

Relatório Técnico: Desafio CT-Mon RNP

Vandirleya Barbosa¹ e Melissa Alves¹

¹Universidade Federal do Piauí (UFPI)
Picos – PI – Brasil

{vandirleya.barbosa, melissaalves}@ufpi.edu.br

Abstract. *This work presents an analysis and forecasting approach for the CT-Mon/RNP Challenge, focused on monitoring high-traffic and dynamic networks. Using metrics data such as DASH, RTT, and Traceroute, we propose a predictive model based on the Random Forest algorithm to forecast means and standard deviations in time series. The preprocessing involved organizing data into temporal windows, while the model was evaluated using the MAPE metric. The submissions were tested on the Kaggle platform, where the results demonstrated progressive improvements and performance surpassing the baseline.*

Resumo. *Este trabalho apresenta uma análise e previsão para o Desafio CT-Mon/RNP, focado no monitoramento de redes de alto tráfego e dinamicidade. Utilizando dados de métricas como DASH, RTT e Traceroute, propomos um modelo preditivo baseado no algoritmo Random Forest para prever médias e desvios padrão em séries temporais. O pré-processamento envolveu a organização de dados em janelas temporais, enquanto o modelo foi avaliado com a métrica MAPE. As submissões realizadas foram testadas na plataforma Kaggle, onde os resultados demonstraram melhorias progressivas e desempenho superior ao baseline.*

1. Introdução

O avanço dos serviços de streaming tem revolucionado o consumo de conteúdo digital, impulsionando a demanda por entregas de vídeos de alta qualidade de forma ágil e eficiente. Essa mudança de paradigma é o domínio ascendente das plataformas de streaming, que rapidamente ascenderam à vanguarda das preferências de consumo de conteúdo e estão liderando a indústria do entretenimento [Nguyen et al. 2023]. Em junho de 2024, o streaming representou 40,3% do uso total de TV nos Estados Unidos [Nielsen 2024]. Este cenário traz consigo a necessidade de infraestruturas tecnológicas avançadas e estratégias de gerenciamento de redes cada vez mais sofisticadas, capazes de lidar com o crescente tráfego de dados sem comprometer a experiência do usuário. A evolução dessas tecnologias tornou-se um pilar fundamental na entrega de serviços de qualidade em um ambiente marcado por padrões de uso intensivo e dinâmico.

Para atender a essas demandas, as redes modernas enfrentam desafios significativos, incluindo a necessidade de decisões ágeis baseadas em métricas de desempenho como congestionamento, latência e disponibilidade de recursos. Com o aumento contínuo de fontes de dados de fluxo de tráfego e o desenvolvimento de tecnologia de inteligência artificial, construir um modelo de previsão de fluxo de tráfego em toda a rede preciso e orientado por dados se torna viável [Chen and Chen 2022]. A capacidade de prever

esses indicadores não só permite uma melhor utilização da infraestrutura existente, mas também abre caminho para ajustes dinâmicos que otimizam a entrega de conteúdo. Assim, o uso de ferramentas analíticas fundamentadas em dados tornou-se indispensável, possibilitando a identificação de gargalos e a implementação de estratégias que atendam às necessidades crescentes de usuários e provedores de serviços [Silva 2015].

Nesse contexto, este relatório explora a aplicação de técnicas avançadas para monitoramento e previsão de métricas de rede, com foco na seleção de servidores ideais para entrega de streaming. A abordagem foi aplicada no âmbito do Desafio CT-Mon/RNP, que propõe a análise de dados reais provenientes da RedeIpê, uma infraestrutura de alto tráfego utilizada por instituições de ensino e pesquisa no Brasil. O desafio envolve a previsão de métricas como DASH, RTT e Traceroute, a partir de séries temporais estruturadas, e exige soluções capazes de lidar com o comportamento dinâmico das redes e volumes significativos de informações. A nossa proposta busca abordar os desafios técnicos associados ao comportamento dinâmico das redes, promovendo soluções que aliam eficiência operacional e qualidade de serviço. Espera-se que os resultados contribuam tanto para a melhoria das estratégias de gerenciamento quanto para sua aplicabilidade em cenários reais, fortalecendo a gestão de redes em ambientes complexos e desafiadores.

As seções subsequentes são organizadas da seguinte forma. A Seção 2 se aprofunda no referencial teórico. A Seção 3 descreve a metodologia usada. A Seção 4 apresenta os resultados obtidos. Por fim, a Seção 5 apresenta a conclusão do trabalho.

2. Referencial Teórico

Esta seção apresenta os conceitos essenciais para a compreensão deste trabalho. As subseções a seguir abordam os tópicos de mineração de dados, o desafio CT-Mon RNP e o aprendizado de máquina, incluindo informações sobre o funcionamento, o algoritmo e a métrica avaliada que serão utilizados neste estudo.

2.1. Mineração de Dados

A mineração de dados é um processo crítico no campo da ciência de dados, focado na descoberta de padrões e informações valiosas a partir de grandes volumes de dados. Além disso, tem se consolidado como uma ferramenta poderosa para a descoberta de informações que revelam estruturas de conhecimento e orientam decisões em condições de incerteza limitada [da Costa Côrtes et al. 2002]. Esse processo utiliza técnicas estatísticas, matemáticas e computacionais para explorar, analisar e transformar dados brutos em conhecimento aplicável. Essa análise indutiva permite identificar tendências e relações que poderiam passar despercebidas em uma avaliação tradicional [Mena 1999]. Uma etapa fundamental da mineração de dados é a limpeza e integração dos dados, que prepara as informações para análise. Em seguida, algoritmos de aprendizado de máquina ou regras de associação são aplicados para identificar correlações, previsões ou classificações que auxiliam na tomada de decisões. O crescimento exponencial de dados em áreas como redes sociais, IoT e aplicações empresariais tornou a mineração de dados indispensável para a criação de soluções inovadoras e competitivas.

Um dos principais desafios da mineração de dados é lidar com dados complexos e de diferentes fontes. Dados estruturados, semi-estruturados e não estruturados exigem

abordagens distintas para análise e interpretação. Desse modo, a mineração de dados pode ser resumida como o processo de extração de conhecimento ou a mineração de informações em grandes volumes de dados [Jiawei and Micheline 2006]. Além disso, a mineração de dados também precisa abordar questões relacionadas à privacidade e à segurança dos dados, pois a exposição inadequada pode levar a violações éticas e legais. O que aponta que técnicas de mineração de dados são amplamente usadas para descobrir relações inesperadas entre dados [Papakyriakou and Barbounakis 2022], destacando sua capacidade de transformar dados brutos em conhecimento valioso. A capacidade de extrair informações relevantes, mantendo a integridade e a confidencialidade, torna a mineração de dados uma área desafiadora, mas essencial para a análise preditiva, a descoberta de padrões e o suporte às tomadas de decisões informadas.

2.2. Desafio CT-Mon RNP

O desafio CT-Mon RNP, promovido pelo Comitê Técnico de Monitoramento da Rede Nacional de Ensino e Pesquisa¹, foca em análises de dados de monitoramento de redes com alto volume e dinamicidade. Em especial, os dados de streaming representam um cenário crítico, pois refletem medidas em tempo real, como RTT (tempo de ida e volta), Dash (dados de streaming em várias resoluções) e Traceroute (caminho percorrido pelos pacotes). Esses dados são coletados continuamente, exigindo ferramentas de processamento que lidem com sua natureza temporal e seu grande volume. O tratamento eficiente desses fluxos é vital para a identificação de anomalias, previsões de desempenho e suporte à otimização de infraestrutura. Dados de streaming trazem também desafios técnicos como latência e consistência, requerendo soluções robustas para entrega e processamentos eficazes. Para enfrentar esses desafios, é importante adotar modelos que permitam avaliar e prever a qualidade do serviço entregue ao usuário, considerando os fatores que impactam diretamente sua experiência.

O Índice de Qualidade da Experiência (QoE) de Streaming é um modelo eficiente para prever a QoE em serviços de vídeo, pois considera fatores como a degradação instantânea da qualidade, os eventos de interrupção da reprodução e as interações entre esses elementos [Duanmu et al. 2017]. Esse desafio não se limita apenas promover a inovação no monitoramento de redes, mas também busca incentivar o desenvolvimento de modelos analíticos capazes de fornecer informações práticas para a gestão eficiente da infraestrutura. Os participantes utilizam dados reais e estruturados em arquivos JSON para prever métricas relacionadas ao desempenho da RedeIpê². Este evento estimula aplicações em aprendizado de máquina, engenharia de *features* e análise de séries temporais e outras relacionadas à precisão preditiva. Assim, o CT-Mon RNP não apenas expõe os participantes a desafios complexos, mas também promove soluções para problemas reais enfrentados em infraestrutura crítica de redes acadêmicas e institucionais.

2.3. Aprendizado de Máquina

O aprendizado de máquina é uma área que visa desenvolver sistemas capazes de aprender e melhorar seu desempenho a partir de dados. De forma mais ampla, pode ser definido como uma área de estudo que proporciona aos computadores a capacidade de aprender sem serem explicitamente programados [Wiederhold and McCarthy 1992]. Esse pro-

¹<https://plataforma.rnp.br/ct-mon>

²<https://redeipe.rnp.br/>

cesso envolve etapas como coleta, limpeza, análise e treinamento em dados representativos. A escolha de algoritmos adequados depende do tipo de problema a ser resolvido, como classificação, regressão ou *clustering*. A integração do aprendizado de máquina com o processamento de dados tem se tornado fundamental em diversas áreas, desde diagnósticos médicos até previsões de mercado. Um aspecto essencial é a engenharia de *features* [Domingos 2012], que consiste em identificar atributos significativos para melhorar o desempenho dos modelos, reduzindo assim o erro e aumentando a generalização dos resultados obtidos.

Entre os algoritmos mais relevantes do aprendizado de máquina destaca-se o Random Forest [Rigatti 2017]. O Random Forest é um algoritmo popular de aprendizado supervisionado baseado em um conjunto de árvores de decisão. Ele utiliza o conceito de bagging, que combina várias árvores independentes para produzir uma previsão final robusta. Cada árvore no Random Forest é treinada em uma amostra diferente do conjunto de dados e utiliza um subconjunto aleatório de variáveis [Ramadhan et al. 2017], promovendo diversificação e reduzindo o risco de *overfitting*. Graças à sua capacidade de lidar com dados de alta dimensionalidade e relações não lineares, o Random Forest é amplamente usado em tarefas como classificação, regressão e análise de dados temporais, mostrando-se eficaz em contextos como previsão de desempenho em redes e diagnósticos em saúde.

A métrica Mean Absolute Percentage Error (MAPE) é amplamente utilizada para avaliar a precisão de modelos preditivos [Kim and Kim 2016] em contextos diversos, incluindo o aprendizado de máquina. O MAPE é calculado pela média do erro absoluto percentual entre os valores previstos e os valores reais. Sua fórmula é dada por, onde representa os valores reais e os valores previstos. Embora seja intuitivo e fácil de interpretar, o MAPE apresenta limitações em cenários com valores reais próximos de zero, podendo distorcer os resultados. Apesar disso, é uma métrica amplamente adotada para modelos preditivos em séries temporais, fornecendo uma perspectiva clara sobre o desempenho relativo em termos percentuais. A fórmula do MAPE é dada por:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

Onde n representa o número total de observações; y_i é o valor real da i -ésima observação; \hat{y}_i indica o valor previsto para a i -ésima observação; e $|\cdot|$ representa o valor absoluto. Para o cálculo da métrica subtrai-se o valor previsto (\hat{y}_i) do valor real (y_i) para determinar o erro. Em seguida, divide-se o erro pelo valor real (y_i), transformando-o em um percentual relativo. O próximo passo é calcular o valor absoluto, ignorando a direção do erro. Depois, somam-se todos os valores percentuais absolutos e divide-se pelo número total de observações (n). Finalmente, multiplica-se o resultado por 100 para expressar o erro como uma porcentagem.

3. Metodologia

Esta seção apresenta a metodologia utilizada no desenvolvimento deste trabalho. Inicialmente, são descritos os dados e ferramentas utilizados, seguidos pela explicação do pré-processamento e das técnicas aplicadas para modelagem e análise. Em seguida, discute-se

a abordagem adotada para a resolução do problema e avaliação dos resultados. Por fim, são apresentados os ajustes realizados ao longo do processo.

3.1. Base de Dados

A Tabela 1 apresenta as informações do conjunto de dados utilizado no Data Challenge CT-Mon/RNP, que foi coletado entre nós da RedeIpê. Este conjunto engloba dois pontos de presença (PoPs) designados como clientes, localizados em Salvador (BA) e Rio de Janeiro (RJ), e quatro como servidores, situados em Fortaleza (CE), Brasília (DF), Vitória (ES) e Teresina (PI). As medições realizadas incluem três tipos principais de dados: DASH, RTT e Traceroute. Cada par cliente-servidor é monitorado continuamente, gerando arquivos no formato JSON, organizados em janelas temporais de 5 minutos para as requisições DASH. Esses dados fornecem uma base abrangente para a criação de modelos preditivos, destacando a dinâmica das redes e o comportamento de suas conexões e são utilizados para o treinamento. Adicionalmente, existem os dados de teste coletados entre 22/6/2024 e 1/7/2024, utilizados posteriormente para avaliar as soluções propostas no desafio. Estes dados, transformados em consultas, abrangem todas as medições de DASH, RTT e Traceroute realizadas durante o período de teste. As consultas exigem a previsão da média e do desvio padrão das requisições DASH para os dois próximos intervalos de medição, destacando a complexidade do desafio.

Table 1. Clientes e servidores do Data Challenge CT-Mon/RNP.

| Cliente | Servidor | Tipo de Dados |
|---------|----------|-----------------------|
| BA | CE | DASH, RTT, Traceroute |
| BA | DF | DASH, RTT, Traceroute |
| BA | ES | DASH, RTT, Traceroute |
| BA | PI | DASH, RTT, Traceroute |
| RJ | CE | DASH, RTT, Traceroute |
| RJ | DF | DASH, RTT, Traceroute |
| RJ | ES | DASH, RTT, Traceroute |
| RJ | PI | DASH, RTT, Traceroute |

3.2. Resolução do Problema

O processo de resolução do problema começou com o download e extração da base de dados necessária para o desafio. O objetivo era prever a média e o desvio padrão das requisições DASH para os próximos dois intervalos de medição (5 minutos e 10 minutos no futuro), utilizando dados históricos entre pares cliente-servidor. O fluxo de trabalho foi dividido em duas etapas principais: pré-processamento e treinamento.

Pré-processamento – Inicialmente, os dados foram carregados para análise a partir de arquivos JSON organizados em diretórios representando clientes e servidores. Para cada par cliente-servidor, foram extraídos os rates das requisições DASH, e esses valores

foram ordenados temporalmente. Em seguida, calculamos a média e o desvio padrão dos rates em janelas de 5 minutos (correspondendo a 15 requisições por intervalo), gerando uma série temporal regularizada. Para atender à estrutura do desafio, agrupamos as séries temporais em conjuntos de 12 medições consecutivas. As 10 primeiras medições de cada grupo (representando 50 minutos) foram usadas como entrada (X), enquanto as últimas 2 medições (equivalente a 5 minutos e 10 minutos no futuro) foram usadas como rótulos (y). Cada vetor de entrada incluiu características como: (i) Identificação do cliente e servidor. (ii) Média e desvio padrão dos rates nas primeiras 10 medições. (iii) Valores específicos das médias e desvios padrão da 10ª medição. Os rótulos (y) foram preenchidos com as médias e desvios padrão das 11ª e 12ª medições.

Treinamento – Para o modelo preditivo, utilizamos o algoritmo Random Forest Regressor, devido à sua robustez em lidar com dados tabulares e alta dimensionalidade. O modelo foi ajustado utilizando o conjunto de dados completo de treinamento, e avaliamos seu desempenho utilizando a métrica MAPE. Na etapa de teste, os dados fornecidos para cada par cliente-servidor foram processados para extrair as características necessárias, replicando a lógica usada no treinamento. As predições foram feitas para as médias e desvios padrão das próximas medições. Os resultados foram organizados em um arquivo CSV no formato exigido pelo desafio, contendo as colunas id, mean_1, stdev_1, mea_2 e stdev_2.

3.3. Submissões e Melhorias

As submissões desenvolvidas neste trabalho tiveram sua validação realizada na plataforma Kaggle³, onde o desafio permanece aberto para competidores. Para cada submissão, realizamos alterações no código, buscando melhorias e redução nos valores do MAPE. Os resultados são exibidos como score na plataforma, e cada equipe pode realizar até 5 submissões por dia, permitindo verificar sua posição no ranking, bem como se superou ou não o resultado esperado da *baseline*. Até o momento, realizamos 21 submissões, embora nem todas as alterações realizadas em cada submissão tenham sido documentadas. Os parâmetros ajustados ao longo dessas tentativas estão apresentados na Tabela 2.

Table 2: Mudanças documentadas das submissões.

| Alteração | Parâmetros do RF | Mudanças | Resultado |
|----------------------------|------------------|--|--|
| 1 | n_estimators=200 | — | Ruim. Abaixo do baseline. Espaço para melhorias. |
| 2 | n_estimators=200 | Mudança na forma de obtenção dos dados do dataset. | Mediano. Acima do baseline. Espaço para melhorias. |
| Continua na próxima página | | | |

³<https://www.kaggle.com/competitions/open-data-challenge-ct-mon-rnp/overview>

Table 2 – Continuação da página anterior

| Alteração | Parâmetros do RF | Mudanças | Resultado |
|------------------|--|---|--|
| 3 | n_estimators=200, max_features=0.3, max_depth=20, min_samples_split=2, min_samples_leaf=10 | Utilizamos os parâmetros do RF da equipe Chameleon. | Bom. Acima do baseline. Espaço para melhorias. |
| 4 | n_estimators=200, max_features=0.3, max_depth=20, min_samples_split=2, min_samples_leaf=10 | Alteramos o formato da divisão de treino e teste nos dados de treino. | Bom. Acima do baseline. Espaço para melhorias. |
| 5 | n_estimators=200, max_features=0.3, max_depth=20, min_samples_split=2, min_samples_leaf=10 | Resolvemos um warning recebido da biblioteca NumPy no pré-processamento. | Bom. Acima do baseline. Espaço para melhorias. |
| 6 | n_estimators=200, max_features=0.3, max_depth=20, min_samples_split=2, min_samples_leaf=10 | Remoção do ruído nos dados. | Bom. Acima do baseline. Espaço para melhorias. |
| 7 | n_estimators=200, max_features=0.3, max_depth=20, min_samples_split=2, min_samples_leaf=10 | Utilizamos toda a base de treino para o treinamento do modelo, sem dividi-la. | Bom. Acima do baseline. Espaço para melhorias. |
| 8 | n_estimators=600, max_features=0.5, max_depth=20, min_samples_split=2, min_samples_leaf=20, random_state=42 | Alteramos os parâmetros do RF. | Excelente. Acima do baseline. Espaço para melhorias. |
| 9 | n_estimators=800, max_features=0.5, max_depth=20, min_samples_split=2, min_samples_leaf=20, random_state=42 | Alteramos os parâmetros do RF. | Mediano. Acima do baseline. Espaço para melhorias. |
| 10 | n_estimators=600, max_features=0.5, max_depth=20, min_samples_split=2, min_samples_leaf=20, random_state=42 | Fizemos alterações no pré-processamento. | Mediano. Acima do baseline. Espaço para melhorias. |

4. Resultados

Nesta seção, apresentamos os resultados obtidos a partir dos testes realizados durante o desenvolvimento deste trabalho. Os dados são organizados de forma a destacar o desempenho das diferentes amostras avaliadas, oferecendo uma visão clara do progresso em relação ao objetivo final. Os resultados são analisados e comparados ao benchmark estabelecido pelo desafio, com base no desempenho do modelo em relação ao *ranking* atual. A Tabela 3 apresenta os resultados das amostras avaliadas e nossa respectiva classificação qualitativa do desempenho. A avaliação foi baseada no *ranking* atual da competição, onde o melhor resultado registrado (*top 1*) é de 0.0802131562. Quanto mais próximo do *top 1* o valor obtido, melhor foi o desempenho do modelo avaliado. Assim, resultados que se aproximam de 0.0802131562 foram classificados como “Excelente”, enquanto valores ligeiramente superiores foram categorizados como “Bom”. Valores intermediários e superiores ao *baseline* receberam a classificação de “Mediano”, e aqueles significativamente distantes do benchmark foram considerados “Ruim”.

Table 3. Resultados dos testes realizados e sua respectiva avaliação.

| Amostra | Resultado | Nossa Avaliação do teste |
|---------|--------------|--------------------------|
| 1 | 0.0817763788 | Excelente |
| 2 | 0.0824439887 | Bom |
| 3 | 0.0829704432 | Bom |
| 4 | 0.0850721901 | Bom |
| 5 | 0.0944079488 | Mediano |
| 6 | 0.0950516130 | Mediano |
| 7 | 0.0988451337 | Ruim |
| 8 | 0.1009221255 | Ruim |
| 9 | 0.1010708254 | Ruim |
| 10 | 0.1041059489 | Ruim |

5. Conclusão

Este trabalho propôs uma abordagem baseada em aprendizado de máquina para o Desafio CT-Mon/RNP, com foco na previsão de métricas de redes dinâmicas de alto tráfego. Os resultados obtidos demonstraram a eficácia da estratégia proposta, consolidando o uso do modelo Random Forest como uma ferramenta robusta e eficiente para lidar com dados de alta dimensionalidade e séries temporais. A metodologia destacou-se por sua capacidade de operar em cenários com grandes volumes de informações, entregando resultados significativamente superiores ao baseline e próximos ao benchmark estabelecido. A etapa de pré-processamento desempenhou um papel fundamental, permitindo a organização e estruturação dos dados em janelas temporais, o que garantiu maior precisão nas previsões geradas. Adicionalmente, o uso da plataforma Kaggle viabilizou uma

avaliação dinâmica e contínua dos resultados, oferecendo percepções em tempo real que foram cruciais para os ajustes realizados durante o desenvolvimento. Esses aprimoramentos constantes refletiram-se no desempenho progressivo observado nas submissões. Os achados deste estudo contribuem para o avanço das estratégias de gerenciamento de redes acadêmicas e institucionais, apresentando soluções práticas e aplicáveis a cenários reais. Futuras pesquisas podem explorar a integração de modelos híbridos ou de redes neurais profundas, com o intuito de refinar ainda mais as previsões e ampliar as possibilidades de aplicação. Assim, este trabalho reforça a importância de abordagens analíticas no enfrentamento de desafios complexos, alinhando-se às demandas crescentes das infraestruturas de rede contemporâneas.

References

- [Chen and Chen 2022] Chen, Y. and Chen, X. M. (2022). A novel reinforced dynamic graph convolutional network model with data imputation for network-wide traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, 143:103820.
- [da Costa Côrtes et al. 2002] da Costa Côrtes, S., Porcaro, R. M., and Lifschitz, S. (2002). *Mineração de dados-funcionalidades, técnicas e abordagens*. PUC.
- [Domingos 2012] Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.
- [Duanmu et al. 2017] Duanmu, Z., Zeng, K., Ma, K., Rehman, A., and Wang, Z. (2017). A quality-of-experience index for streaming video. *IEEE Journal of Selected Topics in Signal Processing*, 11:154–166.
- [Jiawei and Micheline 2006] Jiawei, H. and Micheline, K. (2006). *Data mining: concepts and techniques*. Morgan kaufmann.
- [Kim and Kim 2016] Kim, S. and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3):669–679.
- [Mena 1999] Mena, J. (1999). *Data mining your website*. Digital Press.
- [Nguyen et al. 2023] Nguyen, T.-V., Nguyen, N. P., Kim, C., and Dao, N.-N. (2023). Intelligent aerial video streaming: Achievements and challenges. *Journal of Network and Computer Applications*, 211:103564.
- [Nielsen 2024] Nielsen (2024). Time spent streaming surges to over 40 percent in june 2024. Acesso em: 20 jan. 2025.
- [Papakyriakou and Barbounakis 2022] Papakyriakou, D. and Barbounakis, I. (2022). Data mining methods: A review. *International Journal of Computer Applications*.
- [Ramadhan et al. 2017] Ramadhan, M. M., Sitanggang, I. S., Nasution, F. R., and Ghifari, A. (2017). Parameter tuning in random forest based on grid search method for gender classification based on voice frequency. *DEStech transactions on computer science and engineering*, 10(2017).
- [Rigatti 2017] Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, 47(1):31–39.

- [Silva 2015] Silva, D. M. d. (2015). Avaliação de algoritmos baseados em aprendizagem de máquina para classificação de tráfego em redes sdn. Monografia de bacharelado, Universidade Federal do Rio Grande do Sul, Porto Alegre. Orientador: Prof. Dr. Alberto Egon Schaeffer-Filho.
- [Wiederhold and McCarthy 1992] Wiederhold, G. and McCarthy, J. (1992). Arthur samuel: Pioneer in machine learning. *IBM Journal of Research and Development*, 36(3):329–331.