VAN DER AALST, 2008). Portanto, no presente estudo uma NPR chamada regressão de Kernel (FAN; GIJBELS, 2018) é utilizada como modelo preditivo, uma vez que a regressão de Kernel oferece grande versatilidade para lidar com diversos tipos de dados, como observado no estudo de Van Dongen *et al* (2008).

Há diversas ferramentas que podem ser utilizadas para a implementação dessas técnicas. Alguns programas como Celonis[1] e o Disco[2] são bem utilizados no mercado, mas possuem custos elevados para o uso não corporativo (KEBEDE; DUMAS, 2015). Dessa forma, a fim de realizar o estudo de caso no presente trabalho, foi utilizada a ferramenta Prom[3] e a linguagem de programação Python[4].

O Prom é uma ferramenta de código aberto especialmente adaptada para apoiar o desenvolvimento de projetos de mineração de processo. Com o auxílio das extensões *Project Log on Dotted Chart* e *Mine With Visual Miner* da ferramenta, foi realizada uma análise exploratória dos dados utilizados.

Para a aplicação em Python, a fim de facilitar o uso da linguagem foram utilizadas as bibliotecas Numpy[5] e Pandas[6] para realizar o pré-processamento e a modelagem dos modelos de regressão, as bibliotecas *Matplotlib*[7] e Seaborn[8] para montar as visualizações dos resultados e a biblioteca *Scikit-Learn*[9] para normalizar e separar os dados de treino e teste.

1.3 OBJETIVOS

O presente trabalho tem como **objetivo geral** investigar a aplicação das ferramentas de mineração de processos para gerenciar proativamente o SLA. Para isso, também são analisados diferentes modelos de previsão e implementados em um estudo de caso. Os objetivos específicos do estudo são:

1) Conhecer as metodologias para extração de conhecimento acerca do tempo de duração dos chamados;

2) Conhecer diferentes metodologias para a previsão da duração de casos em andamento;

---

[1] https://www.celonis.com/

[2] https://fluxicon.com/disco/

[3] https://www.promtools.org/doku.php

[4] https://docs.python.org/pt-br/

[5] https://numpy.org/doc/

[6] https://pandas.pydata.org/docs/

[7] hhttps://matplotlib.org/2.0.2/api/pyplot_api.html

[8] http://seaborn.pydata.org/introduction.html

[9] https://scikit-learn.org/stable/

3) Selecionar as metodologias relevantes de mineração de processo e de predição de tempo remanescente de chamados e discutir seus benefícios para o modelo;

4) Aplicar a mineração de processos e o modelo preditivo em um estudo de caso.

1.4 METODOLOGIA

O presente estudo segue a metodologia indicada na Figura 1.

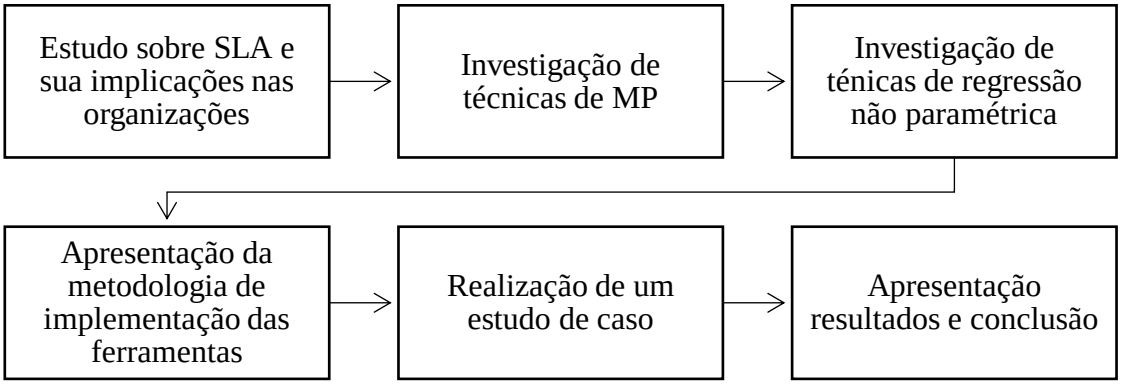

Figura 1- Metodologia do estudo
Fonte: O Autor

Em busca de atingir os objetivos apresentados na seção 1.3 foram realizados os seguintes passos:

1) O estudo sobre o SLA e suas implicações nas organizações foi conduzido através de uma revisão da literatura sobre o tema. A pesquisa realizada considerou livros e artigos sobre gestão de TI, com destaque para Wu e Buyya (2012) e Halvey e Melby (2007).

2) As técnicas de MP foram também estudadas através de pesquisas bibliográficas com destaque para os conceitos e as técnicas de MP para a previsão de duração dos casos em andamento. A maior parte da revisão se baseou em artigos acadêmicos publicados pelo grupo de especialistas em mineração de processo da Universidade de Tecnologia de Eindhoven, criadores de diversos conceitos sobre a mineração de processos. Também foram utilizados outros artigos diversos que ampliaram os primeiros conceitos introduzidos através de novas técnicas e abordagens do assunto;

3) Para a investigação das técnicas de NPR foram utilizadas literaturas sobre os conceitos básicos da NPR e da regressão de Kernel;

4) O método proposto para a utilização da MP em conjunto com o modelo preditivo é detalhado neste trabalho e os passos para a implementação e avaliação desse modelo

são discutidos. O modelo utilizado foi baseado no estudo desenvolvido por Van Dongen *et al.* (2008);

5) É apresentado um estudo de caso. O estudo de caso utiliza de um RE com dados reais sobre o processo de atendimento de TI de uma empresa de grande porte. Para a mineração dos processos e aplicação dos modelos preditivos, é utilizado a linguagem Python;

6) Por fim, são apresentados os resultados e as conclusões do estudo. Essa etapa contará com uma análise das diferentes técnicas de MP empregadas e seus resultados no estudo de caso.

1.5 ESTRUTURA DO TRABALHO

O presente trabalho foi dividido em 5 capítulos.

O capítulo 1 contém a introdução do estudo que será realizado, a contextualização do problema sendo abordado, a delimitação do que será e não será abordado no trabalho, a metodologia do estudo e do estudo de caso que será implementado além da apresentação dos objetivos gerais e específicos sendo buscados;

No capítulo 2 é apresentada a revisão da literatura em que o trabalho está embasado, dentro das seguintes áreas: acordos de nível de serviço, registro de eventos e mineração de processos, além de alguns tópicos sobre abordagens de previsão de processos e modelos preditivos.

No capítulo 3 é apresentada a metodologia de implementação dos modelos de MP em conjunto com o NPR.

No capítulo 4 discute-se um estudo de caso em que técnicas de MP e preditivas são utilizadas para estimar o tempo remanescente de chamados de TI, oferecendo uma oportunidade para auxiliar na gestão proativa do SLA.

Finalmente, o capítulo 5 apresenta as conclusões do trabalho, lições aprendidas e um direcionamento para o futuro.

1.6 SUMÁRIO DO CAPÍTULO

Para auxiliar na gestão proativa do SLA, um estudo sobre a previsão do tempo de conclusão de chamados em andamento é relevante. O presente trabalho tem como objetivo geral

investigar a aplicação das ferramentas de mineração de processos para gerenciar proativamente o SLA. Para isso técnicas de MP e NPR são estudadas e uma metodologia conectando os temas é apresentada e aplicada em um estudo de caso.

2. REVISÃO DE LITERATURA

Esse capítulo aborda o referencial teórico que serve de base para a pesquisa proposta. Inicialmente, é definido o conceito de acorde de nível de serviço (SLA) e suas implicações na área de TI. Em seguida, são apresentados os conceitos de mineração de processos (MP) e de transformações do registro de eventos (RE). Finalmente, são apresentados os fundamentos matemáticos por trás dos modelos preditivos que são aplicados no estudo de caso.

2.1 ACORDO DE NÍVEL DE SERVIÇO (SLA)

Um acordo de nível de serviço é um compromisso mútuo e define as expectativas referentes ao serviço pelo prestador do serviço e pelo consumidor do serviço. Essas garantias referenciam qual transação precisa ser executada e qual o nível de qualidade esperado na execução (WU; BUYYA, 2012).

Os acordos dos serviços de TI, tratam principalmente de questões relacionadas ao desempenho e orientam as informações coletadas durante a execução do processo. A maioria dos contratos incluem expectativas sobre os serviços solicitados, métricas para definir as metas de limite para esses SLAs, métodos padrão para medir cada SLA e indicadores para desvios dos SLAs reais aos acordados (HALVEY; MELBY, 2007).

Um SLA deve abordar os seguintes pontos (WU; BUYYA, 2012):

- Propósito: descreve as razões por trás da criação do acordo;
- Partes envolvidas: descreve quais são as partes envolvidas no acordo e quais os seus respectivos papéis (provedor e consumidor);
- Período de validade: define o período em que o acordo terá validade, delimitados por um período de início e conclusão do acordo;
- Escopo: define e detalha todos os serviços que estão sendo abrangidos pelo acordo em questão;
- Restrições: define os passos que devem ser tomados para que o nível de serviço possa ser fornecido;
- Objetivos do nível de serviço (SLO): o nível de serviço em que o usuário do serviço e o provedor acordam, normalmente inclui os indicadores que devem ser inclusos no acordo, como disponibilidade, desempenho e confiabilidade. Cada indicador de nível de serviço deve possuir um valor alvo que deve ser alcançado.

– Penalidades: define o que deve ocorrer caso o provedor do serviço tenha desempenho inferior ao acordado e seja incapaz de cumprir com os objetivos estipulados. Se o acordo ocorrer com um provedor de serviços externo, uma opção de rescindir o contrato caso níveis de serviço inaceitáveis sejam apresentados deve ser incluído.

– Serviços opcionais: define os serviços que normalmente não são demandados pelos usuários, mas podem ser exigidos como exceção;

– Exclusões: especifica o que não é coberto pelo acordo;

– Administração: descreve os processos criados no SLA para atender e medir seus objetivos, além de definir a responsabilidade organizacional para supervisionar cada um deles processos.

Grande parte dos SLAs de TI ainda são formalizados através de arquivos de texto. Para identificar as violações de tais acordos, os clientes acompanham indicadores de desempenho dos serviços sendo prestados. Para monitorar são estabelecidos limites mínimos e máximos para cada métrica, bem como sanções que podem ser aplicadas ao prestador do serviço a partir de uma violação dos acordos definidos. Incentivos ou penalidades auxiliam encorajar um desempenho superior ou desencorajar um desempenho ruim (PINHEIRO, 2015).

As métricas utilizadas no SLA podem ser categorizadas como aquelas que fazem referência à execução do processo e as que fazem referência à interrupção do processo. Na primeira, o objetivo é uma integração positiva da tecnologia no processo de negócios do cliente. Por outro lado, SLA com métricas de interrupção buscam reduzir o impacto negativo que a tecnologia tem com o negócio do cliente (EDGINGTON; RAGHU; VINZE, 2010).

As métricas de execução do processo incluem disponibilidade, precisão, atualidade, qualidade, capacidade e desempenho de um sistema. O objetivo dessas métricas pode ser visto como o desejo de uma integração perfeita da tecnologia no processo de negócios do cliente, no entanto nem todos os aspectos de TI são totalmente integrados. Existem algumas, como a instalação de upgrades e o treinamento do usuário que, ao interromper os negócios gerais do cliente, têm como objetivo proporcionar melhorias na execução do processo. As atividades de interrupção do processo costumam ser o foco das CS de TI. As métricas de interrupção incluem métricas como resolução de problemas, tempo de recuperação, incidências de chamadas repetidas, tempo de reparo, porcentagem de chamadas abandonadas e incidentes de segurança. Essas atividades impactam negativamente o processo geral de negócios do cliente e, uma vez resolvidas, apenas permitem que o processo interrompido seja retomado (EDGINGTON; RAGHU; VINZE, 2010).

Ademais, os tempos entre o início e a conclusão de um processo de atendimento são informações úteis para avaliar a eficiência operacional do processo. Esse tempo, conhecido como tempo de ciclo, é um contribuidor crucial para a experiência geral do usuário, produtividade, receita organizacional e imagem da marca. No setor de serviços, tempos de ciclo longos podem frustrar os usuários que precisam esperar para terem suas demandas atendidas. O tempo de ciclo longo em um centro de atendimento ao cliente pode resultar na perda de clientes ou publicidade negativa para uma organização. Além disso, o tempo de ciclo mais longo pode significar oportunidades para melhorar a eficiência de um processo existente em termos de economia de custo e tempo e de aumentar a satisfação dos clientes. Logo, o tempo de ciclo é, predominantemente, usado como uma meta de desempenho e faz parte dos SLAs como uma das principais métricas que garantem a qualidade da entrega dos serviços aos clientes (RAMAKRISHNAN; KAUR, 2020).

2.2 MINERAÇÃO DE PROCESSOS (MP)

Atualmente, a maioria das organizações utilizam sistemas de informação para dar suporte a execução dos seus processos. Exemplos de sistemas de informação são WMS, CRM, ERP, entre outros. Esses sistemas, armazenam registros sobre todos os casos em que algum processo foi executado nessas plataformas, a quantidade de execuções realizadas, as pessoas ou sistemas envolvidos na execução, entre outros detalhes. Esses registros são o ponto de partida da mineração de processos, comumente denominados RE (VAN DER AALST; *et al.*, 2009). Diversos componentes podem compor um RE, entre eles, os seguintes se destacam (VAN DER AALST; *et al.*, 2010):

- Evento/Atividade/Tarefa: se refere a uma etapa bem-definida do processo. Essa etapa é executada inteiramente por um recurso e é registrada de forma sequencial conforme foi ocorrida;
- Caso: uma instância de um trabalho realizado que contém um grupo de atividades. Cada caso possui um identificador único;
- Recurso: se refere a uma ou mais pessoas ou máquinas que executam um evento específico. O recurso pode não ter executado o evento, mas foi utilizado na sua execução;
- Carimbo de data/hora: data/hora em que determinado processo de iniciou. Em alguns registros podem ser expressos também as datas de finalização de cada evento. Quando

não há tempo de conclusão, supõe-se que uma atividade finalizou assim que a próxima começou. Quando há uma diferença entre o fim de uma atividade e o início da outra há um tempo de espera entre as atividades;

– Traço: sequência de eventos que compõem determinado caso. Auxilia a entender quais atividades foram realizadas em um caso até aquele momento.

– Outros detalhes sobre os eventos ou casos sendo executados podem ser adicionados a fim de fornecer mais detalhes para os usuários ou solucionadores. A Tabela 1 apresenta um exemplo de RE.

ID do Caso	Tarefa	Status Evento	Criado por	Início (dd/mm/aaaa hh:mm:ss)	Extra
1	Arquivar Multa	Finalizado	Anne	20/07/2004 14:00:00	...
2	Arquivar Multa	Finalizado	Anne	20/07/2004 15:00:00	...
1	Enviar Recibo	Finalizado	system	20/07/2004 15:05:00	...
2	Enviar Recibo	Finalizado	system	20/07/2004 15:07:00	...
3	Arquivar Multa	Finalizado	Anne	21/07/2004 10:00:00	...
3	Enviar Recibo	Finalizado	system	21/07/2004 14:00:00	...
4	Arquivar Multa	Finalizado	Anne	22/07/2004 11:00:00	...
4	Enviar Recibo	Finalizado	system	22/07/2004 11:10:00	...
1	Processo de Pagamento	Finalizado	system	24/07/2004 15:05:00	...
1	Finalizar Caso	Finalizado	system	24/07/2004 15:06:00	...
2	Enviar Lembrete	Finalizado	Mary	20/08/2004 10:00:00	...
3	Enviar Lembrete	Finalizado	John	21/08/2004 10:00:00	...
2	Processo de Pagamento	Finalizado	system	22/08/2004 09:05:00	...
2	Finalizar Caso	Finalizado	system	22/08/2004 09:06:00	...
4	Enviar Lembrete	Finalizado	John	22/08/2004 15:10:00	...
4	Enviar Lembrete	Finalizado	Mary	22/08/2004 17:10:00	...
4	Processo de Pagamento	Finalizado	system	29/08/2004 14:01:00	...
4	Finalizar Caso	Finalizado	system	29/08/2004 17:30:00	...
3	Enviar Lembrete	Finalizado	John	21/09/2004 10:00:00	...
3	Enviar Lembrete	Finalizado	John	21/10/2004 10:00:00	...
3	Processo de Pagamento	Finalizado	system	25/10/2004 14:00:00	...
3	Finalizar Caso	Finalizado	system	25/10/2004 14:01:00	...

Tabela 1 – Exemplo de um registro de eventos
Fonte: Adaptado de van der Aalst; *et al.*, 2009

MP é uma área de estudos que se situa entre a mineração de dados e a modelagem e análise de processos. A MP tem três principais objetivos: o de descobrir, monitorar e melhorar processos reais através da extração (mineração) de conhecimento de RE disponíveis em sistemas de informação. Dentro dessa área de estudo estão situadas as seguintes técnicas: descobrimento de um modelo de processos a partir de um RE, verificação de conformidade do processo real e do mapeado, aprimoramento de modelos de processo, previsão de casos dos processos, entre outros (VAN DER AALST; *et al.*, 2010). Na Figura 2 estão ilustradas as entradas e saídas das três técnicas mais comuns: descoberta, conformidade e aprimoramento.

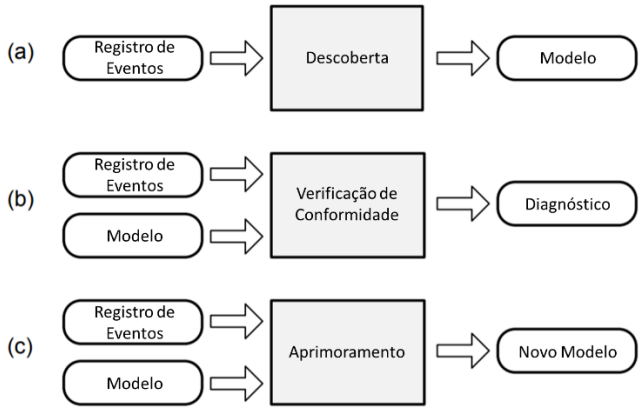

Figura 2 - Entradas e saídas dos três tipos básicos de mineração de processos a) descoberta b) conformidade e c) aprimoramento
Fonte: Adaptado de van der Aalst, *et al*, 2011a

As técnicas de MP de descoberta utilizam de um RE e produzem um modelo do processo que o representa. Este modelo é uma representação que pode ser expressa em diversos formatos como o de um fluxograma ou rede de Petri e reflete o processo que está sendo seguido de acordo com os casos registrados no RE. A descoberta de processos é a técnica de mineração de processos mais proeminente (VAN DER AALST; *et al*, 2011a). A Figura 3 ilustra a rede de Petri de um modelo de processo extraído do RE da Tabela 1.

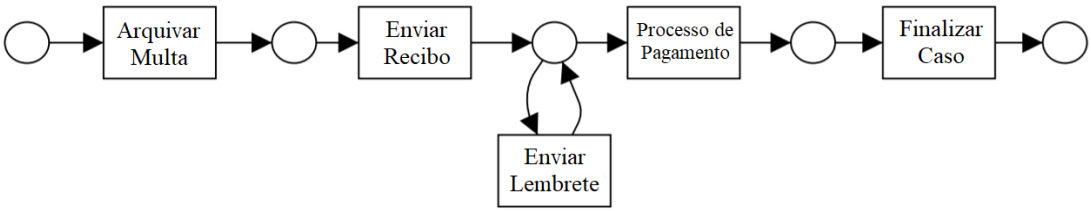

Figura 3 – Exemplo de um registro de eventos
Fonte: Adaptado de van der Aalst; *et al.*, 2009

Já as técnicas de conformidade utilizam de um modelo de processo já existente e o compara com os dados de um RE do mesmo processo. A verificação de conformidade pode ser usada para verificar se a realidade, conforme registrada no RE, está de acordo com o modelo projetado e vice-versa (VAN DER AALST; *et al*, 2011a).

Os processos de aprimoramento buscam ampliar ou melhorar um modelo de processo já existente usando informações adicionais sobre o processo real registrado em um RE. Enquanto a verificação de conformidade mede o alinhamento entre o modelo e a realidade, este terceiro tipo de técnica de MP visa alterar ou estender o modelo (VAN DER AALST; *et al*,

2011a). O presente trabalho explora outra técnica de mineração de processo de grande relevância, à previsão de processos de negócio.

2.3 PREVISÃO DE CASOS EM ANDAMENTO

O monitoramento preditivo de processos também conhecido como previsão de processos de negócios, é outro ramo da mineração de processos que visa prever os alvos predefinidos de interesse usando as métricas das atividades dos casos em execução. O principal conceito subjacente do gerenciamento preditivo de processos de negócios é a antecipação do objetivo de interesse definido pelo usuário com o uso das atividades de processo dos casos em execução (MEHDIYEV; FETTKE, 2021). Alguns parâmetros podem ser previstos usando essas técnicas, entre eles:

- Previsão do próximo evento (PRAVILOVIC; APPICE; MALERBA, 2013)
- Previsão do resultado de um caso em andamento (se uma requisição será aprovada ou negada por exemplo) (DE LEONI; *et al.*, 2016)
- Previsão do tempo remanescente de casos em andamento (VAN DER AALST; *et al*., 2011b) (VAN DONGEN; CROOY; VAN DER AALST, 2008) (LEITNER *et al*, 2009)

No tópico da previsão da duração de casos a partir de um RE, muitos métodos diferentes também foram utilizados. Van Dongen *et al.* (2008) extraíram estimadores do RE e utilizaram da regressão não paramétrica para prever o tempo restante até a conclusão de um caso. Van der Aalst *et al.* (2011b) propuseram anotar um sistema de transição com informações sobre a ocorrência de um evento específico ou o tempo até um determinado evento. Weber *et al.* (2007) descreveram um método de aprendizado preguiçoso (*lazy learning*) para calcular recomendações baseadas em um RE que atuam em uma otimização específica, por exemplo, minimizar o tempo de ciclo ou maximizar o lucro.

Apesar das diversas abordagens, todas utilizam inicialmente da MP para extrair informações relevantes dos RE e assim treinar modelos preditivos. Há uma grande variedade de informações ou medidas que podem ser mineradas do processo e utilizadas nos modelos. O tempo médio de ciclo é frequentemente usado como um estimador para o tempo de ciclo restante. Usando o início e a conclusão de uma sequência é trivial definir o tempo de ciclo médio de um RE, que é apenas a soma das durações de todos os casos dividido pelo número de casos (VAN DONGEN; CROOY; VAN DER AALST, 2008).

RE são populados por eventos sequenciais, onde certos eventos precedem outros. Dessa forma, a qualquer momento, uma instância de processo está em uma atividade que depende das atividades que foram previamente realizadas. As atividades que antecederam a atual em um caso são denominadas de prefixo do caso (VAN DER AALST; *et al*., 2011b). Dessa forma, uma medição de ocorrência de atividade é tal que para cada prefixo (não vazio) de um caso, uma medição é feita, consistindo no número de ocorrências que um evento específico ocorreu naquele caso (VAN DONGEN; CROOY; VAN DER AALST, 2008).

O resultado da extração dessa medida a partir do RE pode ser representado por uma tabela. Cada linha dessa tabela representa um momento em que um caso se encontra, e as colunas o número de vezes que aquele caso realizou cada atividade. Dessa forma, a Tabela 2 exemplifica coletas da medição de ocorrência de atividade a partir do RE apresentado na Tabela 1.

ID do Caso	Arquivar Multa	Enviar Recibo	Enviar Lembrete	Processo de Pagamento	Finalizar Caso
1	1	1	0	1	1
2	1	1	1	1	1
3	1	1	3	1	1
4	1	1	2	1	1

Tabela 2 – Exemplo da medida de ocorrências das atividades para cada caso
Fonte: O Autor

Outra medida que pode ser utilizada como estimador considera a duração em que determinado caso passou em cada atividade já realizada. Dessa forma, para cada prefixo (não vazio) de um caso, é medido o tempo de duração médio que o caso passou naquela atividade. Variações dessa medida, como a soma dos tempos de duração daquele caso em cada atividade também podem ser explorados. Dessa maneira, o domínio obtido pelos parâmetros não são números naturais, como na apresentada acima, mas reais não negativos (VAN DONGEN; CROOY; VAN DER AALST, 2008).

De forma análoga à medida de quantidade de eventos, é possível montar uma tabela com os estimadores extraídos. Nesta tabela, cada linha contém um momento de um caso, e cada coluna o tempo médio ou a soma dos tempos decorridos naquela atividade. A Tabela 3 exemplifica coletas da medida de duração média (em dias) das atividades a partir do RE apresentado na Tabela 1.

ID do Caso	Arquivar Multa	Enviar Recibo	Enviar Lembrete	Processo de Pagamento	Finalizar Caso
1	0,045	4,000	0	0,001	0
2	0,005	30,787	1,962	0,001	0
3	0,167	30,833	21,722	0,001	0
4	0,007	31,167	3,476	0,145	0

Tabela 3 – Exemplo da medida de duração média (em dias) das atividades para cada caso
Fonte: O Autor

Ademais, outras medidas podem ser mineradas a partir do RE. Medidas classificatórias, como os recursos utilizados, colaboradores que participaram ou local onde foi realizado cada atividade também podem ser utilizados como parâmetros nos modelos. Medidas referentes à hora do dia, ou época do ano que determinado caso foi realizado também podem fornecer informações relevantes para uma previsão. Para isso, deve ser levado em conta nos modelos de previsão, qual o tipo de dado de entrada que está sendo utilizado para prever a duração remanescente.

2.4 REGRESSÃO NÃO-PARAMÉTRICA (NPR)

Regressão é uma técnica utilizada para ajustar uma função à um grupo de medidas. Há dois tipos de regressão mais comuns, nomeadamente paramétrica, onde a função é assumida como tendo uma certa forma (por exemplo linear, exponencial, quadrática etc.) e não paramétrica, onde nenhuma suposição é feita sobre a função ajustada às medidas (VAN DONGEN; CROOY; VAN DER AALST, 2008).

Regressão paramétrica consiste em adequar uma função conhecida inicialmente (a priori) à pontos de dados. Na regressão paramétrica, a parte que não pode ser explicada pela função é considerada ruído, isto é, um erro aleatoriamente distribuído (FAN; GIJBELS, 2018).

Para aplicações de comportamentos conhecidos, a escolha de uma função que se adeque satisfatoriamente aos dados apresentados é simples, no entanto, esse não é sempre o caso. Na figura 4, são apresentadas quatro regressões paramétricas diferentes aplicadas à dados de uma simulação de impacto. Neste exemplo, pode ser observado de forma empírica, que dependendo da equação escolhida a priori, a regressão pode não ser uma boa representação das medidas plotadas (FAN; GIJBELS, 2018).

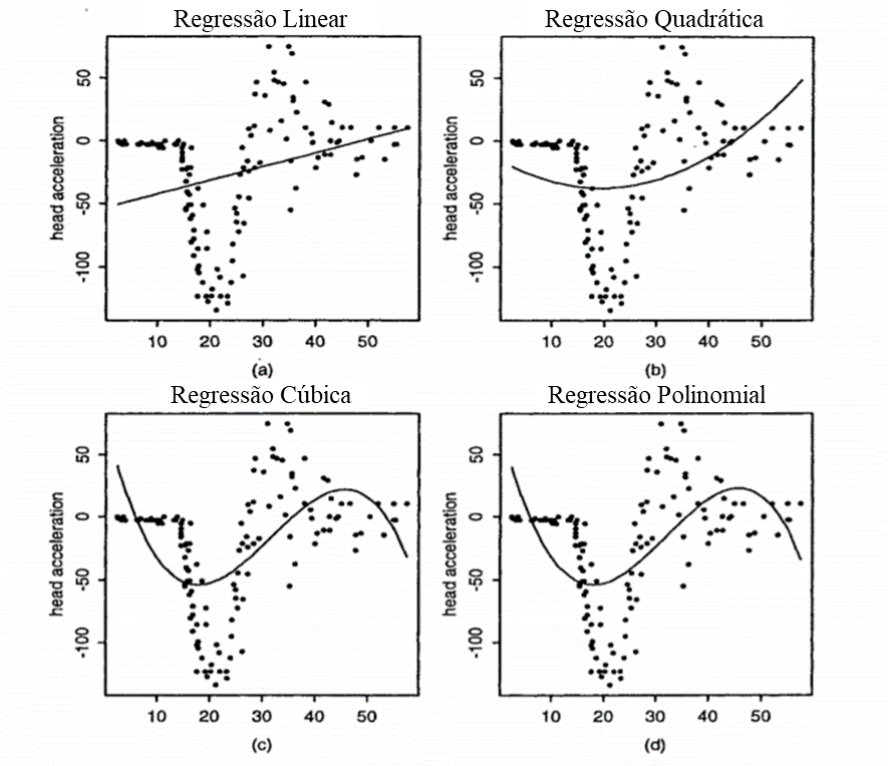

Figura 4 - Comparação entre diferentes regressões paramétricas
Fonte: Adaptado de Fan, Gijbels, 2018

Dessa forma, quando o modelo representativo de determinados dados é desconhecido, a NPR, isto é, a regressão que não considera nenhum conhecimento a priori, pode ser uma opção mais adequada. Na Figura 5, são comparadas duas regressões: paramétrica linear (na esquerda) e uma regressão não paramétrica (na direita).

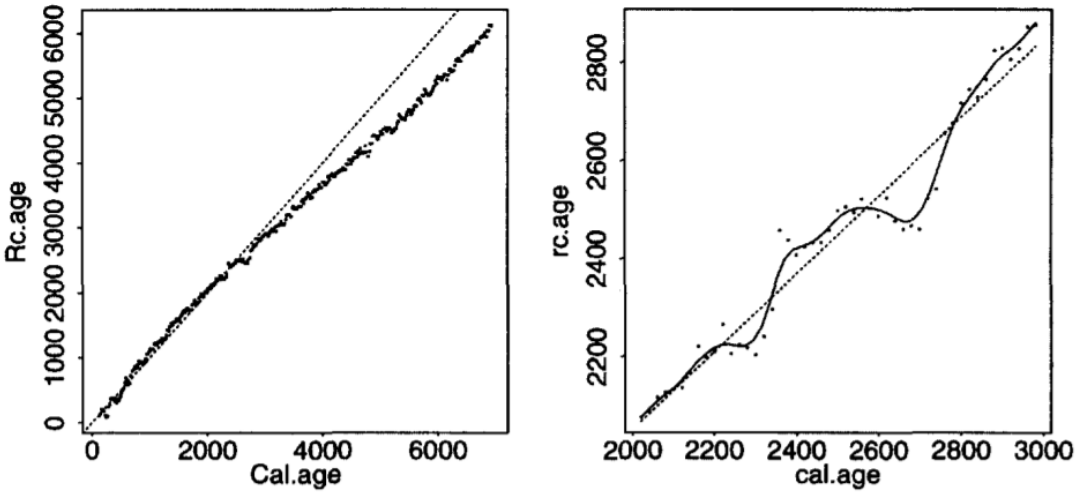

Figura 5 – Comparação entre uma regressão linear e uma regressão não paramétrica
Fonte: Bowman, A.W, Azzalini, 1997

Em outras palavras, as técnicas de NPR realizam uma regressão que considera alguns/todos os pontos plotados para dar maior flexibilidade ao seu formato (WITTEK, 2014). Das técnicas de NPR, a regressão de Kernel é uma das mais utilizadas dado que esta fornece uma vasta seleção de estimadores podendo ser usadas em aplicações diversas (FAN; GIJBELS, 2018).

Dada uma função de regressão que não possui uma forma específica, a regressão de Kernel estima para cada ponto de teste, um valor de resposta, através de uma média ponderada dos pontos ao redor do ponto de teste. O peso atribuído a cada ponto é determinado por uma função denominada de função de Kernel e os pontos sendo considerados para a média são dependentes do parâmetro de suavização (largura de banda) utilizado (FAN; GIJBELS, 2018).

Dessa forma, dada uma função de Kernel qualquer, o parâmetro de suavização *h* controla a janela de dados em que são atribuídos pesos pela função Kernel e, consequentemente, representa o grau de suavização da regressão. As regiões sombreadas na Figura 6 são representações gráficas da janela em que a função de Kernel coloca um peso apreciável na construção de uma média local. Para parâmetros de suavização mais altos, o estimador resultante perde alguns detalhes da curvatura dos dados, como pode ser observado no gráfico da esquerda. À medida que o parâmetro de suavização diminui, o estimador começa a rastrear os dados até que acaba interpolando os pontos observados. Dessa forma, a largura da banda governa a complexidade do modelo, isto é, se a função é mais ou menos influenciada por cada ponto específico (BOWMAN; A.W; AZZALINI, 1997).

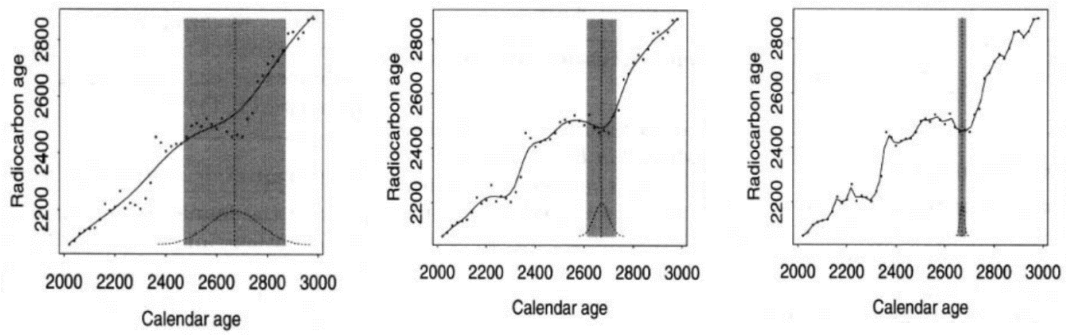

Figura 6 – Regressões não paramétricas considerando parâmetros de suavização diferentes
Fonte: Adaptado de Bowman, A.W, Azzalini, 1997

Para a atribuição de pesos pela função de Kernel, pontos mais distantes de determinado ponto de teste, devem impactar menos na estimativa de um valor de resposta. Dessa forma, podem ser atribuídos aos pontos dentro da largura de banda, pesos que representam o quão importante aquele ponto é para a estimativa da regressão (FAN; GIJBELS, 2018).

Há diversas funções ou estimadores, que podem ser utilizados na atribuição desses pesos, na Tabela 4 são apresentadas algumas funções de Kernel. As funções de Kernel $\mathrm{K}(x, x_i)$ possuem os parâmetros: ponto de teste x, demais pontos x_i, parâmetro de suavização h e a distância entre os pontos $\mathrm{d}(x, x_i)$ (IZBICKI; DOS SANTOS, 2020).

Nome	**Função**
Uniforme	$\mathrm{K}(x, x_i) = 1 \; \forall \; (\mathrm{d}(x, x_i) \leq h)$
Gaussiano	$\mathrm{K}(x, x_i) = \left(\sqrt{2\pi}\right)^{-1} * \exp\left\{-\frac{d^2(x, x_i)}{2h^2}\right\}$
Epanechnikov	$\mathrm{K}(x, x_i) = 0{,}75 * \left(1 - \frac{\mathrm{d}^2(x, x_i)}{h^2}\right) \forall (\mathrm{d}(x, x_i) \leq h)$

Tabela 4 – Funções de Kernel
Fonte: Adaptado de Demir, Toktamis (2010)

O Kernel uniforme atribui o mesmo peso para todas as observações a uma distância menor que h, isto é, todos os pontos dentro da largura de banda recebem o mesmo peso para a média ponderada. Já o Kernel de Epanechnikov atribui pesos diferentes de acordo com a distância até x_i, observações mais próximas recebem um peso maior. O Kernel gaussiano é a única função apresentada que atribui um peso a todos os pontos observados, no entanto, dado sua formulação, os pontos muitos distantes recebem um peso muito baixo, o que mantém em parte o efeito do parâmetro de suavização (IZBICKI; DOS SANTOS, 2020). As funções de Kernel Gaussiana e de Epanechnikov, costumam ser boas estimadoras e são as mais utilizadas na prática (DEMIR; TOKTAMIS, 2010).

Dessa forma, a regressão de Kernel pode ser utilizada como um modelo preditivo, em que diversas variáveis x são fornecidas e o modelo faz uma previsão de quanto deve ser a variável alvo. Utilizando dados históricos, é possível extrair as variáveis de entrada, prever a variável alvo e compará-la com os valores reais registrados. Assim, pode ser realizada a avaliação das regressões ajustadas em busca do melhor modelo.

No capítulo 3, são abordados como o modelo preditivo se conecta com as técnicas de mineração de processos, e os critérios de avaliação dos modelos preditivos são apresentados.

2.5 NORMALIZAÇÃO DOS DADOS

A normalização é uma técnica que pode ser aplicada como parte da preparação de dados para modelos preditivos. O objetivo da normalização é alterar os valores numéricos em conjunto de dados para usar uma escala comum, sem distorcer as diferenças nos intervalos de valores ou perder informações. A normalização também é necessária em alguns algoritmos para modelar os dados corretamente (BATISTA, 2003).

Por exemplo, se um conjunto de dados contém uma coluna com valores que variam de 0 a 1 e outra coluna com valores que variam de 10.000 a 100.000. A grande diferença na escala dos números pode causar uma desvalorização das escalas menores ao combinar os valores como medidas durante a modelagem. A normalização evita esses problemas, alterando os valores de forma a manter a distribuição geral e as proporções nos dados de origem, aplicando uma escala em todas as colunas numéricas usadas no modelo (BATISTA, 2003).

Dessa forma, a normalização é um processo em que a unidade de medida dos dados é desconsiderada. Existem dois tipos mais comuns de normalização. O primeiro tipo, se origina da álgebra linear e trata os dados como vetores em um espaço multidimensional. Nesse contexto, normalizar os dados é transformar o vetor de dados em um novo vetor cuja norma (comprimento do vetor) é igual a um. O segundo tipo de normalização se origina da estatística e elimina a unidade de medida, transformando os dados em novos valores com uma média de 0 e um desvio padrão de 1. Essa transformação é conhecida como normalização Z-score (ABDI; WILLIAMS, 2010).

2.6 DADOS DE TREINO E TESTE

No processo de regressão são utilizados dados históricos para a criação da curva de regressão que então pode ser utilizada para prever a variável alvo quando esta ainda não é conhecida. Para validar se o modelo consegue representar bem o processo uma parte das variáveis do histórico são utilizadas somente para testar o modelo que foi criado. Dessa forma, os dados tendem a ser divididos em dois grupos um de treino do modelo de regressão e o outro para testar o modelo após sua criação (IZBICKI; DOS SANTOS, 2020).

Uma boa prática para definir quais amostras serão utilizadas para compor o conjunto de treino e o conjunto de teste é definir aleatoriamente. Dessa forma, um gerador de números aleatório pode ser utilizado a fim de se definir essa divisão. Esse procedimento evita o problema

da função regredida ser criada sobre um subgrupo específico de dados, e portanto, não representar bem o procedimento como um todo (IZBICKI; DOS SANTOS, 2020).

Outro parâmetro analisado é o percentual de dados que deve ser utilizado como dados de treino e como dados de teste. Alguns valores de referência são divisões 70/30, 80/20 e 90/10, onde o primeiro valor é o percentual de dados utilizados no treino e o segundo número o percentual de dados utilizados no teste do modelo. Para uma otimização completa do modelo, diversos valores podem ser testados e avaliados até que a divisão ótima seja encontrada (NGUYEN; *et al.*, 2021).

2.7 SUMÁRIO DO CAPÍTULO

O SLA é um acordo entre o prestador e o consumidor de um serviço que define as expectativas e responsabilidades de ambas as partes. Esse serviço realizado pode ser descrito por um processo e armazenado em um RE. As análises realizadas sobre um RE pertencem a área de conhecimento da mineração de processos (MP), em que uma das suas áreas é a previsão do tempo de conclusão dos processos, o que é útil para a gestão proativa do SLA. Dessa forma, é possível extrair métricas relevantes do RE que podem ser aplicadas a modelos preditivos como o NPR para que o tempo remanescente de casos em andamento sejam previstos.

3. METODOLOGIA DE IMPLEMENTAÇÃO DOS MODELOS

Este capítulo aborda a metodologia utilizada na implementação das técnicas de MP e dos modelos preditivos. A mesma metodologia foi utilizada na execução do estudo de caso também abordado no capítulo 4 desse trabalho. A metodologia apresentada na Figura 6 pode ser replicada para diversas aplicações e foi adaptada do estudo de Van Dongen *et al.* (2008). Dessa forma, esta metodologia pode ser generalizada para qualquer processo que possua um RE.

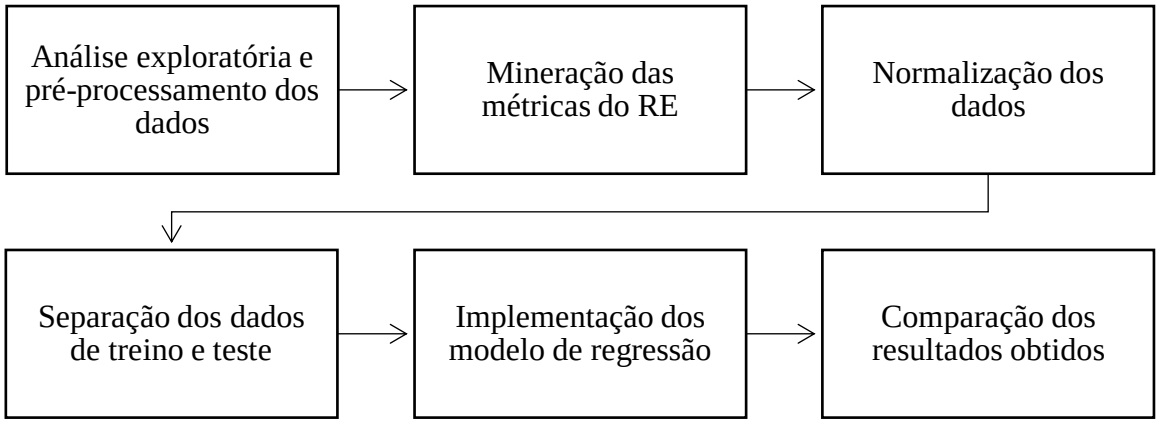

Figura 6 – Metodologia utilizada no estudo de caso
Fonte: Adaptada de Van Dongen *et al.* (2008)

3.1 ANÁLISE EXPLORATÓRIA E PRÉ-PROCESSAMENTO DOS DADOS

A etapa de análise exploratória e pré-processamento teve como objetivo entender o processo que os chamados passam e de realizar as transformações iniciais necessárias nos dados para a mineração nas etapas subsequentes. Os dados de entrada dessa etapa devem ser um registro de eventos (RE) contendo as atividades realizadas ao longo de diversos casos diferentes.

A análise iniciou com uma decomposição de todos os casos e de suas atividades conforme suas datas de abertura e fechamento. Assim é possível observar os tempos despendidos em cada atividade dos diversos casos analisados. Para isso foi utilizada a extensão *Project Log on Dotted Chart* ferramenta do ProM.

Para entender o processo que dá origem ao RE, foi minerado o fluxograma que descreve o processo através da utilização da extensão *Mine With Visual Miner* da ferramenta ProM.

Para analisar em mais detalhes cada uma das atividades, foram plotados histogramas do tempo decorrido para cada *status* e suas respectivas médias e desvios padrão. Por fim, foram extraídas de cada caso, a duração em horas úteis do processo e plotadas em histogramas por *status*. Para isso foram utilizadas as bibliotecas *Pandas*, *Matplotlib* e *Seaborn* da linguagem *Python*.

3.2 MINERAÇÃO DAS MÉTRICAS DO RE

Com os dados preparados, para que seja realizada a implementação dos modelos de regressão, primeiro foi preciso preparar as métricas que serviram como dados de entrada no modelo. Para isso, foram selecionadas as métricas propostas pelo trabalho de Van Dongen *et al.* (2008): o tempo decorrido até o momento no chamado, a quantidade de vezes que um chamado passou por cada atividade e a soma do tempo despendido em cada atividade. Esta extração foi realizada através de um código na linguagem *Python* criado pelo autor, utilizando as bibliotecas *Pandas* e *Numpy*.

3.3 NORMALIZAÇÃO DOS DADOS

A etapa de normalização teve como objetivo colocar todas as métricas utilizadas em uma escala única. Para isso todas as métricas foram colocadas em uma escala centrada na média e com as métricas representadas em unidades de variância. Essa implementação foi realizada através da função *scale* da biblioteca *Scikit-Learn* em linguagem *Python*.

3.4 SEPARAÇÃO DOS DADOS DE TREINO E TESTE

Esta etapa tem como objetivo separar as métricas mineradas em dois grupos, os dados de treino e teste, com o intuito de obter uma regressão que melhor represente dados ainda não introduzidos pelo modelo. Para isso, cada caso e suas métricas foram separados aleatoriamente em dois grupos. O maior grupo, os dados de treino, conteve 70% dos casos e foram utilizados no ajuste dos modelos de regressão. Os dados de teste, representaram 30% dos dados totais. Nestes, os modelos criados foram implementados e se obteve previsões do tempo remanescente de cada momento de diversos chamados.

3.5 IMPLEMENTAÇÃO DO MODELO DE REGRESSÃO

A etapa de implementação do modelo de regressão conteve duas principais partes, o ajuste dos modelos de regressão aos dados de treino e a avaliação dos modelos obtidos com os dados de teste. O objetivo dessa etapa foi de implementar diversas regressões não-paramétricas de Kernel com variados hiper parâmetros, a fim de achar o modelo que melhor conseguisse prever o tempo remanescente de casos em andamento.

Para obter uma linha de base ou um ponto de referência para os resultados, foi desenvolvido um modelo de regressão simples, utilizando apenas a média do tempo remanescente por atividade para realizar as previsões. Para cada caso, foi considerado somente a atividade sendo realizada naquele momento e o tempo remanescente daquele caso até a sua conclusão. Foi então realizada a média dos tempos remanescentes por atividade, obtendo previsões simples do tempo remanescente para quando os casos se encontram em cada uma das atividades. Essas médias foram utilizadas para realizar a previsão do tempo remanescente de todos os dados de teste e foram calculadas as diferenças absolutas entre os valores reais e os valores previstos encontrados.

Ademais, foram ajustados modelos de regressão de Kernel com duas funções de Kernel a de Epanechnikov e a Gaussiana. Para encontrar o parâmetro de suavização que gera a regressão que melhor se encaixa nos dados de treino, foi realizado uma busca heurística, onde o parâmetro foi variado entre 0,1 e 50 em iterações de 0,1 para cada uma das duas funções. No total, foram geradas 1.000 combinações de regressões diferentes. Cada regressão ajustada foi utilizada para obter previsões dos tempos remanescente dos dados de teste e para cada uma das previsões foi calculada a diferença absoluta do valor previsto para o valor real, que serviu como parâmetro de avaliação dos melhores modelos.

A implementação da regressão de Kernel foi realizada através da linguagem de programação *Python*. Para a regressão de Kernel Gaussiana, foi utilizada uma classe desenvolvida por Mehta (2020). A mesma classe foi adaptada pelo autor e foi utilizada para a regressão de Kernel de Epanechnikov. As duas regressões são apresentadas na Figura 7.

```python
import numpy as np

'''Class for Gaussian Kernel Regression'''
class GKR:

    def __init__(self, x, y, b):
        self.x = np.array(x)
        self.y = np.array(y)
        self.b = b

    '''Implement the Gaussian Kernel'''
    def gaussian_kernel(self, z):
        return (1/np.sqrt(2*np.pi))*np.exp(-0.5*z**2)

    '''Calculate weights and return prediction'''
    def predict(self, X):
        kernels = np.array([self.gaussian_kernel((np.linalg.norm(xi-X))/self.b) for xi in self.x])
        weights = np.array([len(self.x) * (kernel/np.sum(kernels)) for kernel in kernels])
        return np.dot(weights.T, self.y)/len(self.x)

'''Class for Epanechnikov Kernel Regression'''
class EKR:

    def __init__(self, x, y, b):
        self.x = np.array(x)
        self.y = np.array(y)
        self.b = b

    '''Implement the Epanechnikov Kernel'''
    def epanechnikov_kernel(self, z):
        r = (0.75*(1-np.power(z,2)))
        if abs(z) > 1: r = 0
        return r

    '''Calculate weights and return prediction'''
    def predict(self, X):
        kernels = np.array([self.epanechnikov_kernel((np.linalg.norm(xi-X))/self.b) for xi in self.x])
        weights = np.array([len(self.x) * (kernel/np.sum(kernels)) for kernel in kernels])
        return np.dot(weights.T, self.y)/len(self.x)
```

Figura 7 – Classe de funções de regressão de Kernel Gaussiana e de Epanechnikov
Fonte: Adaptado de Mehta, 2020

3.6 COMPARAÇÃO DOS RESULTADOS OBTIDOS

O erro absoluto médio (MAE), é um valor obtido realizando a média das diferenças entre os valores previstos pela regressão e o valor real do tempo remanescente dos casos dos dados de teste. Nesta etapa, foram calculados o MAE's para cada uma das regressões ajustadas e os valores encontrados foram comparados.

A fim de analisar os efeitos que cada função de Kernel e o uso de diferentes métricas tem sobre o resultado, forem selecionadas as regressões com o menor MAE entre as diferentes combinações realizadas. Dessa forma, das 1.000 regressões que foram ajustadas, foram selecionadas 8 regressões, combinando as 4 métricas de entrada com as 2 regressões de Kernel testadas.

Para conseguir comparar e entender melhor as características de cada modelo, o MAE foi explorado de duas formas. Na primeira, os dados de teste foram agrupados conforme o *status* em que se encontravam e MAE foi calculado por *status*. Na segunda, foi calculado o MAE por *status* para o modelo de referência, e os outros modelos foram representados como percentuais do modelo de referência, assim facilitando uma análise comparativa entre os modelos.

Ademais, os modelos com a maior precisão também foram comparados quanto a sua capacidade de auxiliar na previsão de violações do SLA, através do percentual de acerto na classificação de chamados violados.

Por fim, foram realizadas as conclusões sobre os resultados de cada modelo e realizadas as recomendações sobre os usos de cada modelo e função de Kernel.

3.7 SUMÁRIO DO CAPÍTULO

A metodologia da implementação dos modelos de MP e das regressões foi planejada em seis partes com base no trabalho de Van Dongen *et al.* (2008). Inicialmente é realizada uma análise exploratória e pré-processamento do RE para entender o processo que está sendo seguido e preparar os dados para as etapas subsequentes. São então mineradas e normalizadas as métricas que servem como dados de entrada nos modelos de regressão. Ademais, as métricas são separadas em dois grupos, os dados de treino e teste. Diversos modelos de regressão são ajustados às métricas, a fim de encontrar o modelo que produza o menor erro absoluto para os dados de teste. Por fim, todos os resultados obtidos são comparados e analisados.

4. RESULTADOS OBTIDOS NO ESTUDO DE CASO

Nesse capítulo são apresentados os dados de entrada utilizados no estudo de caso e os resultados obtidos através da implementação da metodologia apresentada no Capítulo 3.

4.1 ANÁLISE EXPLORATÓRIA E PRÉ-PROCESSAMENTO DOS DADOS

Para o presente trabalho foram utilizados dados de um registro de eventos (RE) da equipe de TI de uma empresa de grande porte, contendo informações sobre um chamado de compra de equipamento de TI como *notebooks* e *smartphones*. Ao precisar de um desses equipamentos, representantes de unidades de todo o Brasil pode acessar a central de serviços (CS) e abrir um chamado para serem atendidos pelo time de TI responsável. Estes chamados possuem um acordo de nível de serviço (SLA) de 108 horas úteis, portanto, precisam ser atendidos nesse período ou os solucionadores sofrem penalidades em suas metas para o ano.

O processo realizado na oferta de compra de equipamentos da CS fica armazenado em um RE contendo as seguintes informações:

- ID do chamado: identificador único que cada chamado de TI possui, representa cada caso de realização do processo;
- Oferta do chamado: nome do serviço que está sendo prestado nesse chamado. Para este estudo, estão sendo utilizados somente dados da oferta de compra de equipamentos de TI;
- *Status* do chamado: em qual parte do processo o chamado se encontra. Representa cada um dos eventos/atividades possíveis de serem realizadas em um chamado. Esse campo pode receber os seguintes valores:
    - *Aberto*: o primeiro *status* que todo chamado recebe. Significa que foi recebido pelo solucionador de TI, mas este ainda não começou a atuar na resolução do chamado;
    - *Em Tratamento*: ocorre enquanto o grupo de solucionadores está trabalhando no chamado para atender à solicitação realizada;
    - *Pendente Usuário*: ocorre quando há alguma informação que o time de TI precisa do usuário. Quando o usuário responder o que foi solicitado o chamado sai desse *status*;

- *Em Validação*: *status* em que a solução proposta pelo time de TI é validada pelo usuário requisitante. Caso o usuário não valide a solução proposta o chamado é aceito de forma tácita;
- *Solucionado*: ocorre quando a solução proposta é aceita e o chamado é finalizado.

- Hora de início: data e hora em que o chamado entrou naquele *status*;
- Hora de fim: data e hora em que o chamado saiu daquele *status*.

O processo sendo analisado sofreu mudanças em junho de 2021. Por isso estão sendo usados casos de 75 chamados diferentes, com o primeiro sendo iniciado em 25/06/2021 e o último finalizado dia 12/11/2021. A Tabela 5 apresenta o exemplo de um dos 75 chamados, conforme suas trocas de *status* foram registradas no RE.

ID	**Oferta**	***Status***	**Início**	**Fim**
R4906956	Compra de Equipamento	*Aberto*	01/07/2021 18:06	06/07/2021 13:14
R4906956	Compra de Equipamento	*Em Tratamento*	06/07/2021 13:14	06/07/2021 18:08
R4906956	Compra de Equipamento	*Pendente Usuário*	06/07/2021 18:08	07/07/2021 16:01
R4906956	Compra de Equipamento	*Em Tratamento*	07/07/2021 16:01	08/07/2021 12:18
R4906956	Compra de Equipamento	*Pendente Usuário*	08/07/2021 12:18	13/07/2021 10:42
R4906956	Compra de Equipamento	*Em Tratamento*	13/07/2021 10:42	20/07/2021 16:31
R4906956	Compra de Equipamento	*Em Validação*	20/07/2021 16:31	27/07/2021 17:31
R4906956	Compra de Equipamento	*Solucionado*	27/07/2021 17:31	27/07/2021 17:31

Tabela 5 – Chamado do RE utilizado no estudo de caso
Fonte: O Autor

Com o objetivo de entender a distribuição que os chamados analisados tiveram ao longo do tempo, foi criado um gráfico mostrando os momentos em que os chamados trocaram de *status*, apresentado na Figura 8. Neste gráfico fica perceptível que os chamados se iniciam de maneira uniforme ao longo de todo o período, sem grandes concentrações em determinado momento.

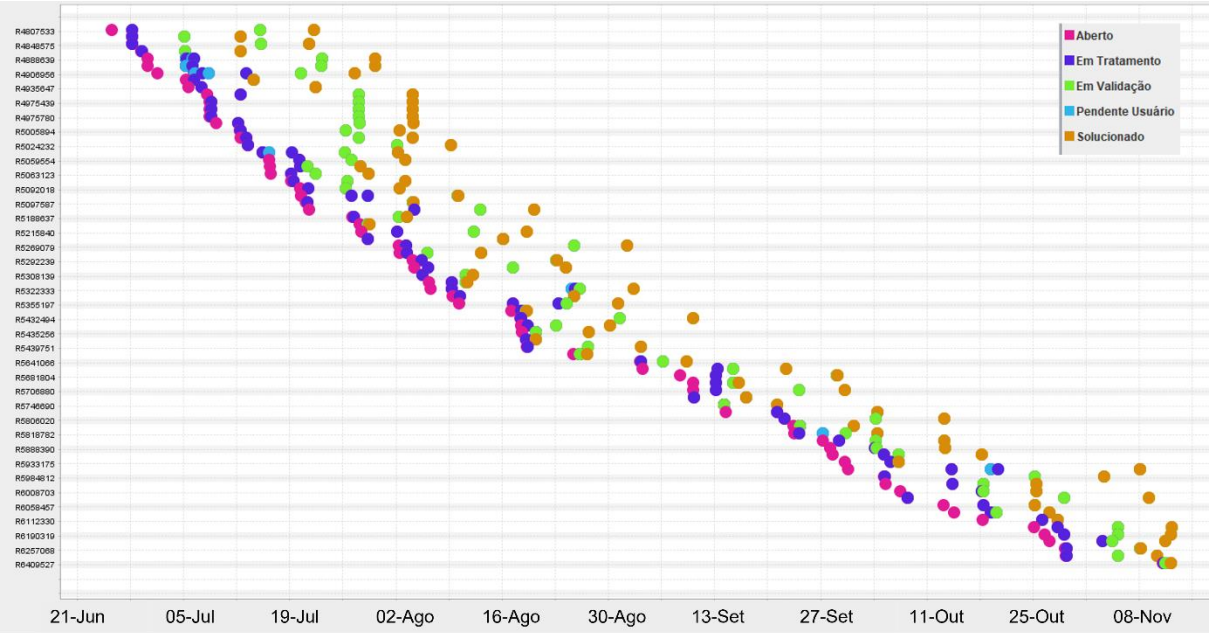

Figura 8 –Atividades dos chamados ao longo do ano
Fonte: O Autor

A Figura 9 apresenta um resultado similar ao apresentado na Figura 8. No entanto, o eixo-x desse gráfico contém o tempo decorrido desde o início de cada chamado, ordenado do chamado mais curto ao mais longo em duração. O chamado mais curto do período analisado durou 23,37 horas e o mais longo durou 38 dias e 14,97 horas.

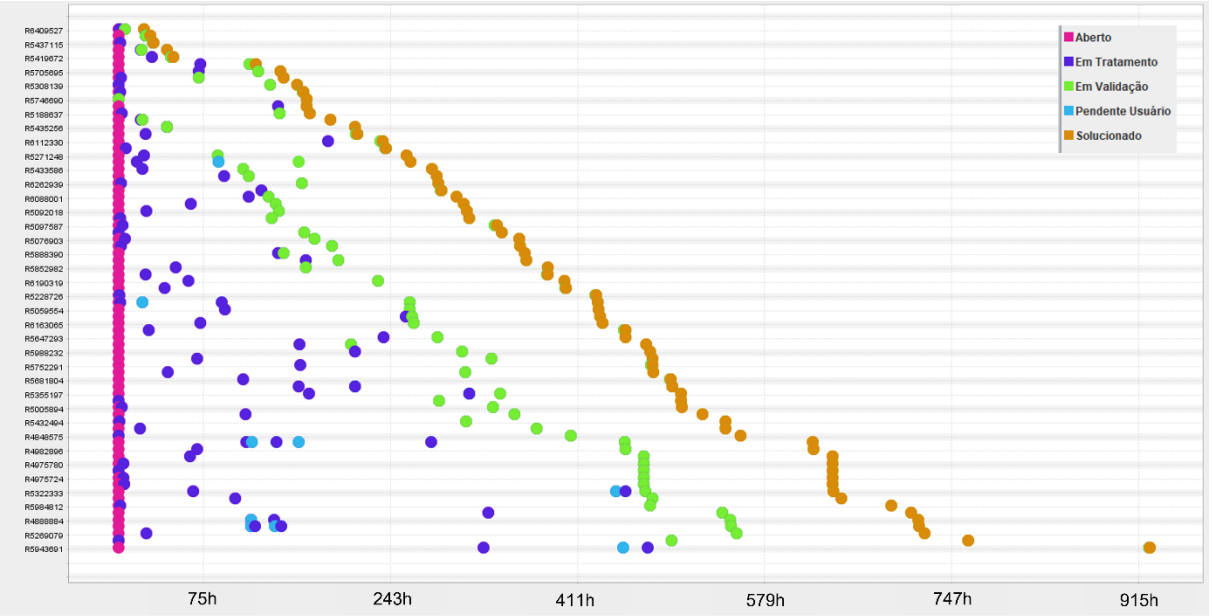

Figura 9 – Atividades dos chamados ao longo de sua duração
Fonte: O Autor

Dessa forma, é possível perceber que a duração da resolução de um chamado varia bastante nos diversos casos. Visualmente pode ser observado que o chamado de duração mais

extensa não segue a mesma linha de tendência que os demais chamados, indicativo de um possível *outlier* do processo. Ademais, pode ser observado que na maioria dos chamados, o *status Em Validação* teve duração bem consistente para os diferentes chamados, o que pode ser atribuído ao tempo máximo que um chamado pode permanecer nesse *status* antes de ser solucionado tacitamente pelo sistema.

Para entender em mais detalhes o processo sendo seguido pelo RE foi elaborado um fluxograma do processo, apresentado na Figura 10. Nesta figura pode-se observar que todos os casos passam pelos *status Aberto*, *Em Validação* e *Solucionado*. A principal alteração sofrida por alguns dos casos foi um *loop* que ocorreu entre os *status Em Tratamento* e *Pendente Usuário*, o que ocorre quando um solucionador está tratando o chamado mas precisa de alguma informação ou evidência do usuário.

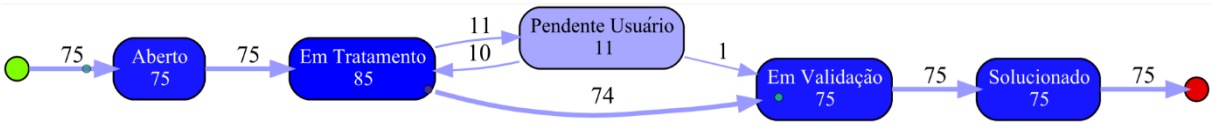

Figura 10 – Fluxograma do processo minerado do RE
Fonte: O Autor

A falta de processos repetitivos, a quantidade pequena de atividades e a linearidade do processo, que sempre começa e termina nos mesmos *status*, indicam uma simplicidade no processo. Dessa forma, a informação das atividades já realizadas em determinado chamado serão bem parecidas de chamado para chamado.

Para que cada *status* fosse analisado com mais detalhes foram criados histogramas com a duração de cada chamado naquele *status*, além de suas respectivas médias e desvios padrão, conforme a Figura 11.

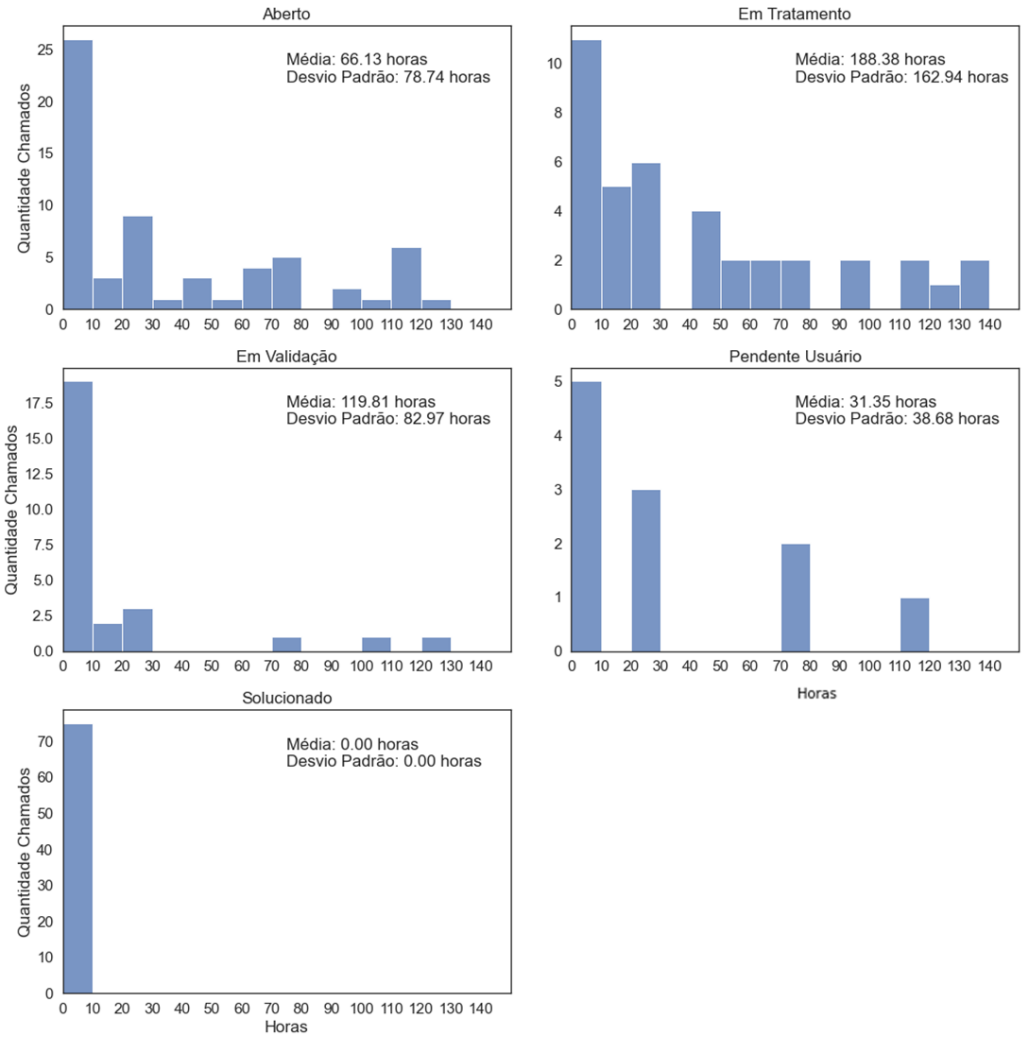

Figura 11 – Histogramas da duração dos chamados por *status*
Fonte: O Autor

As durações dos chamados em cada *status* possuem uma variabilidade alta para os casos analisados. Esse efeito ocorre pois os chamados são tratados pelos times da TI das 9:00 às 18:00 nos dias de semana que não são feriados nacionais. Dessa forma, um chamado que foi aberto em uma sexta-feira fica mais horas em aberto do que um aberto na segunda-feira. Ademais, o cálculo do SLA também não considera as horas em que o time de TI não está em atendimento.

Dado o objetivo do estudo de prever o tempo remanescente de chamados em andamento, faz mais sentido a previsão do tempo remanescente em horas úteis. Para isso, na Figura 12 foram apresentados os mesmos histogramas que os da Figura 11 mas com a duração em horas úteis.

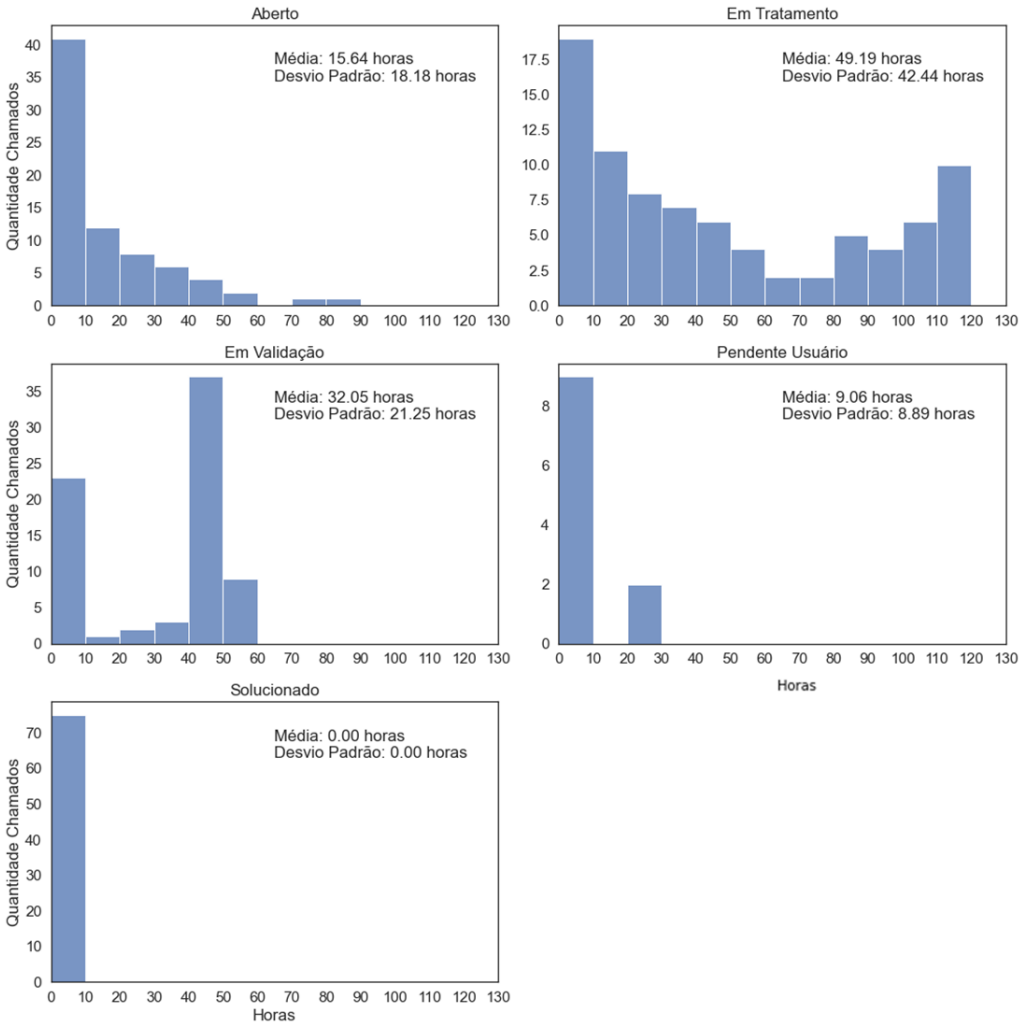

Figura 12 – Histogramas da duração em horas úteis dos chamados por *status*
Fonte: O Autor

Como pode ser observado nesta figura, o *status* que mais varia é o *Em Tratamento*, com desvio padrão de 42,44 horas de duração. O *status* que menos varia é o *status Solucionado*, uma vez que ao entrar nesse *status* o chamado conclui no mesmo segundo. Dado que todos os chamados passaram pelo *status Solucionado* e em todos o tempo despendido foi de 0 horas, este não fornece nenhuma informação relevante sobre o tempo remanescente do chamado, que é concluído imediatamente ao ser solucionado.

4.2 MINERAÇÃO DAS MÉTRICAS DO RE

A etapa de mineração das métricas do RE extrai do RE informações que são utilizadas nos modelos de regressão para a previsão do tempo remanescente até a conclusão de um

chamado. Foram minerados do registro de eventos as seguintes métricas, conforme o estudo realizado por Van Dongen *et al.* (2008):

- Tempo decorrido: o tempo decorrido até aquele momento no chamado;
- Ocorrência das atividades: o número de vezes que cada chamado passou por determinado *status*;
- Duração das atividades: a soma do tempo dispendido em cada *status* que o chamado passou.

Para o estudo, também foi considerado um conjunto das demais métricas, que juntou as três métricas apresentadas em uma tabela única para ser usada como dados de entrada do modelo de regressão. Dessa forma, o resultado obtido nessa etapa foram 4 tabelas em que cada linha apresenta as métricas quando os chamados entraram em cada atividade. Como exemplo, as Tabelas 6, 7, 8 e 9 apresentam as métricas mineradas a partir do chamado de ID R4906956, o mesmo utilizado na Tabela 5. Ou seja, são apresentadas na Tabela 6 a métrica de tempo decorrido, na Tabela 7 a métrica de ocorrência de atividades, na Tabela 8 a métrica de duração de cada atividade e na Tabela 9 o conjunto de todas as métricas.

ID	*Status*	Tempo Decorrido	Tempo Remanescente
R4906956	Aberto	0,00	161,52
R4906956	Em Tratamento	22,23	139,28
R4906956	Pendente Usuário	27,00	134,52
R4906956	Em Tratamento	34,02	127,50
R4906956	Pendente Usuário	39,30	122,22
R4906956	Em Tratamento	64,70	96,82
R4906956	Em Validação	115,52	46,00
R4906956	Solucionado	161,52	0

Tabela 6 – Métrica do tempo decorrido (ID R4906956)
Fonte: O Autor

ID	*Status*	Aberto	Em Tratamento	Pendente Usuário	Em Validação	Solucionado	Tempo Remanescente
R4906956	Aberto	1	0	0	0	0	161,52
R4906956	Em Tratamento	1	1	0	0	0	139,28
R4906956	Pendente Usuário	1	1	1	0	0	134,52
R4906956	Em Tratamento	1	2	1	0	0	127,50
R4906956	Pendente Usuário	1	2	2	0	0	122,22
R4906956	Em Tratamento	1	3	2	0	0	96,82
R4906956	Em Validação	1	3	2	1	0	46,00
R4906956	Solucionado	1	3	2	1	1	0

Tabela 7 – Métrica de ocorrência de atividades (ID R4906956)

Fonte: O Autor

ID	Status	Aberto	Em Tratamento	Pendente Usuário	Em Validação	Solucionado	Tempo Remanescente
R4906956	Aberto	0	0	0	0	0	161,52
R4906956	Em Tratamento	22,23	0	0	0	0	139,28
R4906956	Pendente Usuário	22,23	4,77	0	0	0	134,52
R4906956	Em Tratamento	22,23	4,77	7,02	0	0	127,50
R4906956	Pendente Usuário	22,23	10,05	7,02	0	0	122,22
R4906956	Em Tratamento	22,23	10,05	32,42	0	0	96,82
R4906956	Em Validação	22,23	60,87	32,42	0	0	46,00
R4906956	Solucionado	22,23	60,87	32,42	46,00	0	0

Tabela 8 – Métrica de duração das atividades em horas (ID R4906956)
Fonte: O Autor

ID	*Status*	Tempo Decorrido	Aberto	Em Tratamento	Pendente Usuário	Em Validação	Solucionado	Dur: Aberto	Dur: Em Tratamento	Dur: Pendente Usuário	Dur: Em Validação	Dur: Solucionado	Tempo Remanescente
R4906956	Aberto	0,00	1	0	0	0	0	0	0	0	0	0	161,52
R4906956	Em Tratamento	22,23	1	1	0	0	0	22,23	0	0	0	0	139,28
R4906956	Pendente Usuário	27,00	1	1	1	0	0	22,23	4,77	0	0	0	134,52
R4906956	Em Tratamento	34,02	1	2	1	0	0	22,23	4,77	7,02	0	0	127,50
R4906956	Pendente Usuário	39,30	1	2	2	0	0	22,23	10,05	7,02	0	0	122,22
R4906956	Em Tratamento	64,70	1	3	2	0	0	22,23	10,05	32,42	0	0	96,82
R4906956	Em Validação	115,52	1	3	2	1	0	22,23	60,87	32,42	0	0	46,00
R4906956	Solucionado	161,52	1	3	2	1	1	22,23	60,87	32,42	46,00	0	0

Tabela 9 – Conjunto de todas as métricas (ID R4906956)
Fonte: O Autor

4.3 NORMALIZAÇÃO DOS DADOS

Todos as métricas utilizadas como dados de entrada do modelo são valores numéricos. Dessa forma, para que exista uma compatibilidade nas escalas utilizadas, todas as métricas foram normalizadas conforme o método descrito na metodologia (seção 3.3). Os resultados dessa etapa foram as mesmas quatro tabelas apresentadas na seção 4.2 mas com os valores normalizados.

4.4 SEPARAÇÃO DOS DADOS DE TREINO E TESTE

Nesta etapa cada uma das tabelas foi dividida em quatro. Inicialmente, todos os dados foram divididos aleatoriamente em dois subgrupos contendo 70% dos casos e outro contendo 30%. O maior, os dados de treino, foi utilizado na implementação do modelo de regressão e o menor, os dados de teste, para medir os resultados de cada modelo.

Para o modelo de regressão ser implementado é necessário a separação dos dados de entrada da regressão (os X's da função) e do resultado que se busca prever (o Y da função).

Para o presente estudo, os dados de entrada do modelo foram as métricas mineradas e normalizadas, e os dados de saída foram os tempos remanescentes de cada um dos casos.

4.5 IMPLEMENTAÇÃO DO MODELO DE REGRESSÃO

4.5.1 MODELO DE REFERÊNCIA

Para analisar os resultados foi primeiro calculado um modelo simples que serviu de referência para uma análise comparativa com os modelos de regressão de Kernel utilizados. Esse modelo foi realizado através do cálculo da média do tempo remanescente em cada *status* dos dados de treino. A Figura 12 apresenta a média dos tempos remanescentes por *status*. Dessa forma, considerando os casos dos dados de treino, quando um chamado entra no *status Aberto*, este demora em média 99,1 horas até sua conclusão. Apesar do *status Pendente Usuário* ocorrer após ser *Aberto* e ficar *Em Tratamento*, os chamados dos dados de treino tiverem o maior tempo remanescente em média nesse *status*. Isso pode ocorrer pois nem todos os chamados analisados ficaram *Pendente Usuário*, como pode ser observado na Figura 10. Já na etapa *Em Validação* o chamado já está encaminhado para o estado final, em que o tempo remanescente é zero, pois ele é solucionado imediatamente.

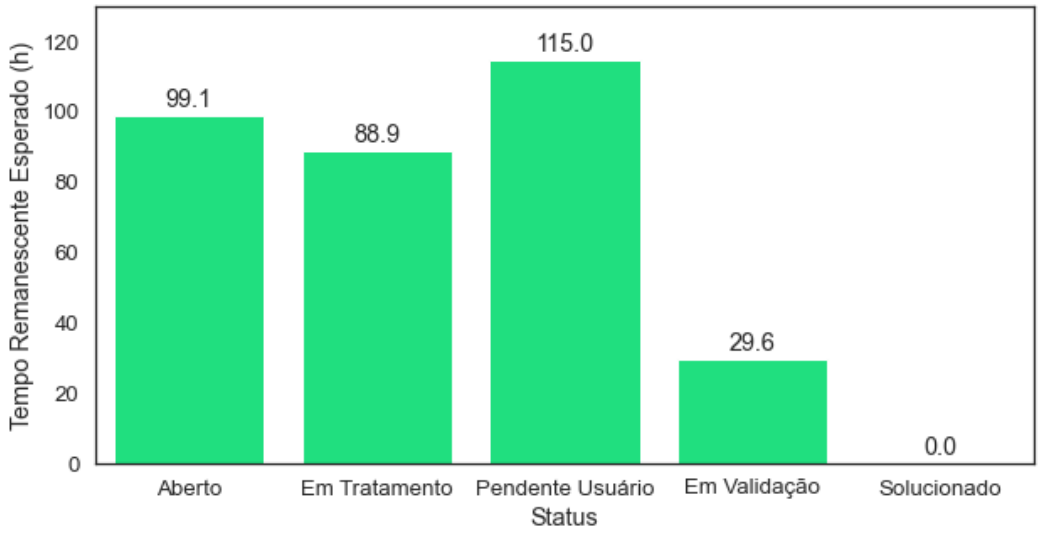

Figura 13 – Tempo remanescente médio por *status*
Fonte: O Autor

Dessa forma, utilizando somente a média como modelo preditivo, foram realizadas as estimativas dos tempos remanescentes dos chamados pertencentes aos dados de teste. Para cada uma das previsões realizadas foram calculados o MAE entre o tempo remanescente previsto e o tempo remanescente real. Na Figura 14 são apresentados os erros absolutos médios dos casos quando estes estavam em cada *status* e do modelo como um todo.

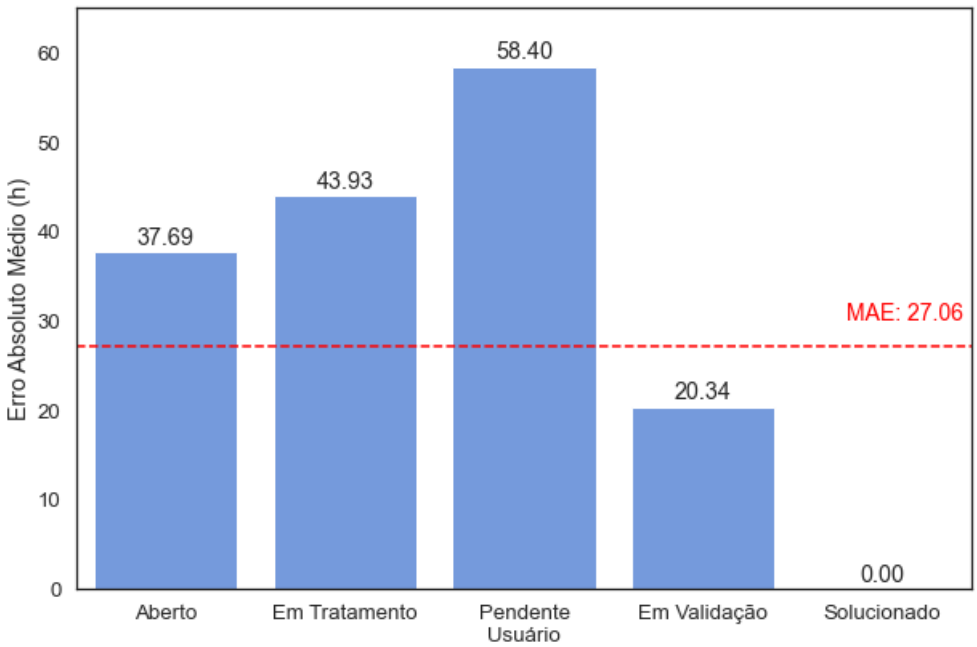

Figura 14 – MAE por status do modelo de referência
Fonte: O Autor

Para esse modelo que foi utilizado como a referência para o resto do estudo, o MAE obtido foi de 27,06 horas. Para o presente estudo, esse foi o erro considerado como base para comparar com os demais modelos de regressão de Kernel aplicados. Portanto, um modelo cujo MAE excede 27,06 horas é em média menos preciso que a utilização da média do tempo remanescente por *status* como preditor.

4.5.2 REGRESSÕES DE KERNEL

A implementação dos modelos preditivos foi realizada com duas funções de Kernel, a de Epanechnikov e a Gaussiana. Para encontrar o parâmetro de suavização que gere a regressão que melhor se encaixe nos dados de treino foi realizada uma busca heurística onde o parâmetro de suavização foi variado entre 0,1 e 50 em iterações de 0,1.

Métricas	Função de Kernel	*h*	MAE
Conjunto	Gaussiana	0,2	24,27
Ocorrência das Atividades	Epanechnikov	2,3	26,07
Conjunto	Epanechnikov	2,9	26,75
Ocorrência das Atividades	Gaussiana	0,9	27,20
Duração das Atividades	Gaussiana	0,2	27,98
Tempo Decorrido	Epanechnikov	0,1	31,13
Duração das Atividades	Epanechnikov	1,4	31,18
Tempo Decorrido	Gaussiana	0,1	32,36

Tabela 10 – Melhores resultados das regressões de Kernel
Fonte: O Autor

Na Tabela 10 estão indicados os melhores resultados obtidos em relação ao MAE para diferentes métricas e parâmetros de suavização (*h*). Foram analisadas as métricas do tempo decorrido daquele chamado até aquele momento, de ocorrência de atividades que conta o número de vezes que cada chamado passou em cada *status*, da duração que cada chamado teve por *status* e do conjunto de todas as métricas em uma tabela única.

A função de Kernel Gaussiana teve os melhores resultados para a aplicação aos dados. Dentre as diferentes métricas que foram testadas nos modelos, a métrica que utilizou o conjunto de todas as informações disponíveis teve o menor MAE para a função Gaussiana (24,77 horas).

Já os melhores parâmetros de suavização foram concentrados em valores mais baixos. Para a função Gaussiana o parâmetro de suavização mais alto foi de 0,9 e para a função de Epanechnikov foi de 2,9, considerando o conjunto das métricas e a ocorrência das atividades, respectivamente. Dos parâmetros de suavização utilizados na busca heurísticas estes estão entre os 5,8% parâmetros mais baixos. Larguras de banda mais baixas levam a regressões menos suaves, mas que se adaptam melhor às características dos dados, dessa forma, as regressões tendem a ser mais complexas quanto mais baixo o parâmetro de suavização (BOWMAN; A.W; AZZALINI, 1997). A Figura 15 apresenta os erros obtidos na regressão de Kernel Gaussiana distribuídos pelos *status* em que os chamados estavam.

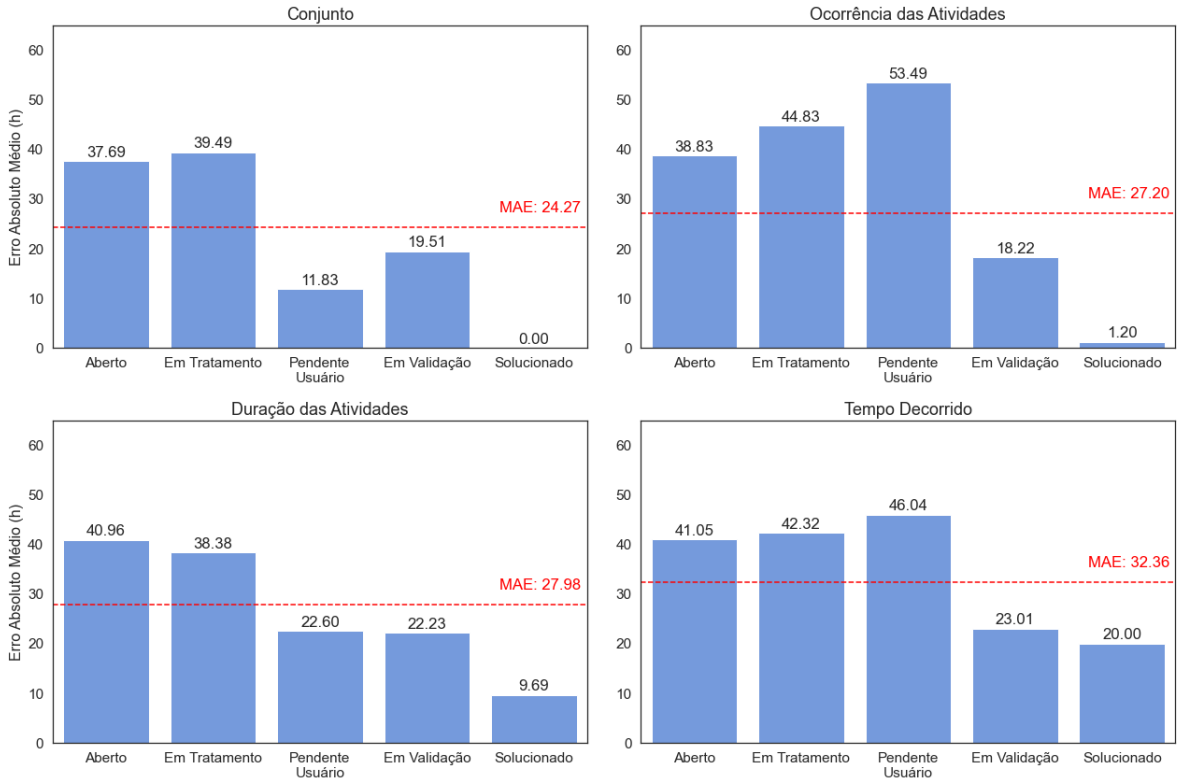

Figura 15 – Erros médios por status para a regressão de Kernel Gaussiana
Fonte: O Autor

Utilizando somente o tempo decorrido como estimador do tempo remanescente, nenhum dado de entrada indica quais *status* o chamado passou até aquele momento. Na figura 9 da análise exploratória pode-se observar que há uma grande variação dos tempos decorridos em cada um dos *status*. Dessa forma, um chamado que durou uma certa quantidade de tempo pode estar em diversos *status*. Sem a informação do *status* que o chamado se encontra há uma perda de uma informação importante, o que resultou no pior das MAE's.

Utilizando somente a duração dos tempos despendidos em cada *status*, o erro obtido foi mais elevado do que o obtido no modelo de referência, com um MAE de 27,98 horas. O modelo de referência, por utilizar a média do tempo remanescente por *status*, sempre prevê corretamente chamados no *status Solucionado*, que já foram concluídos. Utilizando o tempo de duração em cada *status*, no entanto, o modelo teve um erro de 9,69 horas na previsão do tempo remanescente de chamados solucionados.

Utilizando somente a quantidade de vezes que um chamado passou por cada *status*, o MAE obtido pelo modelo foi de 27,20 horas, levemente menor do que o obtido no modelo de referência. Com essa métrica, o modelo tem a informação que os chamados foram solucionados, o que resultou em um erro bem pequeno na previsão do tempo remanescente de chamados nesse

status. Isso auxiliou essa métrica a ter um MAE menor do que a utilização da métrica de duração das atividades. No entanto, essa métrica teve um resultado pior na previsão de chamados no *status* de *Pendente Usuário*, com o maior erro obtido em 53,49 horas. Este resultado se assemelha muito ao obtido no modelo de referência, que também não conseguiu prever chamados nesse *status* com previsão, resultando em seu maior erro também.

A utilização do conjunto das métricas resultou no melhor resultado com um MAE de 24,27 horas. Neste caso, pode ser observado que o *status Solucionado* que possui tempo remanescente de zero, sempre é previsto com sucesso. Ademais, o *status Pendente Usuário*, que possui a pior precisão no modelo de referência, teve uma das melhores nesse modelo, com um MAE comparativamente baixo, de 11,83 horas. Ademais, o único *status* que não teve um resultado superior ao do modelo de referência foi no *status Aberto*, no qual o MAE foi idêntico ao obtido pelo modelo de referência em 37,69 horas. Apesar do erro absoluto ser idêntico na média, as previsões realizadas foram diferentes entre os dois modelos.

A Figura 16 apresenta um gráfico comparativo entre as diferentes métricas da regressão de Kernel Gaussiana com o modelo de referência. Para isso, os erros médios por *status* foram normalizados, e são representados como o percentual do erro médio do modelo de referência (MAE). No eixo-x, o único *status* que não é representado foi o *Solucionado*, pois no modelo de referência este foi zero, o que resultaria em uma divisão por zero.

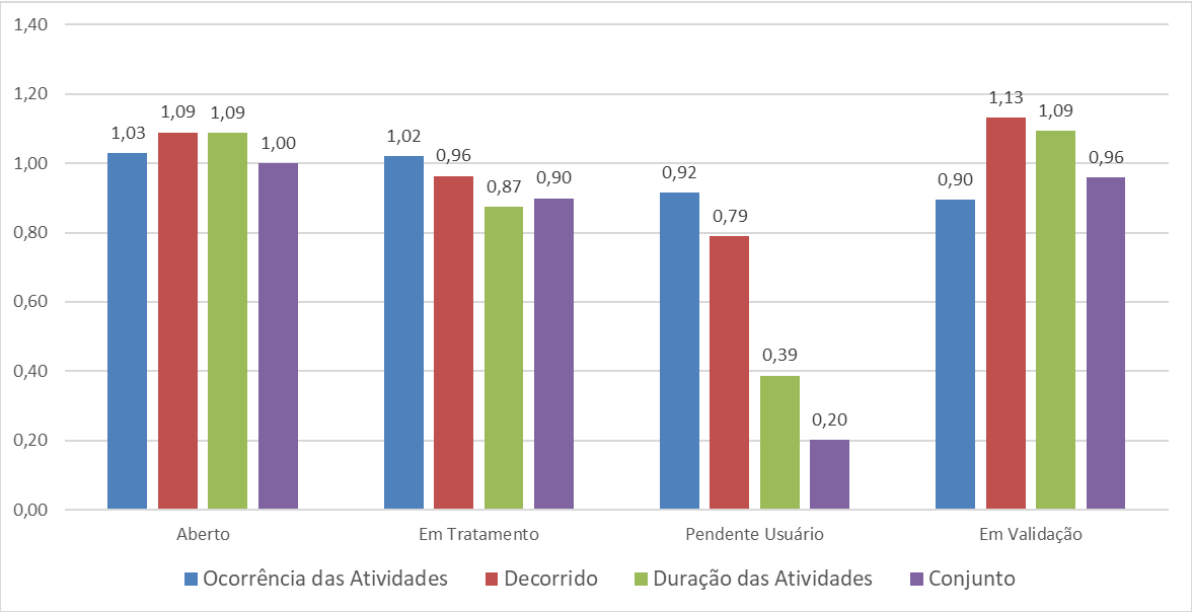

Figura 16 – Percentual de erro entre a regressão de Kernel Gaussiana e o modelo de referência
Fonte: O Autor

Dessa forma, valores acima de um correspondem a MAE's mais elevados que o modelo de referência e abaixo MAE's menores do que o de referência. Para o primeiro *status* recebido,

Aberto, onde ainda há poucas informações sobre os chamados, a utilização de uma métrica simples como a média do tempo remanescente foi superior ou igual à todas as regressões de Kernel realizadas. No entanto, para os demais *status* há uma grande diferença entre a utilização de um modelo simples como o de referência para o modelo não-paramétrico de Kernel.

A métrica de ocorrência de atividades, apesar de ter tido um MAE geral menor que o do modelo de referência, ficou muito próximo deste para os diferentes casos. Já a métrica em conjunta obteve os melhores resultados, exceto quando comparada com a duração das atividades *Em Tratamento*, que obteve um resultado levemente superior.

Os principais ganhos do modelo de Kernel Gaussiana sobre o modelo de referência foram no *status Pendente Usuário*. Este é o único *status* onde há uma divergência do processo realizado, como é observado no fluxograma da Figura 10. Com essa camada de complexidade adicional, a métrica conjunta que utiliza a maior quantidade de informações obteve o melhor resultado com um MAE de 11,83 horas, 20% das 58,40 horas obtidas no modelo de referência.

Para todas as regressões de Kernel de Epanechnikov (KE) o MAE obtido foi mais elevado do que o melhor resultado encontrado com a regressão de Kernel Gaussiana. A Figura 17 apresenta todos os MAE's obtidos por *status*.

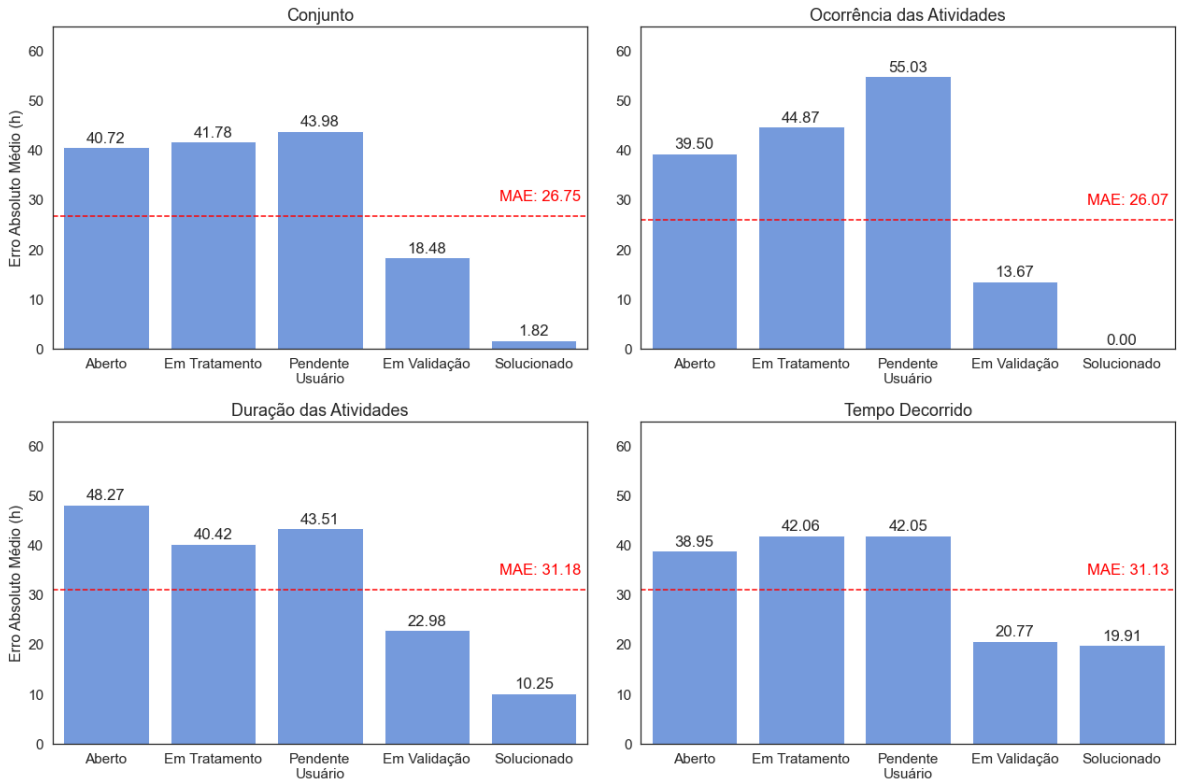

Figura 17 – Erros médios por status para a Regressão de Kernel de Epanechnikov
Fonte: O Autor

Diferentemente do resultado obtido com a função de Kernel Gaussiana, o melhor resultado encontrado foi com a utilização da métrica de ocorrência de atividades com MAE de 26,07 horas e o pior com a utilização do tempo decorrido com 31,13 horas.

O modelo que utilizou a métrica do tempo decorrido teve uma classificação ruim dos casos já solucionados, cuja previsão deveria ser zero horas para todos os casos, resultando em um desempenho similar à Gaussiana e inferior ao modelo de referência.

Já o conjunto das métricas, não classificou todos os casos solucionados corretamente, levando a um MAE de 1,82 horas. Esse foi um resultado interessante uma vez que somente com a métrica de Ocorrência de Atividades, o modelo previu todos os chamados solucionados corretamente. Neste caso, a maior quantidade de informações teve um efeito oposto ao observado com a função Gaussiana, e resultou em um MAE mais elevado. Contudo, ambas as métricas ainda obtiveram resultados superiores ao modelo de referência.

Na Figura 18, todos os modelos criados foram normalizados conforme o modelo de referência. Diferentemente do modelo Gaussiano, as regressões de Epanechnikov tiveram resultados próximos ao modelo de referência. Quando no *status* em *Aberto*, a métrica de duração das atividades foi consideravelmente pior que os demais modelos, apresentando um MAE 22% maior que o do modelo de referência. A ocorrência das atividades teve valores próximos do modelo de referência para os três primeiros *status*, mas depois teve uma precisão mais elevada na previsão de chamados *Em Validação*.

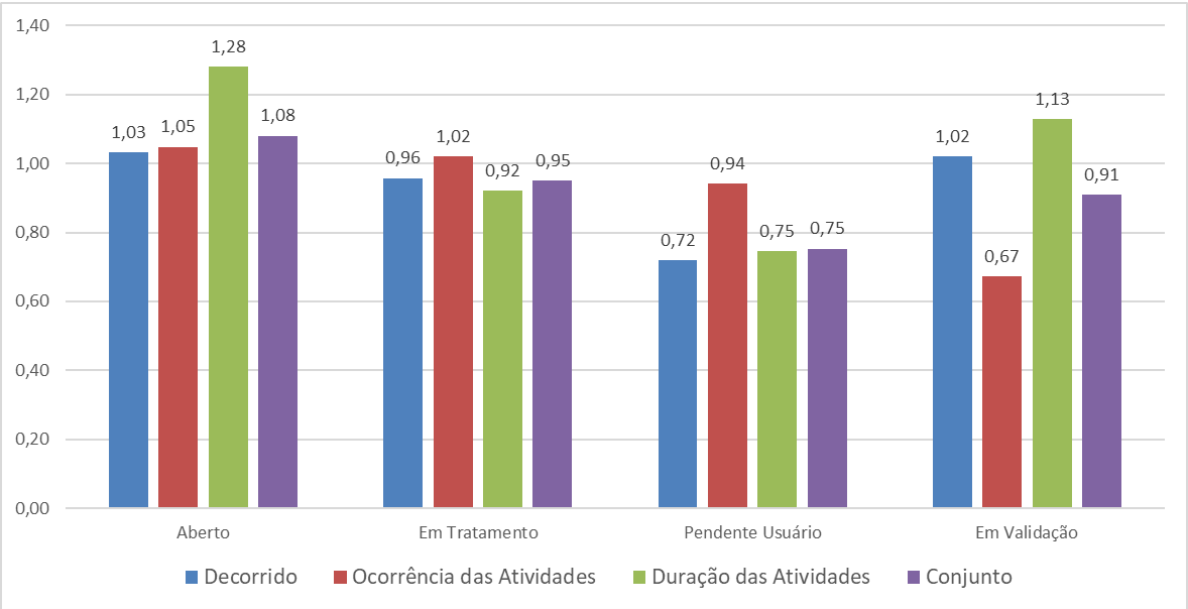

Figura 18 – Percentual de erro entre a regressão de Kernel de Epanechnikov e o modelo de referência
Fonte: O Autor

As previsões do tempo remanescente conseguem ser utilizadas pelos solucionadores para entender quais chamados podem cometer infrações no SLA. Para isso, é preciso somar o tempo decorrido do chamado com a previsão do tempo remanescente e comparar se esse valor infringiria ou não o SLA. Para o presente estudo, essa classificação entre violado e não violado foi realizada para cada um dos chamados usando os diferentes modelos elaborados. Dessa forma, pode-se comparar os resultados obtidos com os dados históricos para criar uma métrica de percentual de chamados classificados corretamente. Essa métrica, pode ser utilizada para entender se um modelo corretamente preveria uma violação e possibilitar o time de solucionadores de atuarem de forma proativa no chamado.

A Figura 19 apresenta o percentual de chamados classificados corretamente com o modelo de referência e os melhores resultados obtidos com as funções Gaussiana e de Epanechnikov.

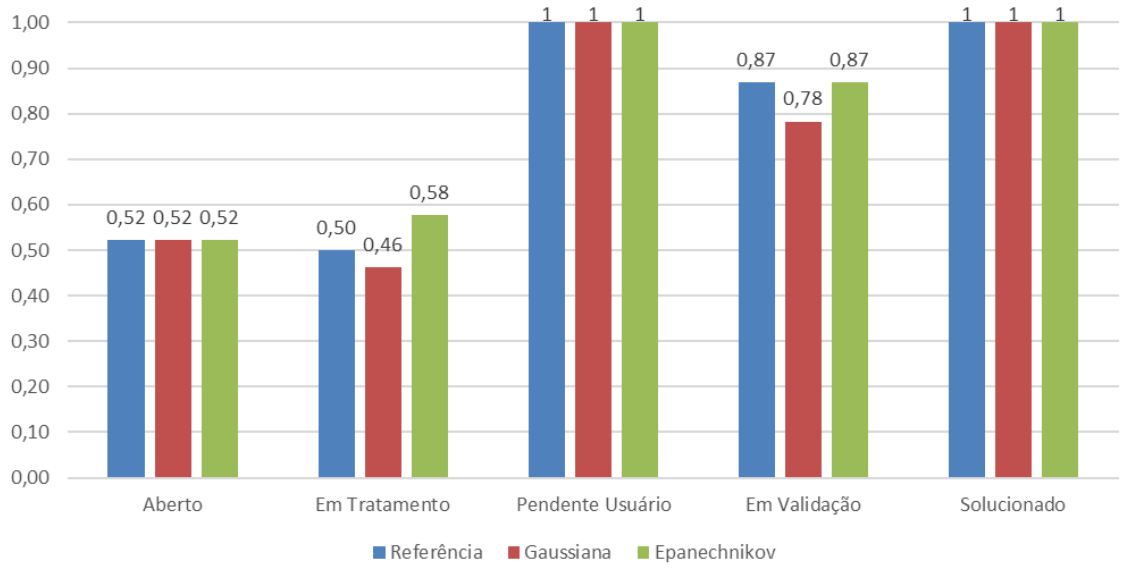

Figura 19 – Percentual de chamados classificados corretamente
Fonte: O Autor

Na figura pode ser observado que no status em *Aberto* todos os modelos conseguiram classificar a violação de SLA corretamente em 52% dos casos. No entanto, nos status *Em Tratamento* e *Em Validação* os modelos divergiram. O melhor modelo para ambos os casos foi o de Epanechnikov que classificou os chamados *Em Tratamento* com precisão de 58% e E*m Validação* com precisão de 87%. Nos status *Pendente Usuário* e *Solucionado* todos os modelos classificaram corretamente os casos em andamento, indicando que qualquer um destes pode ser utilizado obtendo-se o mesmo resultado.

Embora a abordagem do modelo Gaussiano tenha sido a melhor opção na regressão, fornecendo o menor MAE, pode-se observar na Figura 19 que isso não se repete para a classificação. Isso devido ao foco do trabalho em prever o tempo remanescente dos chamados e não de classificá-los entre violados e não-violados. Para gerar um modelo preditivo puramente focado em classificar entre violados e não violados, podem ser realizados mais estudos com enfoque em modelos de classificação em detrimento do modelo de regressão utilizado nesse estudo.

4.6 SUMÁRIO DO CAPÍTULO

Através da análise exploratória foi observado que o processo seguido pelo RE utilizado no estudo pertence a um processo simples com poucas etapas e divergências. Também foi descoberto a necessidade da utilização da duração em horas úteis para os cálculos realizadas, dado o horário de trabalho do time de solucionadores de TI e o cálculo das infrações de SLA. Todas as regressões de Kernel, fizeram uso de parâmetros de suavização (h) com valores baixos para obter seus melhores resultados. Dos modelos comparados, a função Gaussiana com a métrica conjunta teve a maior precisão com um MAE de 24,27 horas. A regressão de Kernel de Epanechnikov teve sua maior precisão com a métrica de ocorrência das atividades, obtendo MAE de 26,07 horas. Dessa forma, ambos os modelos foram superiores ao modelo de referência analisada. No entanto, em uma análise da utilização da previsão para classificar chamados que violaram seu SLA, todos os modelos tiveram resultado muito similar, apontando que todos poderiam ser utilizados para esse fim.

5. CONCLUSÕES

O presente trabalho teve como objetivo investigar a aplicação das ferramentas de mineração de processos e NPR para auxiliar a gerenciar proativamente o SLA e analisar diferentes modelos de previsão, implementando-os em um estudo de caso. Para isso, o SLA e suas implicações nas organizações foram estudados, foram apresentadas diferentes técnicas de MP e de NPR e os modelos construídos foram implementados em um estudo de caso com dados reais.

Dessa forma, o melhor resultado obtido para a previsão do tempo remanescente foi com a utilização da função Gaussiana com o conjunto de métricas, que obteve um MAE de 24,27 horas. A função de Epanechnikov, em seu melhor caso, obteve erro similar em 26,07 horas, ambas abaixo do modelo de referência utilizado (27,06 horas). No entanto, quando os modelos foram utilizados para prever as violações no SLA estes obtiveram resultados similares, apontando que todos poderiam ser utilizados para esse fim.

Ademais, as regressões de Kernel fizeram uso de parâmetros de suavização (h) com valores baixos para obter seus melhores resultados. Valores baixos de h podem gerar uma interpolação dos dados, provocando um sobre ajuste e prejudicando a precisão do modelo quando utilizado nos dados de teste. Dessa forma, pode ser concluído que os dados de treino e teste são similares, fruto de um RE descrito por um processo simples que foi utilizado no estudo.

Os resultados obtidos nesse trabalho, sugerem que a MP e os modelos de NPR podem ser utilizados para prever tempo remanescente de chamados de TI e auxiliar no gerenciamento proativo do SLA. No entanto, dado a simplicidade do processo utilizado, mais estudos devem ser realizados em processos mais complexos onde há uma maior quantidade de atividades e divergências. Ademais, outros modelos de regressão e métricas também podem ser avaliados com o objetivo de encontrar melhores resultados do que os obtidos.

REFERÊNCIAS

ABDI, Herve; WILLIAMS, Lynne J.. **Normalizing data**. Encyclopedia of research design 1. 2010.

BATISTA, Gustavo Enrique de Almeida Prado**. Pré-processamento de dados em aprendizado de máquina supervisionado**. Diss. Universidade de São Paulo. 2003.

BOWMAN, A.W.; AZZALINI, A. **Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations**. Vol. 18. OUP Oxford. 1997.

DE LEONI, M., *et al.* **A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs**. Inf. Syst. 56, 235–257. 2016.

DEMIR, Serdar; TOKTAMIS, Öniz. **On the adaptive Nadaraya-Watson kernel regression estimators**. Hacettepe Journal of Mathematics and Statistics 39.3. 429-437. 2010.

EDGINGTON, T.M., RAGHU, T.S.; VINZE, A.S. **Using process mining to identify coordination patterns in IT service management**. Decision Support Systems 49.2: 175-186. 2010.

FAN, Jianqing; GIJBELS, Irene. **Local polynomial modelling and its applications: monographs on statistics and applied probability 66**. Routledge, 2018.

FREITAS, Marcos André dos Santos. **Avaliações Fundamentos do Gerenciamento de Serviços de TI: Preparatório para a certificação ITIL Foundation**. 2. ed. [S. l.]: Brasport, 2011. 405 p.

HALVEY, John K.; MELBY, Barbara Murphy. **Business process outsourcing: Process, strategies, and contracts**. John Wiley & Sons, 2007.

IZBICKI, Rafael; DOS SANTOS, Tiago Mendonça. **Aprendizado de máquina: uma abordagem estatística**. Rafael Izbicki. 2020.

JUNIOR, Gilmar Teixeira. **Modelagem de Sistemas de Informação para Mineração de Processos: Características e Propriedades das Linguagens**. p. 129 , 2017.

KEBEDE, Musie; DUMAS, M.. **Comparative evaluation of process mining tools**. University of Tartu. 2015.

LEITNER, Philipp, *et al*. **Runtime prediction of service level agreement violations for composite services**. Service-oriented computing. ICSOC/ServiceWave 2009 workshops. Springer, Berlin, Heidelberg. 2009.

MAGER, Christian. **Analysis of Service Level Agreements using Process Mining techniques**. 2014.

MEHDIYEV, Nijat; FETTKE, Peter. **Explainable Artificial Intelligence for Process Mining: A General Overview and Application of a Novel Local Explanation Approach for Predictive Process Monitoring**. Interpretable Artificial Intelligence: A Perspective of Granular Computing 937: 1. 2021.

MEHTA, Kunj. **Kernel Regression from Scratch in Python**. Towards Data Science. [S.I.] 2020. Disponível em: <https://github.com/kunjmehta/Medium-Article-Codes/blob/master/gaussian-kernel-regression-from-scratch.ipynb>. Acesso em: 20 dez. 2021.

NETO, José Inácio Jaeger, et al. **A percepção dos gestores de TI em relação às práticas de governança de TI adotadas em empresas do Rio Grande do Sul**. Revista Eletrônica de Sistemas de Informação 8.1. 2009.

NGUYEN, Quang Hung, et al. **Influence of data splitting on performance of machine learning models in prediction of shear strength of soil.** Mathematical Problems in Engineering 2021. 2021.

PINHEIRO, Carlos Carvalho; *et al.*. **Modelo Autônomo de Mineração de Dados para Identificação de Violações de SLAs e Baixa QoS para os Serviços de TI**. 10.13140/RG.2.1.2616.3045. 2015.

PRAVILOVIC, Sonja; APPICE, Annalisa; MALERBA, Donato. **Process mining to forecast the future of running cases**. International workshop on new frontiers in mining complex patterns. Springer, Cham, 2013.

RAMAKRISHNAN, Raghu; KAUR, Arvinder. **Performance evaluation of web service response time probability distribution models for business process cycle time simulation**. Journal of Systems and Software 161: 110480. 2020.

VAN DER AALST, *et al.*. **Workflow mining: Discovering process models from event logs**. IEEE transactions on knowledge and data engineering 16.9. 1128-1142. 2004.

VAN DER AALST, Wil MP; *et al.*. **ProM: The process mining toolkit**. BPM (Demos) 489.31. 2009.

VAN DER AALST, Wil MP; *et al.*. **Beyond process mining: From the past to present and future**. International Conference on Advanced Information Systems Engineering. Springer, Berlin, Heidelberg. 2010.

VAN DER AALST, Wil MP; *et al*.. **Process mining manifesto**. International conference on business process management. Springer, Berlin, Heidelberg. 2011a.

VAN DER AALST, *et al*.. **Time prediction based on process mining**. Information systems 36.2. 450-475. 2011b.

VAN DONGEN, Boudewijn F.; CROOY, Ronald A., VAN DER AALST, Wil MP. **Cycle time prediction: When will this case finally be finished?**. OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, Berlin, Heidelberg. 2008.

WEBER, B., *et al*. **Supporting flexible processes through recommendations based on history**. BETA Working Paper Series, WP 212, Eindhoven University of Technology. 2007.

WITTEK, Peter. **Quantum machine learning: what quantum computing means to data mining.** Academic Press, 2014.

WU, Linlin; BUYYA, Rajkumar. **Service level agreement (SLA) in utility computing systems**. **Performance and dependability in service computing: Concepts, techniques and research directions**. IGI Global, 2012. 1-25.

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA

Termo de Declaração de Autenticidade de Autoria

Declaro, sob as penas da lei e para os devidos fins, junto à Universidade Federal de Juiz de Fora, que meu Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia de Produção é original, de minha única e exclusiva autoria. E não se trata de cópia integral ou parcial de textos e trabalhos de autoria de outrem, seja em formato de papel, eletrônico, digital, áudio-visual ou qualquer outro meio.
Declaro ainda ter total conhecimento e compreensão do que é considerado plágio, não apenas a cópia integral do trabalho, mas também de parte dele, inclusive de artigos e/ou parágrafos, sem citação do autor ou de sua fonte.
Declaro, por fim, ter total conhecimento e compreensão das punições decorrentes da prática de plágio, através das sanções civis previstas na lei do direito autoral[1] e criminais previstas no Código Penal[2], além das cominações administrativas e acadêmicas que poderão resultar em reprovação no Trabalho de Conclusão de Curso.

Juiz de Fora, 19 de fevereiro de 2022.

Felipe Carrullo Pumar	201649055
NOME LEGÍVEL DO ALUNO (A)	Matrícula
Felipe Carrullo Pumar	179.693.997-89
ASSINATURA	CPF

[1] LEI N° 9.610, DE 19 DE FEVEREIRO DE 1998. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências.
[2] Art. 184. Violar direitos de autor e os que lhe são conexos: Pena – detenção, de 3 (três) meses a 1 (um) ano, ou multa.