UNIVERSIDADE DA INTEGRAÇÃO INTERNACIONAL DA LUSOFONIA AFRO-BRASILEIRA

INSTITUTO DE ENGENHARIAS E DESENVOLVIMENTO SUSTENTÁVEL

CURSO DE ENGENHARIA DE COMPUTAÇÃO

THIAGO QUEIROZ DA SILVA

ANÁLISE COMPARATIVA DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DOS CUSTOS DE MEDICAMENTOS PARA DOENÇAS REUMÁTICAS DO SUS

REDENÇÃO - CE

2025

THIAGO QUEIROZ DA SILVA

ANÁLISE COMPARATIVA DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DOS CUSTOS DE MEDICAMENTOS PARA DOENÇAS REUMÁTICAS DO SUS

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Engenharia de Computação, do Instituto de Engenharias e Desenvolvimento Sustentável da Universidade da Integração Internacional da Lusofonia Afro-Brasileira, como requisito parcial para a obtenção do Título de Engenheiro da Computação.

Orientador: Prof. Dr. John Hebert da Silva Felix.

REDENÇÃO - CE

2025

Universidade da Integração Internacional da Lusofonia Afro-Brasileira
Sistema de Bibliotecas da UNILAB
Catalogação de Publicação na Fonte.

Silva, Thiago Queiroz da.

S586a

Análise comparativa de algoritmos de aprendizagem de máquina para predição dos custos de medicamentos para doenças reumáticas do SUS / Thiago Queiroz da Silva. - Redenção, 2025.
54f: il.

Monografia - Curso de Engenharia De Computação, Instituto De Engenharias E Desenvolvimento Sustentável, Universidade da Integração Internacional da Lusofonia Afro-Brasileira, Redenção, 2025.

Orientador: Prof. Dr. John Hebert da Silva Felix.

1. Aprendizagem de máquina. 2. Predição de custos. 3. Doenças reumáticas. 4. Sistema Único de Saúde (SUS). I. Título

CE/UF/BSCA CDD 006.31

THIAGO QUEIROZ DA SILVA

ANÁLISE COMPARATIVA DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DOS CUSTOS DE MEDICAMENTOS PARA DOENÇAS REUMÁTICAS DO SUS

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Engenharia de Computação, do Instituto de Instituto de Engenharias e Desenvolvimento Sustentável da Universidade da Integração Internacional da Lusofonia Afro-Brasileira, como requisito parcial para a obtenção do Título de Engenheiro da Computação.

Aprovada em: 27/11/2025.

BANCA EXAMINADORA

_____
Prof. Dr. John Hebert da Silva Félix (Orientador)
Universidade da Integração Internacional da Lusofonia Afro-Brasileira (Unilab)

_____
Prof. Dr. Antonio Alisson Pessoa Guimarães
Universidade da Integração Internacional da Lusofonia Afro-Brasileira (Unilab)

_____
Prof. Dr. Antonio Carlos da Silva Barros
Universidade da Integração Internacional da Lusofonia Afro-Brasileira (Unilab)

Dedico este trabalho primeiramente a Deus e a todas as pessoas que de alguma forma contribuíram para sua realização. Em especial, a minha mãe Toinha, ao meu pai Gilvan e a minha irmã Thais por todo amor, apoio e carinho.

AGRADECIMENTOS

Agradeço primeiramente a Deus, porque se eu cheguei até aqui foi porque ele me deu forças, foi com a fé nele que pude ter paz, segurança e orientação.

Segundo a minha mãe e ao meu pai que são minha base e meu alicerce e estiveram comigo me apoiando em todos os momentos da minha graduação e da minha vida. Obrigado por todo apoio, cuidado e incentivo durante essa minha trajetória. Saiba que eu os amo acima de tudo e se sou o que sou hoje eu devo a minha base familiar e todo meu sucesso e conquista se deve também ao fato de eu estar buscando sempre o melhor para eles.

A minha irmã gêmea Thais, por toda parceria e união, em que toda a trajetória é a que mais fala palavras que me fazem não querer desistir, ela sempre está me colocando em um patamar de destaque que às vezes eu acho que não mereço, mas que ela afirma que estou.

Ao meu professor e orientador John Herbert, que embarcou comigo nessa jornada e afirmo o quanto ele merece meu sincero agradecimento pelo compromisso e por me guiar com firmeza, sou grato pela paciência que teve comigo.

A minha família, minha avó Maze e Maria e aos meus avôs Chico Liosmar e in memoriam Chico Coco, as minhas tias, aos meus tios, meus primos, minhas primas, a todos da comunidade onde eu moro, da igreja onde eu faço parte, da minha cidade Barreira.

A todo corpo docente do IEDS que passou nessa minha trajetória e aqui faço questão de citar todos, Márcia Roberta, José Cleiton, Silvia Helena, Tales Paiva, Ada Amélia, Cicero Saraiva, Alexandre Cunha, Nicolas Martins, Sergio Servilha, Antonio Carlos, Sabi Yari, Halisson, Carlos Caceres, Luiz Martins, Rejane, Manoel Ribeiro, Ligia, Olimpio, Allberson Bruno, Ranoyca, Gustavo Henn, Vandilberto, Luis Rigo, José Willamy, Herivelton, Jonas Costa, André Luis, vocês também fizeram parte dessa história e não seria justo eu não agradecer a vocês, sei que não fui o melhor aluno, talvez o pior, porém eu espero que, no futuro, vocês possam olhar para mim e ter orgulho em dizer: “Ele foi meu aluno.”

A todos os meus professores do Ensino Infantil, Fundamental e Médio por que se eu cheguei até aqui é por que passei nas mãos de gigantes e de uma boa base.

A todos meus amigos e aqui o texto é maior, por que se teve pessoas que aguentaram diariamente foram estes, desculpem pelo meu estresse e grosseria, por ter sido tão chato, por isso agradeço por cada palavra de apoio, paciência, risadas, conhecimentos, ajuda… eu não conseguiria ter chegado até onde eu cheguei se não fosse vocês. Estes foram minha rede de

apoio, companhia de estudos, motivo de risadas e força nos dias difíceis. Obrigado por dividirem tempo, conhecimento e coração, essa conquista também se deve a eles, porque transformaram meu cansaço em forças. Além disso, devo citar aqui o nome de dois - Átila e Leonardo - que nunca mediram esforços para me ajudar.

Agradeço a todos os grupos de pesquisa que fiz parte SMARTMETER, BND, PROGROBO e de forma especial ao GEPPAA, grupo de Astronomia que é liderado pelo professor Michel Lopes, este se tornou um grande amigo, agradeço por sempre estar disposto a ajudar e também obrigado a todos os colegas deste grupo que foram de muita importância para mim.

Agradeço a Unilab e ao meu Instituto o IEDS e aqui entra toda a equipe administrativa, também as pessoas que fazem parte para ela funcionar porteiros, faxineiras, pessoal do RU, motorista, técnicos, guardas, bibliotecárias, pessoal das chaves, as meninas da cantina, já que devo deixar claro que o trabalho de vocês sustenta nossos estudos com dignidade e acolhimento.

Gratidão a todos!

Esse trabalho é para o meu eu do passado que enfrentou todos os desafios e conseguiu chegar até aqui e para meu do eu do futuro que irá colher esses frutos.

“Senhor, Fazei-me um instrumento de vossa paz”

Oração de São Francisco

RESUMO

O grande crescimento do uso e gastos de biofármacos e limitação de recursos públicos para doenças reumáticas no SUS tornam os modelos de predição importantes para fornecer decisões de incorporação, negociação de preços e alocação de recursos no SUS, já que existem evidências de bilhões de reais em despesas públicas nos últimos anos. Este trabalho teve como objetivo realizar uma análise comparativa de algoritmos de aprendizado de máquina para predição de custos de medicamentos para doenças reumáticas no SUS. A metodologia envolveu a coleta e processamento de dados do DATASUS e do Banco de Preços em Saúde (BPS) referentes ao período de 2020 a 2024. Três algoritmos foram implementados em linguagem Python: Random Forest (RF), Multi-Layer Perceptron (MLP) e K-ésimo Vizinho mais Próximo (KNN), avaliados por meio das métricas R^2, RMSE (Raiz do Erro Quadrático Médio) e MAE (Erro Absoluto Médio). Os resultados indicaram que o Random Forest apresentou o melhor desempenho global, com R^2 de 0,89, RMSE de R$ 2.200,07 e MAE de R$ 947,70, demonstrando maior poder explicativos e menores erros de predição em comparação aos demais modelos. O KNN obteve desempenho intermediário com R^2 de 0,83, enquanto o MLP apresentou os resultados mais limitados com R^2 de 0,6626 e ambos algoritmos, alternaram conforme a faixa de custo considerada, com o KNN se destacando em alguns cenários e a MLP em outros, em relação ao RMSE e MAE. Conclui-se que o Random Forest é o algoritmo mais adequado para predição de custos de medicamentos no contexto do SUS com dados administrativos heterogêneos, evidenciando o potencial do aprendizado de máquina como ferramenta de apoio à gestão da Assistência Farmacêutica, com possibilidade de extensão dessa abordagem para outros grupos de medicamentos e condições clínicas.

Palavras-chave: Aprendizagem de Máquina; Predição de Custos; Doenças Reumáticas; SUS.

ABSTRACT

The significant growth in the use and spending of biopharmaceuticals, along with the limited public resources available for rheumatic diseases in the SUS, makes prediction models important tools for supporting decisions on technology incorporation, price negotiations, and resource allocation within the system, given the evidence of billions of reais in public expenditures in recent years. This study aimed to conduct a comparative analysis of machine learning algorithms for predicting medication costs for rheumatic diseases in the SUS. The methodology involved collecting and processing data from DATASUS and the Health Price Database (BPS) for the period from 2020 to 2024. Three algorithms were implemented in Python: Random Forest (RF), Multi-Layer Perceptron (MLP), and K-Nearest Neighbors (KNN), and evaluated using R^2, RMSE (Root Mean Square Error), and MAE (Mean Absolute Error). The results indicated that Random Forest presented the best overall performance, with an R^2 of 0.89, RMSE of R$ 2,200.07, and MAE of R$ 947.70, demonstrating greater explanatory power and lower prediction errors compared to the other models. KNN achieved intermediate performance with an R^2 of 0.83, while MLP showed the most limited results with an R^2 of 0.6626. Both algorithms alternated in performance depending on the cost range considered, with KNN performing better in some scenarios and MLP in others in terms of RMSE and MAE. It is concluded that Random Forest is the most suitable algorithm for predicting medication costs in the SUS context with heterogeneous administrative data, highlighting the potential of machine learning as a support tool for Pharmaceutical Assistance management, with the possibility of extending this approach to other groups of medications and clinical conditions.

Keywords: Machine Learning; Cost Prediction; Rheumatic Diseases; SUS.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AF	Assistência Farmacêutica
AM	Aprendizado de Máquina
API	Application Programming Interface
BPS	Banco de Preços em Saúde
CEAF	Componente Especializado da Assistência Farmacêutica
CID	Classificação Internacional de Doenças
CLI	Interface de Linha de Comando
CNPJ	Cadastro Nacional da Pessoa Jurídica
CONITEC	Comissão Nacional de Incorporação de Tecnologias no SUS
DATASUS	Departamento de Informática do Sistema Único de Saúde
GPU	Unidade de Processamento Gráfico
KNN	K-ésimo Vizinho mais Próximo
IA	Inteligência Artificial
IoT	Internet das Coisas
MAE	Erro Absoluto Médio
ML	Machine Learning
MLP	Multi-Layer Perceptrons
OMS	Organização Mundial da Saúde
R^2	Coeficiente de Determinação
RF	Random Forest
RMSE	Raiz do Erro Quadrático Médio
SIA	Sistema de Informações Ambulatoriais
SIH	Sistema de Informações Hospitalares
SIM	Sistema de Informações sobre Mortalidade
SINASC	Sistema de Informações sobre Nascidos Vivos
SUS	Sistema Único de Saúde
TPU	Unidade de Processamento Tensorial
VS Code	Visual Studio Code

SUMÁRIO

1. INTRODUÇÃO

O Sistema Único de Saúde (SUS) é um dos maiores sistemas públicos de saúde do mundo, que atende e presta assistência a 215 milhões de brasileiros. No Brasil, 54% do gasto em saúde acontece no setor privado, que atende a apenas 25% da população. O SUS, é exclusivamente responsável por 75% da população, além de realizar serviços voltados para toda a sociedade, conta com apenas 46% dos recursos (Campos, 2018). Esses recursos vêm de impostos e contribuições sociais pagos pela própria sociedade e são repartidos entre União, estados, Distrito Federal e municípios por regras constitucionais específicas.

Nisso entende-se o desafio de viabilizar a expansão do acesso e equilibrar a inovação com os recursos financeiros disponíveis, já que entre os principais fatores críticos, tem um destaque para o crescimento em forma exponencial de gastos com medicações de alto custo, e em particular as destinadas ao tratamento de doenças reumáticas crônicas.

Vale destacar que o acesso insuficiente aos medicamentos está diretamente associado com a piora do estado de saúde, maior uso de terapias adicionais, aumento no número de retornos aos serviços de saúde e gastos adicionais nos tratamentos (Boing *et al.,* 2013). Logo, essas condições que atingem mais de 15 milhões de brasileiros, de qualquer idade, requerem terapias biológicas e imunossupressoras de elevado custo, pressionam ainda mais a sustentabilidade do sistema.

Os medicamentos para doenças reumáticas, especialmente os biofármacos, representam um grande desafio para o orçamento do SUS. Embora garanta um ganho significativo na qualidade de vida e no controle da progressão das doenças, os gastos públicos somaram cerca de R$28 bilhões de reais entre 2012 e 2019 (Rodrigues Filho; Perreira, 2022). Tal informação sinaliza o quanto surge a necessidade de ferramentas que permitam um planejamento e uma previsão apropriada destes gastos para os pacientes, viável prever o gasto farmacêutico com uma razoável acurácia (Linnér *et al.,* 2020).

A capacidade de prever com precisão os custos futuros com medicamentos serve como base para entender o rápido aumento nos custos de compras de produtos farmacêuticos e dessa maneira contribui para fornecer cuidados de saúde equitativos e de qualidade onde há uma necessidade profunda de entender melhor os impulsionadores dos preços dos produtos farmacêuticos (Fazekas; Veljanov; Oliveira, 2024).

A aprendizagem de máquina tem apresentado um desempenho notável em várias áreas de aplicações na área da saúde, porém a aplicação para a predição de custos de medicamentos no SUS ainda é minimamente explorada, com poucos estudos de comparação

sistemática, para validar a aplicação na prática da sua funcionalidade.

O presente estudo faz uma análise de comparação dos algoritmos de Random Forest Regressor, Rede neural MLP (perceptron multicamada) e K-ésimo Vizinho mais Próximo (KNN) para predição dos custos de medicamentos designadas para o tratamento de doenças reumáticas. Estudos recentes têm explorado intensamente dados administrativos e de compras públicas para compreender determinantes de preços de medicamentos e dos custos em saúde, no contexto brasileiro, Kohler et al. (2015) analisou a base Banco de Preços em Saúde (BPS) e mostrou, por meio de modelos de regressão linear, que o aumento da transparência nos preços, por si só, levou a reduções nos valores pagos de medicamentos amplamente utilizados. Ampliando, Fazekas et al. (2024) utilizaram mais de 200 mil registros de compras públicas de medicamentos em oito países latino-americanos e dois estados brasileiros para modelar preços unitários e comparar regressão linear com Random Forest, demonstrando que o modelo apresenta maior poder explicativo e menor erro de previsão. Por fim, Nelson & Arbeeva (2022) ressalta que houve rápida expansão de IA e ML em doenças reumáticas e musculoesqueléticas, mas que muitos clínicos ainda enfrentam dificuldades com os conceitos e a interpretação dos modelos.

Com base no contexto apresentado, este trabalho tem como objetivo desenvolver e treinar três algoritmos de aprendizado de máquina, para verificar qual possui a maior eficácia, no processamento de dados históricos de casos de doenças reumáticas no SUS, através de métricas estatísticas apropriadas e por fim trazer uma avaliação sobre a prática de cada algoritmo.

Espera-se que os resultados obtidos neste trabalho, possam contribuir para o aperfeiçoamento da administração de recursos no SUS, já que ocorre a crescente demanda por serviços de saúde no SUS, aliada à limitação de recursos financeiros e estruturais, e isso exige ferramentas analíticas baseadas em dados para otimizar a alocação de leitos, equipamentos e equipes médicas, tornando a gestão mais eficiente e econômica.

E por fim, a metodologia e os achados deste estudo podem ser ajustados para outros tipos de medicamentos, expandindo o impacto da pesquisa na gestão do sistema público de saúde brasileiro.

Espera-se que a aplicação de modelos preditivos baseados em Machine Learning permita identificar padrões de demanda hospitalar com precisão superior a métodos tradicionais, possibilitando alocação mais eficiente de recursos e redução de custos operacionais.

2. OBJETIVO

2.1. Objetivo Geral

Comparar algoritmos de aprendizado de máquina na predição de custos de medicamentos para doenças reumáticas no SUS, com dados que foram coletados do DATASUS e do BPS (Banco de Preços em Saúde), avaliando o desempenho com métricas de avaliação para saber o algoritmo mais apropriado no contexto analisado.

2.2. Objetivos Específicos

- Fazer o levantamento de dados dos últimos 5 anos do DATASUS via PySUS e dados de valores do Banco de Preços em Saúde, organizando um dataset com variáveis para detecção de casos de pacientes reumáticos.
- Realizar pré-processamento dos dados obtidos para modelagem e tratamento de variáveis categóricas e numéricas.
- Desenvolver, treinar e comparar três modelos de aprendizado de máquina e assim monitorar os seus desempenhos.
- Verificar o desempenho dos modelos com tabelas e gráficos, demonstrando qual modelo é superior, qual é o intermediário e qual é inadequado para o conjunto de dados disponível.
- Discutir sobre implicações práticas, indicando a abordagem como apoio ao planejamento orçamentário e negociação de preços, além de discutir limitações e caminhos futuros de análise a outros medicamentos.

3. REFERENCIAL TEORICO

3.1. Sistema Único de Saúde (SUS)

O Sistema Único de Saúde (SUS) é o sistema público de saúde brasileiro, reservado a serviços de saúde sem intolerância, englobando desde ações de promoção e prevenção até atenção de baixa, média e alta complexidade. Sendo inspirado em valores como igualdade, democracia e emancipação, o SUS está inserido na Constituição, na legislação ordinária e em normas técnicas e administrativas (Paim, 2018).

Conforme Pontes, Oliveira e Gomes (2014), deve-se considerar que o SUS foi criado com base no princípio da universalidade do cuidado, no qual a saúde é tomada como direito de todos e dever do Estado. Além de cuidado clínico, detém responsabilidade na vigilância epidemiológica, sanitária, assistência farmacêutica, e também promove a inclusão de ações e programas do governo para o bem-estar da população e de educação em saúde.

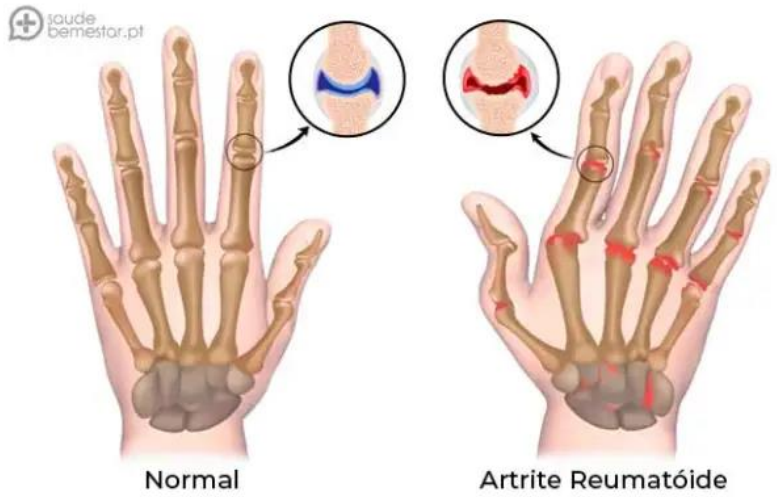
O SUS busca atender toda a população de forma persistente e ordenada, com um foco principal em uma melhor qualidade de vida e redução das desigualdades em saúde. Além disso, possui uma das maiores estruturas e mais complexos sistemas públicos de saúde do mundo, que vão desde simples atendimentos até procedimentos, sendo disponibilizados serviços de alta complexidade, garantindo atendimento em zonas urbanas e remotas do país.

3.2. Doenças Reumáticas

As doenças reumáticas englobam um conjunto de distúrbios que acometem principalmente o sistema musculoesquelético, incluindo articulações, ossos, músculos e tecido conjuntivo (Koo; Lu, 2023).

Dessa forma, o reumatismo representa um conjunto de diferentes doenças que acometem o aparelho locomotor, como, ossos, articulações, cartilagens, músculos, tendões e ligamentos. Estas podem comprometer outras partes e funções do corpo humano, como rins, coração, pulmões, olhos, intestino e até a pele (Brasil, 2013), conforme mostrado na Figura 1.

Figura 1 – Imagem definindo Artrite Reumatoide.



Fonte: Redação SANAR (2025)

Se torna necessário entender essas condições que atingem as limitações de movimento já que essas doenças impactam a qualidade de vida, sendo importante compreender como tratá-las e preveni-las de forma adequada, visto que os sintomas estão entre a a dificuldade para se movimentar ou enrijecimento das articulações, além do recuo da flexibilidade da coluna, limitação em algumas atividade como pentear os cabelos e escovar os dentes, além de dor, inchaço e calor nas articulações (Smolen *et al.,* 2018).

Em outro contexto, as doenças reumáticas e músculoesqueléticas são uma das principais causas de incapacidade no mundo e constituem um enorme encargo socioeconômico (Al Maini *et al.,* 2020). Tais doenças ocorrem em crianças, jovens e adultos podendo atingir pessoas de qualquer idade, em suma estas não são dependentes de cor, sexo ou idade, esse grupo de doenças não é transmissível, não contagioso e acompanhado de dor (Brasil, 2013).

A prevenção e o cuidado são fundamentais visto que algumas doenças reumáticas são crônicas e podem resultar em incapacidade funcional e problemas em órgãos internos caso não sejam controladas. Com o conhecimento se torna provável a detecção de sintomas de forma precoce e uma busca de tratamento médico mais indicado, na qual serão evitados agravamentos.

Ademais, as doenças reumáticas acarretam um considerável impacto em termos físicos, psicológicos e sociais para os pacientes, pertinente a utilização de medidas multidimensionais de qualidade de vida (Oliveira et al., 2009).

Assim ocorre uma necessidade de uma avaliação do impacto da doença reumática na vida do paciente, a fim de contribuir para o acesso aos tratamentos necessários. Uma vez que no estudo de Carga Global de Doenças da Organização Mundial da Saúde (OMS) de 2010, as doenças reumáticas e músculoesqueléticas foram a segunda principal causa de incapacidade

no mundo, medida em anos vividos com incapacidade (Al Maini *et al.*, 2015).

A significativa repercussão econômica das enfermidades reumáticas, em especial relacionadas aos medicamentos para o tratamento torna-se pertinente, a aplicação de recursos, como algoritmos de aprendizado de máquina, para predizer e assim otimizar os custos ao tratamento.

3.3. SUS e Medicamentos

A influência do Sistema Único de Saúde (SUS) na distribuição de medicamentos, de forma principal no contexto das doenças reumáticas, possui desafios e avanços referentes à gestão pública da saúde no Brasil.

Destaca-se que o acesso de medicamentos pelo SUS mostra-se um instrumento de combate às desigualdades e injustiças que resultam da sobreposição de múltiplas formas de discriminação ou desvantagem social, corroborando a ideia de que o SUS é uma política pública eficiente para promover justiça social (Mujica; Bastos; Boing, 2024).

A Política Nacional de Assistência Farmacêutica (AF) estabelece diretrizes claras para o fornecimento desses medicamentos, buscando assegurar o direito à saúde e à justiça distributiva. Essas políticas estabeleceram o acesso gratuito a medicamentos essenciais como direito dos cidadãos brasileiros e efetivaram a AF como política pública de saúde (Mujica; Bastos; Boing, 2024).

Conforme Tavares (2016), o acesso a medicamentos para tratamento de doenças crônicas ocorre para parcela considerável da população brasileira, especialmente para os mais pobres, indicando diminuição das desigualdades socioeconômicas, mas com diferenças entre regiões e entre algumas classes de medicamentos.

O financiamento dos medicamentos no SUS é estruturado por fontes diversas, incluindo orçamento federal, estadual e municipal, o que implica numa complexa responsabilidade entre União, estados e municípios. Esta divisão acaba exigindo gerenciamento efetivo para que ocorra uma garantia de uma sustentabilidade financeira do sistema, assim evitando rupturas no abastecimento. Dessa forma, dado o impacto financeiro da adoção de novas tecnologias em saúde, decidir o que incorporar e quando fazê-lo é um grande desafio (Lima; Brito; Andrade, 2019).

A aquisição de medicamentos representa uma parcela importante dos gastos totais em saúde e a disponibilidade de medicamentos de alto preço tem grande impacto social e relevância terapêutica (Fatel et al., 2021).

Vale destacar que é no CEAF (Componente Especializado da Assistência Farmacêutica) que são disponibilizados os medicamentos de uso ambulatorial de maior preço médio no SUS, incluindo os mais recentemente incorporados pela CONITEC (Comissão Nacional de Incorporação de Tecnologias no Sistema Único de Saúde), como o eculizumabe (Rover *et al.,* 2021).

Em suma, a gestão desses recursos exige atenção, diante do envelhecimento da população e do aumento de tratamentos inovadores, para que ocorra a garantia do acesso universal e a aplicação de tecnologias avançadas.

3.4. Aprendizado de Máquina – Definição

Aprendizado de Máquina (AM) corresponde no desenvolvimento de algoritmos com a capacidade de desenvolver padrões, sem precisar de serem programados para tarefas específicas. As técnicas AM permitem que o computador aprenda com exemplos, ou seja, aprenda por meio de dados (Ludemir, 2019).

Tais algoritmos elaboram modelos preditivos por meio de dados de entrada, em que tentam identificar associações e tendências que possibilitam fazer previsões ou decisões de forma automatizada, tentando minimizar a interferência humana. Segundo Bi et al. (2019), o AM enfatiza a precisão preditiva em vez da inferência dirigida por hipóteses, geralmente lidando com conjuntos de dados grandes e de alta dimensionalidade.

Destaca-se que o processo envolve algumas etapas como a preparação dos dados, o treinamento e também a avaliação de sua performance visando garantir resultados precisos, muito utilizado em áreas como a análise preditiva de custos e tratamento de doenças. Com isso o AM trata de como construir algoritmos que melhore automaticamente por meio da experiência, definida como o processo de melhorar alguma medida de desempenho ao executar uma tarefa, por meio de algum tipo de experiência de treinamento (Jordan; Mitchell, 2015).

3.5. Aplicações De Um Algoritmo De Aprendizado De Máquina

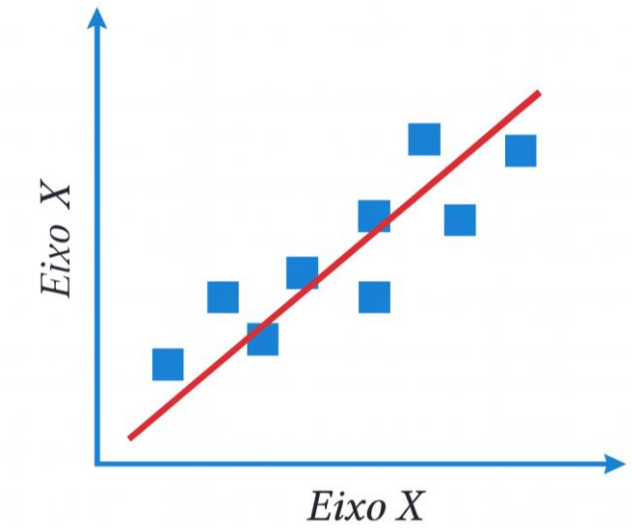
O AM é aplicado hoje em vários setores, incluindo saúde, finanças, agricultura, transporte, energia, varejo, cibersegurança, educação, cidades inteligentes, IoT (Internet das Coisas) e monitoramento ambiental (Sarker, 2021). Além disso, conforme Singh e Gupta (2025), este se destaca nas aplicações em diagnóstico médico e descoberta de fármacos,

detecção de fraude e avaliação de risco, previsão de produtividade agrícola, gestão de tráfego e veículos autônomos, otimização de sistemas de energia, recomendações em comércio eletrônico, detecção de ameaças cibernéticas e personalização de processos educacionais. Em resumo, os algoritmos de AM enquadram-se em uma das áreas de IA cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática (Monard; Baranauskas, 2016).

As tarefas são classificadas em agrupamento de dados, previsão de séries temporais e outras (Cerri, 2017). Na prática, escolhe-se conforme o problema e os dados, que são divididos entre supervisionado, não supervisionados e por reforço. O aprendizado de máquina supervisionado é a busca por algoritmos que raciocinam a partir de instâncias fornecidas externamente para produzir hipóteses gerais (Kotsiantis, 2007).

Estes são treinados com dados rotulados, em que cada entrada é interligada a uma saída que assim se espera. O algoritmo busca aprender com exemplos conhecidos para fazer previsões ou classificações com base em novos dados. Um exemplo de aprendizado de máquina supervisionado é mostrado na Figura 2. Na qual é possível visualizar um exemplo, cada ponto é um par x_i, y_i. O algoritmo escolhe os parâmetros *a* (inclinação) e *b* (intercepto) que diminuem o erro entre os valores que são previstos pela reta e os valores reais dos pontos.

Figura 2 – Ilustração do aprendizado de Machine Learning Supervisionado



Fonte: DataV - Educação Tech (2025)

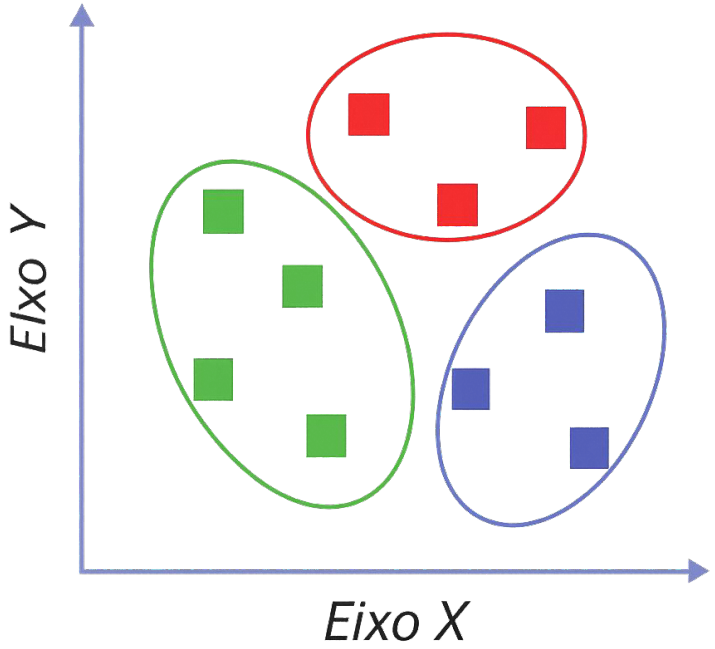
Os algoritmos de aprendizagem não-supervisionada, não é atribuído um rótulo para os dados de saída, este busca padrões e similaridades entre os dados, permitindo identificar

grupos de itens similares (Fontana, 2020).

Neste método, os dados não têm classificações predefinidas, permitindo que o algoritmo descubra padrões e estruturas de modo próprio. Este tenta compreender a organização dos dados de forma independente executando atividades como agrupamento, sendo um processo que une elementos equivalentes de acordo com suas características, a fim de facilitar a identificação de conjuntos diferentes na base de dados e análise de componentes principais, em que essa técnica se baseia em diminuir a complexidade dos dados preservando as informações mais significativas, tornando mais simples a interpretação dos padrões existentes.

Um exemplo de aprendizado de máquina não supervisionado é apresentado na Figura 3. Nesta é possível perceber, que cada quadradinho é um dado $x_i y_i$ e as elipses coloridas mostram grupos formados por similaridade.

Figura 3 – Ilustração do aprendizado de Machine Learning Não Supervisionado



Fonte: DataV - Educação Tech (2025)

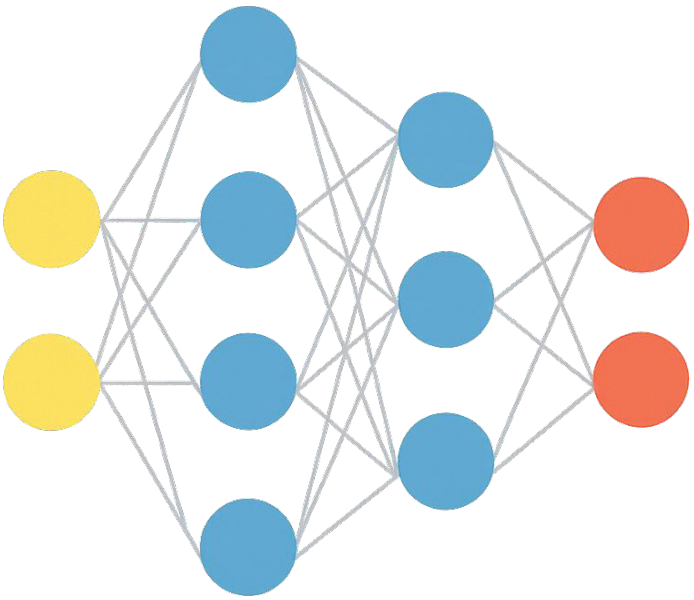
Nos algoritmos por reforço, um agente realiza uma ação em um ambiente e recebe uma recompensa de acordo com o resultado dessa ação, tendo o objetivo do algoritmo receber a maior recompensa possível (Fontana, 2020).

Estes interagem em um ambiente dinâmico, aprendendo a fazer escolhas sequenciais para maximizar recompensas de maneira progressiva. Além disso, exploram esse ambiente, aprendendo por tentativa e erro, recebendo *feedback* na forma de recompensas.

Um exemplo de aprendizado de máquina por reforço é mostrado na Figura 4.

Representando um estado do ambiente, na qual é escolhida uma ação, recebe uma recompensa e atualiza. Assim, entender as aplicações de cada tipo é de grande fundamento para que seja explorado todo o potencial desse estudo e poder impulsionar inovações futuras.

Figura 4 – Ilustração do aprendizado de Machine Learning Por Reforço



Fonte: DataV - Educação Tech (2025)

3.6. Algoritmos De Aprendizado De Máquina Do Estudo

3.6.1. Random Forest

O Random Forest (RF) é um algoritmo de aprendizado que utiliza múltiplas árvores, isto é, o algoritmo utiliza múltiplas árvores de decisão para realizar predições mais robustas e precisas, de regressão e classificação (Schonlau; Zou, 2020).

Com isso, as árvores individuais são construídas com base em amostras bootstrap, que é conjuntos de dados gerados a partir da própria amostra original, por meio de re-amostragem com reposição, sendo assim a construção dela não ocorre na amostra original, chamado de agregação bootstrap ou simplesmente bagging (Schonlau; Zou, 2020).

Dessa forma, cria-se vários subconjuntos aleatórios do conjunto original, para que ocorra o treinamento de uma árvore de decisão independente para cada subconjunto. Para obter a predição final, utiliza-se da agregação das saídas individuais de todas as árvores.

Para a compreensão deve-se entender o funcionamento do algoritmo no qual o processo inicia-se com bootstrap, em seguida o algoritmo cria múltiplas amostras aleatórias da

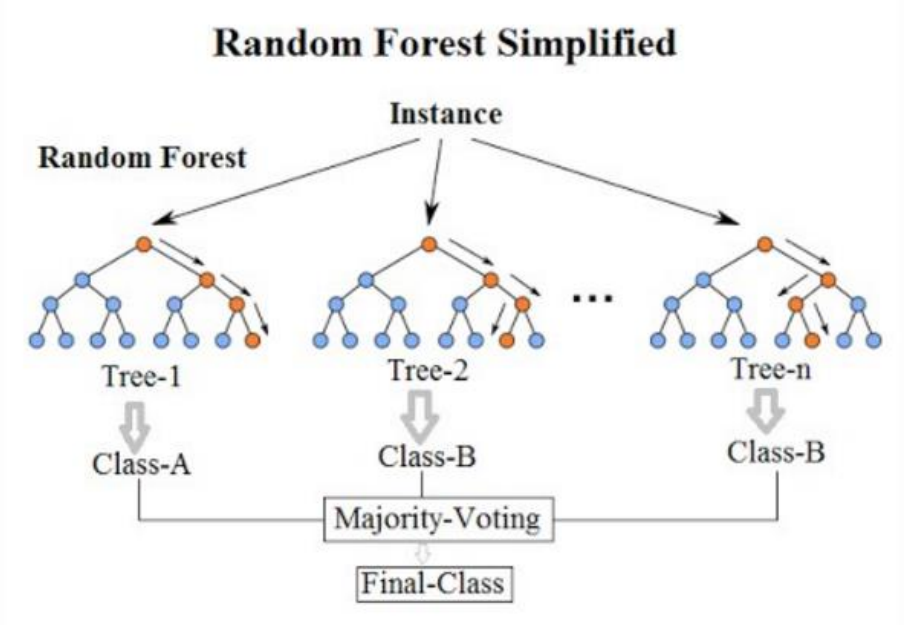
base de dados original com reposição. Isso significa que um mesmo registro pode surgir múltiplas vezes na mesma amostra, para que cada árvore seja treinada com diferentes dados. Essa diversidade de amostras melhora a generalização do modelo.

Para cada bootstrap, o algoritmo desenvolve uma árvore de decisão. Um modelo baseado em árvore envolve o particionamento recursivo do conjunto de dados fornecido em dois grupos com base em um determinado critério até que uma condição de parada predeterminada seja atendida (Schonlau, 2020).

Durante a construção de cada nó da árvore, um subconjunto aleatório é considerado para estabelecer a melhor divisão. Essa seleção aleatória introduz variância entre as árvores, de modo que ocorra a redução da correlação entre elas. As árvores expandem-se até sua profundidade máxima, desse modo permite que se tornem altamente especificadas e se adaptem aos dados que receberam.

A decisão final de classificação é tomada pela média e cada árvore vota para uma associação de classe (Schonlau, 2020). O algoritmo realiza predições para gerar novas amostras consultando cada árvore independente. Cada árvore "vota" na classe que entende como apropriada, e a classe com maior número de votos é retornada como predição final. Na Figura 5 é exemplificado como funciona o algoritmo RF, a partir de uma floresta de árvores de decisão.

Figura 5 – Funcionamento de um algoritmo Random Forest



Fonte: Will Koehrsen (2025)

Cada árvore, na Figura 5, oferece um palpite, a floresta adiciona esses palpites através de uma votação e entrega a classe final. Em vez de uma árvore grande, é criado *n* árvores independentes.

Dessa forma, cada árvore possui um treinamento com uma amostra aleatória com devolução dos dados de treino, com isso as árvores acabam vendo conjuntos diferentes. Por fim, ao decidir o melhor corte em um nó, esta avalia não apenas um subconjunto de variáveis aleatórias.

3.6.2. MLP – Multi-Layer Perceptrons

O Perceptron Multicamadas é um algoritmo de aprendizado composto por três ou mais camadas, sendo uma de entrada, uma ou mais camadas ocultas e uma camada de saída. Uma generalização do perceptron possui conexões entre todos os neurônios de uma camada e todos os neurônios da camada seguinte, isso dá origem ao chamado perceptron multicamadas (Murtagh, 1991).

Tal algoritmo é utilizado para tarefas de classificação e regressão. Esse sistema de neurônios (nós) simples e interconectados acaba representando um mapeamento não linear entre um vetor de entrada e um vetor de saída (Gardner; Dorling, 1998).

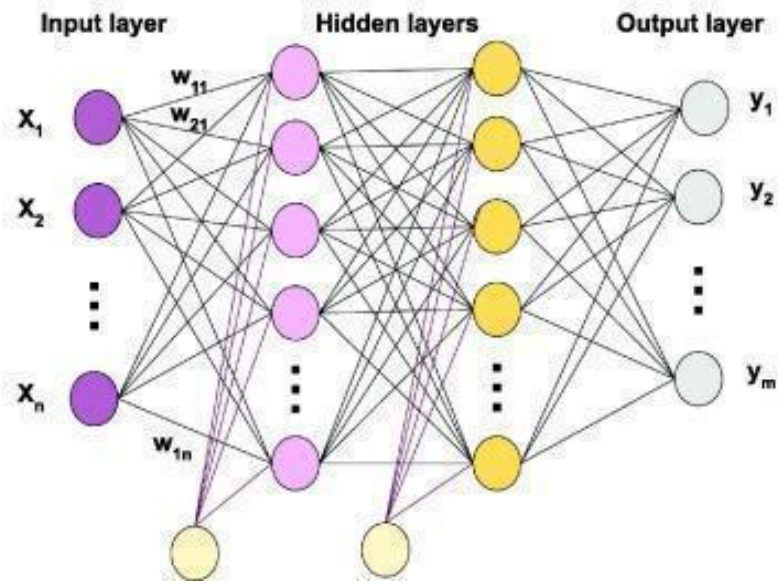
A primeira camada é a de entrada dos dados. Na qual cada nó representa uma característica do dado. A segunda camada de dados ocultos para poder captar padrões não lineares. Nesta camada ocorre o recebimento de entradas de todos os nó da camada anterior e transfere sua saída para os nó da próxima camada que é a camada de saída com a resposta final, podendo ser valores para regressão ou para classificação, conforme é apresentado na Figura 6.

O funcionamento do algoritmo ocorre em dois processos principais: propagação para frente e retropropagação do erro. A operação da rede consiste em começar inserindo o vetor na primeira camada; os sinais são então transmitidos camada a camada até que as unidades de saída emitam y′, a estimativa da saída desejada (Hecht-Nielsen, 1989).

Os dados de entrada são multiplicados por pesos e assim somados em cada nó, junto com um termo de viés. Assim o nó processa esse valor através de uma função de ativação, causando a não linearidade. As saídas da camada anterior se tornam entradas para a seguinte, até a saída final na camada. Primeiro, calcula-se o erro entre a saída da rede e o valor real por meio de uma função de perda. Em seguida, aplica-se o algoritmo de backpropagation, isto é, a

retropropagação do erro, conhecida como modo reverso da diferenciação automática. Nessa etapa, os gradientes são calculados com custo computacional essencialmente igual ao da propagação direta das ativações (Schmidhuber, 2015). De acordo com o que é mostrado na Figura 6 abaixo.

Figura 6 – Explicação de um algoritmo MLP



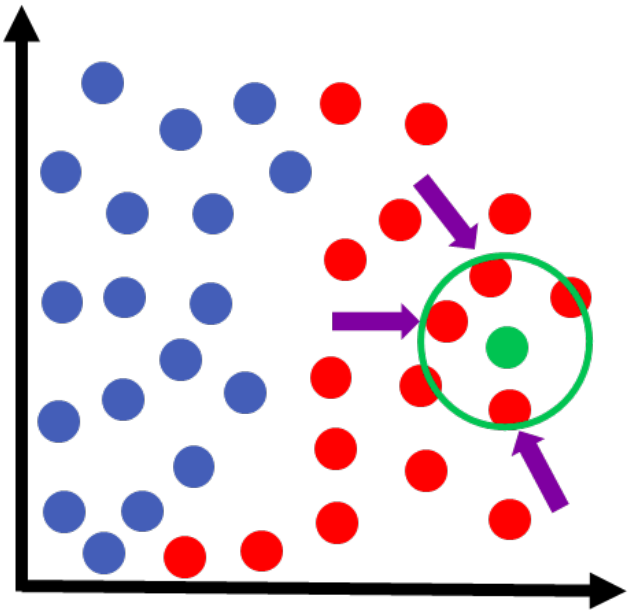
Fonte: ScienceDirect (2025)

Na Figura 6, cada nó é uma entrada e cada linha é um peso, no meio as camadas ocultas com nós que fazem soma ponderada com a ativação. Por último as saídas adequadas ao tipo de problema.

3.6.3. KNN

O algoritmo KNN é um método de aprendizado de máquina supervisionado utilizado principalmente para problemas de classificação e regressão para prever a saída de um novo exemplo (Zhang, 2016). Para prever, o KNN olha para os k exemplos mais parecidos no conjunto de treino e usa o que eles são para decidir, conforme mostrado na figura 07 abaixo.

Figura 7 – Explicação de um algoritmo KNN



Fonte: GeeksforGeeks (2025)

O KNN é chamado de método baseado em instâncias. Conforme Uddin et al. (2022) o algoritmo armazena os dados de treino, incluindo as entradas (características) e os rótulos (classes ou valores), quando um novo ponto precisa ser classificado ou ter seu valor estimado, o algoritmo calcula a distância entre este ponto e todos os pontos do conjunto de treino. Dessa forma, as distâncias mais usadas são a euclidiana, que corresponde à ideia de distância de em várias dimensões, e a distância de Manhattan, que soma os valores absolutos das diferenças em cada atributo.

Depois de calcular essas distâncias, o KNN ordena os exemplos de treino do mais próximo para o mais distante e escolhe os k primeiros. Esses k vizinhos mais próximos são então usados para fazer a predição (Parry et al., 2010). Com isso, em algumas situações, pode-se ponderar esses votos pela distância, dando mais peso aos vizinhos mais próximos. Em regressão, em vez de votação, faz-se a média dos valores associados aos vizinhos.

4. METODOLOGIA

4.1. Pré-Processamento de Dados em Aprendizado de Máquina

Antes de aplicar algoritmos de aprendizado de máquina em bases de saúde, o pré-processamento é necessário para garantir que os dados sejam adequados ao aprendizado. O pré-processamento é essencial para melhorar o desempenho de tarefas de classificação ou predição (Kale; Pandey, 2024). Isso constitui uma etapa de enorme fundamento já que garante uma qualidade e consistência dos dados utilizados.

Em essência, o pré-processamento de dados engloba uma série de procedimentos destinados a aumentar a adequação do conjunto de dados para o aprendizado de máquina (Chahid; Elmiad; Badaoui, 2023).

Por isso se torna necessário fazer codificação de variáveis categóricas: UF (estado), fabricante (responsável pela fabricação), fornecedor (quem fornece o medicamento).

Para a etapa de limpeza e preparação dos dados, inicialmente foram removidas variáveis/colunas consideradas não relevantes para a modelagem preditiva (Código Compra, Descrição Catmat, Unidade Fornecimento, CNPJ Fabricante, Fabricante, CNPJ Fornecedor, Fornecedor, CNPJ Comprador, Nome Instituição, Nome Município, Observações, Seq. Compra Item, Data Homologação, codigo_uf, Tratamento de dados desbalanceados, Agregação temporal dos dados) e agregação temporal (mês/ano).

Em seguida, procedeu-se ao tratamento de valores ausentes: para as variáveis oriundas do SIH-SUS, referentes ao número de casos reumatológicos por unidade federativa e mês e ao gasto total associado, os valores faltantes foram dados com zero. Posteriormente, linhas contendo valores ausentes em outras variáveis foram excluídas, garantindo a consistência das observações utilizadas no treinamento dos modelos.

As planilhas foram modeladas com One-Hot Encoding para variáveis categóricas, a codificação é definida como um vetor preenchido apenas com zeros e uns na posição atribuída ao caractere (He; Chua, 2017).

O RF divide os dados em dois conjuntos 80% para treinamento e 20% para teste, sem um conjunto de validação separado. Enquanto o algoritmo MLP, segue com 40% para teste, e 60% para treinamento. Já o algoritmo KNN, faz a seguinte implementação com 40% dos dados reservados para teste, e os 60% finais utilizados para treinamento. A divisão acontece em duas etapas com o primeiro separa-se o conjunto de teste do restante, e depois o conjunto de validação é extraído do que sobrou.

4.2. Ferramentas Usadas Nesta Pesquisa

As bases de dados em saúde utilizadas foram o DATASUS, ele sendo um repositório que reúne informações sobre mortalidade, internações, registros de doenças, assistência à saúde, dados financeiros, demografia, entre outras informações referentes a área. Dentro do DATASUS foi utilizado os dados do SIA (Sistema de Informações Ambulatoriais)/SUS e SIH (Sistema de Informações Hospitalares)/SUS para os dados ambulatoriais e hospitalares.

Depois foi baixado as informações dos medicamentos BPS, para análises econômicas ligadas ao custo de insumos no SUS.

A implementação envolveu a utilização de ferramentas computacionais, bibliotecas específicas e *frameworks* de código aberto para a criação dos algoritmos. A linguagem de programação adotada foi Python, esta é a linguagem mais preferida para computação científica, ciência de dados e AM, aumentando desempenho e produtividade ao permitir o uso de bibliotecas de baixo nível e APIs de alto nível (Raschka, 2020).

O Pandas é uma biblioteca usada para manipulação de dados tabulares, para as operações de leitura e conversão de vários arquivos em Planilha de Excel (.xlsx e .xls). Para as operações numéricas de alta performance, foi utilizado NumPy, uma biblioteca que dá suporte para arrays multidimensionais e funções matemáticas vetorizadas para os cálculos estatísticos e transformações logarítmicas aplicadas às variáveis-alvo. A biblioteca Glob para identificação de arquivos, o que facilitou a análise de múltiplas fontes de dados de medicamentos de doenças reumáticas.

A biblioteca Scikit-learn foi usada para implementação dos algoritmos para fornecer módulos para todas as etapas do *pipeline* de Machine Learning (ML). O Scikit-learn tornou-se o padrão da indústria em Python para engenharia de atributos e modelagem de ML clássico em conjuntos de dados (Raschka, 2020).

Para as arquiteturas foi usado TensorFlow para fornecer interface python, Keras como parte do TensorFlow. Este é uma biblioteca de código usado em AM, utilizado também Seaborn para gráficos estatísticos elegantes, Pickle para serialização de objetos Python, *OS* para manipulação de caminhos e diretórios, *Warnings* para filtrar avisos em múltiplos scripts e Datetime para manipulação de datas.

A biblioteca PySUS foi usada para automatização de dados do Sistema de Informações Hospitalares do DATASUS, o que permitiu o *download* de internações hospitalares filtradas através de estado, ano, mês e grupo de procedimentos. Portanto ressalta que utiliza-se duas interfaces oferecidas aos usuários: TabNet e TabWin. Já para a leitura dos

arquivos Parquet utilizou-se as bibliotecas PyArrow e FastParquet, para o armazenamento de colunas, otimizando o enorme volume de dados de epidemiologia.

A biblioteca Matplotlib foi utilizada para gerar gráficos, histogramas, gráficos de barras, gráficos de dispersão, com linhas de código.

O módulo Joblib foi utilizado para converter um objeto de um fluxo de dados para que possa ser armazenado em um arquivo, eficaz em linguagem Python, para pipelines treinados. O módulo JSON serviu para exportação das métricas de avaliação em estrutura padronizada e legível, o que tornou fácil a documentação e a reprodutibilidade do estudo.

A biblioteca OpenPyXL fornece o mecanismo de leitura e escrita para arquivos Excel em formato .xlsx. Por sua vez, o Pathlib disponibiliza a interface orientada a objetos para que ocorra a manipulação de caminhos de sistema de arquivos, permitindo a migração entre plataformas.

Além destes, utilizou-se a biblioteca Typing para fornecer anotações para documentação de funções, especificando os tipos de retorno (List, Tuple, Optional, Dict) que melhoram a legibilidade e assim permitem a verificação estática de tipos. O módulo Argparse é implementado para a interface de linha de comando (CLI) com parsing automático de argumentos, validação de tipos e geração de mensagens de ajuda, permitindo parametrização flexível dos scripts de treinamento. Algumas ferramentas são mostradas na Figura 8.

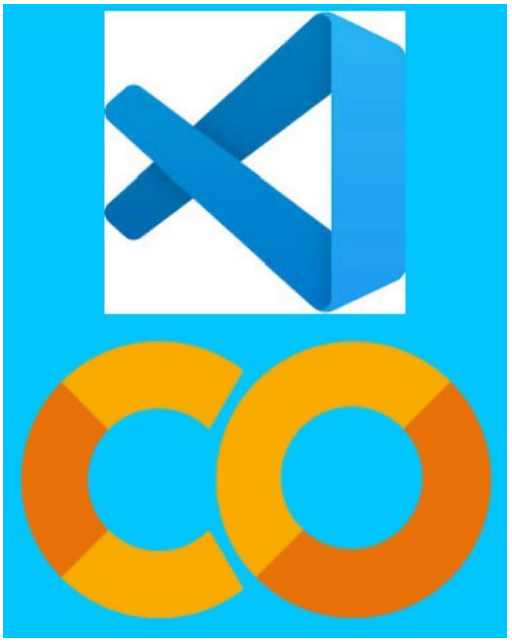
Figura 8 – Ferramentas utilizadas para a pesquisa

Fonte: Autoria Própria (2025)

O desenvolvimento e execução dos algoritmos foi no ambiente Google Colaboratory (Google Colab), uma plataforma gratuita em nuvem disponibilizada pelo Google para execução de desenvolvimento de códigos. Com mais detalhes, esta proporciona um ambiente de notebook dinâmico e compartilhado permitindo a criação e execução de código direto no navegador, sem a necessidade de configuração ou instalação qualquer outro tipo de *software* no computador, fornecendo acesso gratuito a GPUs e TPUs para aceleração de treinamento de deep learning.

O outro ambiente utilizado foi Visual Studio Code, abreviado como VS Code, um editor de código-fonte, muito utilizado para a criação, edição e depuração de código em inúmeras linguagens de programação, conforme mostrado na Figura 9. Possui uma interface amigável, capacidade de expansão e suporte a inúmeras ferramentas e *frameworks*, o que torna o ambiente de grande preferência para o desenvolvimento de algoritmos de aprendizado.

Figura 9 – Ambientes usados na pesquisa



Fonte: Autoria Própria (2025)

4.3. Métricas de Avaliação

4.3.1. Coeficiente de Determinação (R^2)

A métrica Coeficiente de Determinação (R^2) mede a proporção da variância da variável dependente que é explicada (ou previsível) pelas variáveis independentes (Chicco, 2021). O R^2 vai de 0 a 1, onde valores próximos a 1 indicam que o modelo deixa claro a

variação dos dados.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

O y_i representa os valores reais, o $\bar{y}_i$ os valores previsto, o $\bar{y}$ a média desses valores e *n* o número total de observações.

4.3.2. Raiz do Erro Quadrático Médio (RMSE)

A Raiz do Erro Quadrático Médio (RMSE) representa o desvio padrão dos resíduos e proporciona uma medida do erro médio em unidades da variável resposta (Reais). No código, o RMSE foi calculado através da raiz quadrada do erro quadrático médio:

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

O $\hat{y}_i$ é o valor previsto pelo modelo para observação *i*, e o y_i é o valor real da observação *i*, a diferença dos dois é o erro de previsão, *n* o valor das observações. A fórmula indica, em média, quanto os valores previstos se afastam dos valores reais, na mesma unidade da variável original.

4.3.3. Erro Absoluto Médio (MAE)

O Erro Absoluto Médio (MAE) mede o erro médio em valor absoluto entre observações y_i e previsões $\hat{y}_i$, sendo dependente de escala e menos sensível a outliers do que métricas quadráticas (Hyndman, 2006).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2$$

4.4. Fonte de Dados

A coleta de dados ocorreu em três etapas, feita de forma cuidadosa, estruturada e sequencial, para o levantamento sistemático e que abrangesse as informações de grande

relevância sobre medicamentos que são utilizados no tratamento de doenças reumáticas.

Sendo assim, cada etapa da busca foi elaborada com o objetivo de assegurar a confiança e a abrangência dos dados reunidos, isso acabou permitindo uma análise crítica e embasada nos aspectos relacionados à disponibilidade ao seu acesso, utilização e impacto financeiro desses medicamentos no contexto do sistema de saúde.

De início ocorreu a consulta de dados epidemiológicos no Departamento de Informática do Sistema Único de Saúde (DATASUS) que é um sistema desenvolvido pelo Ministério da Saúde do Brasil e disponibiliza informações referentes a mortalidade, internações hospitalares, morbidades e assistência à saúde de toda a população brasileira.

Desde sua criação em 1991, o DATASUS implementou sistemas de informação e suporte de informática necessários para o planejamento, operação e controle aos órgãos ligados aos SUS. Além disso, o DATASUS disponibiliza informações que podem auxiliar na análise da situação sanitária, na tomada de decisões baseadas em evidências e na elaboração de programas de saúde (Koike, 2025).

O sistema integra variadas bases de dados coletados em diferentes sistemas de informação em saúde, para a permissão ao acesso a registros, sejam na parte administrativa e epidemiológicos em escala nacional. A plataforma também disponibiliza dados bem explicados possibilitando análises epidemiológicas bem consistentes, junto oferece estudos de tendências cronológicas, estimativas das desigualdades das regiões do país e monitoramento de indicadores de saúde.

Para a extração e processamento dos dados do DATASUS, foi usada a biblioteca PySUS, desenvolvida em linguagem Python principalmente para facilitar esse acesso aos dados, *download* e manipulação de dados do Sistema Único de Saúde brasileiro.

O PySUS foi eficiente para que ocorresse a superação das limitações acerca das interfaces de consulta do DATASUS, permitindo baixar e fazer a leitura de forma automatizada de inúmeros sistemas de informação em saúde, contendo Sistema de Informações Hospitalares (SIH), Sistema de Informações sobre Mortalidade (SIM), Sistema de Informações sobre Nascidos Vivos (SINASC), dentre vários outros.

Os dados obtidos consideraram o período de janeiro do ano de 2020 a dezembro de 2024, em que abrangeu todos os casos de doenças reumáticas que foram registrados no Sistema de Informações Hospitalares (SIH) do DATASUS durante esse período de tempo, isso totalizou 5 anos de dados epidemiológicos e financeiros coletados.

O motivo dos dados a partir de 2020 se explica pela disponibilidade de registros mais consistentes no DATASUS e no Banco de Preços em Saúde (BPS), a partir da pandemia da

COVID-19, os sistemas acabaram passando por aprimoramentos em sua qualidade e assim ocorreu a padronização dos dados nesse período, vale destacar a demora do processamento para obtenção dos dados referente a saúde.

A população da pesquisa incluiu-se todos os pacientes que teve diagnósticos classificados nos códigos CID-10 correspondentes às doenças reumáticas crônicas, em todas as idade, sexo ou localização geográfica, nas 27 unidades federativas do Brasil, e foram registrados no sistema.

O fim da primeira etapa se deu com os dados extraídos sendo: UF, Faixas de CID-10 para reumáticas crônicas, ANO, MÊS e CASOS. Esta abrangência de variáveis foi para ser feito uma análise precisa e multidimensional no perfil epidemiológico das doenças reumáticas no território nacional, para uma compreensão mais aprofundada do impacto e predominância dessas doenças no sistema de saúde brasileiro.

O início da segunda etapa ocorreu com a realização de uma pesquisa em sites médicos na *Web* para que pudesse ocorrer identificação dos nomes dos medicamentos indicados para o tratamento dessas condições. Esta permitiu estabelecer quais fármacos iriam ser usados na pesquisa.

O início da terceira etapa se deu com o levantamento dos medicamentos, para isso ocorreu a busca na *Web* e depois uma pesquisa nos preços através do Banco de Preços em Saúde (BPS) que é um sistema de informação *online*, de acesso público, que registra, armazena e disponibiliza os preços de medicamentos e produtos de saúde gerenciado pelo Ministério da Saúde, responsável pelo registro e liberação das informações de compras públicas e privadas. A plataforma possui de forma obrigatória desde dezembro de 2017 informações de todos os medicamentos comprados para o abastecimento do SUS.

Este sistema tem o objetivo de mostrar para a população a transparência envolvendo a utilização dos recursos públicos e os preços em todo o território brasileiro de forma principal nas aquisições farmacêuticas, isso acaba funcionando como acompanhamento para minimizar a desigualdade de informações entre fornecedores e gestores.

Os dados extraídos incluem informações bem precisas sobre nome de medicamento, código de compra, fabricante, fornecedor, instituição compradora, valor unitário, quantidade adquirida e valor total das transações.

Na Tabela 1 são mostrados os resultados simplificados dessa busca, informando o nome do medicamento, para qual a sua indicação, o local desses preços, onde este possui menor e maior valor. Sendo feita a coleta dos 27 estados do país.

Tabela 1 – Medicamentos para doenças reumáticas e suas indicações terapêuticas

MEDICAMENTO	INDICADO	MENOR PREÇO	MAIOR PREÇO
ABATACEPTE	Artrite Reumatóide e Artrite Idiopática Juvenil.	FRASCO-AMPOLA - MARABÁ (2021) - 0,18	FRASCO-AMPOLA - SÃO PAULO (2021) - 1.296,71
ADALIMUMABE	Artrite Reumatóide, Artrite Idiopática Juvenil; na segunda linha de tratamento da Artrite Psoríaca.	SERINGA - SÃO PAULO (2020) - 2,75	SERINGA: 0,4ML - SÃO PAULO (2024) - 3.533,00
BARICITINIBE	Artrite reumatoide ativa moderada a grave	COMPRIMIDO - CANÁPOLIS (2021) - 5,68	COMPRIMIDO - CURITIBA (2021) - 157,72
ETANERCEPTE	Artrite Reumatóide, Artrite Psoriásica, Espondilite Anquilosante e Psoríase	FRASCO-AMPOLA - SALTO (2022) - 1,35	SERINGA - VITÓRIA (2020) - 6.010,91
GOLIMUMABE	Tratar Doenças Autoimunes, como Artrite e a Colite Ulcerativa	SERINGA - 2.590,72 (2021) - SÃO PAULO	SERINGA - 3.408,88 (2024) – CURITIBA
INFLIXIMABE	Artrite Reumatoide, Espondilite Anquilosante, Artrite Psoriásica e Psoríase em placa	FRASCO-AMPOLA - RIO DE JANEIRO (2023) - 710,00	FRASCO-AMPOLA - ITAÍ (2022) - 3.416,00
IXEQUIZUMAB	Psoríase em Placas, Artrite Psoriásica e Espondilite Anquilosante	SERINGA - SÃO PAULO (2021) / CURITIBA (2022) - 4.560,3	SERINGA: 1 ML - SÃO PAULO (2024) - 5.580,4
RISANQUIZUMABE	Artrite Psoriásica	SERINGA - CURITIBA (2022) - 7.277,01	SERINGA - TUBARÃO (2020) - 13.686,23
RITUXIMABE	Artrite Reumatoide e Linfomas Não-Hodgkin-m, Linfoma de Grandes Células B e Linfoma Folicular	FRASCO: 50ML - JAGUARIÚNA (2021) - 1,546	FRASCO: 50ML - UBERABA (2024) - 8.400,00
SECUQUINUMABE	Artrite Psoríaca	SERINGA - SÃO	SERINGA -

	Ativa	PAULO (2020) - 8,85	MONTE ALTO (2020) - 5.438,62
TOCILIZUMABE	Artrite Reumatoide Ativa, Moderada a Grave	FRASCO: 10ML - SÃO PAULO (2022) - 0,0001	FRASCO: 4ML - JOÃO PESSOA (2024) - 1.770,85
TOFACITINIBE	Artrite Psoriásica Ativa	COMPRIMIDO - SÃO PAULO (2022) - 46,61	COMPRIMIDO - CURITIBA (2022) - 78,55
UPADACITINIBE	Artrite reumatoide, Artrite psoriásica, Espondiloartrite axial Retocolite ulcerativa, Doença de Crohn	COMPRIMIDO - RIBEIRÃO PRETO (2023) - 142,03	COMPRIMIDO - JOÃO PESSOA (2024) - 6.165,71
USTEQUINUMABE	Psoríase em Placa, Moderada a Grave	FRASCO: 26ML - SÃO PAULO (2022) - 17,848	FRASCO: 26ML - CAMPO GRANDE (2021) - 93.183,93

Fonte: Autoria Própria (2025)

Todas as informações coletadas foram organizadas em planilhas do Microsoft Excel, para que fosse feita a análise estruturada dos dados. Para fazer a predição os modelos treinam de forma distinta por faixa de valor, usando todos os dados juntos. As entradas são ANO, MES, UF, Quantidade Item Compra, Valor Total Compra e os indicadores obtidos do site DATASUS sih_reuma_resumo_estado_ano_mensal e sih_reuma_resumo_estado_ano, com isso os algoritmos aprendem durante o treino e durante as variações verificando o crescimento ao longo do tempo e recebendo todos os dados e prevê o preço com base nos padrões.

O RF usa todos os parâmetros fixos, os dados de entrada são de arquivos Excel e enriquecidos com um CSV de SUS e com três hiperparâmetros, o número de arvores na floresta *(n_estimators)* estabelecida em 100, à semente aleatória *(random_state)* fixada em 42 e o parâmetro *n_jobs* configurado em -1, tudo para garantir equilíbrio e reprodutibilidade.

O MLP recebe como parâmetros, o diretório de entrada com os arquivos Excel, um CSV opcional de dados do SUS, o nome da coluna alvo a ser prevista, a arquitetura da rede MLP, a taxa de aprendizado, o número de épocas de treinamento, a fração de dados reservada para teste e para validação, o diretório de saída, e, a clusterização. O modelo MLP é treinado com todos os dados juntos e o número de modelagens varia de 1 até o número de clusters escolhido.

O KNN, os parâmetros são o diretório de entrada dos Excel, um caminho do CSV do SUS e um diretório de saída para salvar resultados. O alvo é a coluna “Valor Item Compra”. O modelo define quantos vizinhos mais próximos serão considerados para fazer a predição. Este valor não é fixo, o sistema testa automaticamente vários valores diferentes, depois seleciona qual valor apresentou melhor desempenho. O modelo calcula a distância para no fim dizer quais são os vizinhos mais próximos e quais terão maior influência na predição.

Na Tabela 2 são apresentados somente os dados do estado do Ceará, como exemplo da distribuição regional dos medicamentos, não se mostra os demais estados em razão ao enorme volume de informações, o que iria comprometer a leitura do texto. A escolha do Ceará justifica-se pelo local onde ocorreu o estudo, desenvolvido na (UNILAB), campus de Redenção-CE. Porém ressalta-se que todos os dados completos de todas as unidades federativas foram utilizados para análise dos algoritmos.

Tabela 2 – Medicamentos para doenças reumáticas no estado do Ceará

DESCRIÇÃO	ANO	MÊS	LOCAL	QUANTIDADE ITEM COMPRA	VALOR TOTAL COMPRA
ETANERCEPTE, FRASCO-AMPOLA	2022	02	FORTALEZA	679,21	88.800
INFLIXIMABE, FRASCO-AMPOLA	2021	06	JUAZEIRO DO NORTE	3.065,62	14
INFLIXIMABE, FRASCO-AMPOLA	2022	03	JUAZEIRO DO NORTE	3.242,5	24
INFLIXIMABE, FRASCO-AMPOLA	2023	02	JUAZEIRO DO NORTE	2.100	24
RITUXIMABE, FRASCO: 10ML	2022	03	JUAZEIRO DO NORTE	2.110,4	8
RITUXIMABE, FRASCO: 10ML	2023	02	JUAZEIRO DO NORTE	1.982,1	8
RITUXIMABE, FRASCO: 50ML	2022	03	JUAZEIRO DO NORTE	5.388,3	15
RITUXIMABE, FRASCO: 50ML	2023	02	JUAZEIRO DO NORTE	4.851	15
USTEQUINUMABE, FRASCO: 0,5 ML	2022	03	JUAZEIRO DO NORTE	29.479,8	20

USTEQUINUMABE, SERINGA	2023	02	JUAZEIRO DO NORTE	13.289,26	20

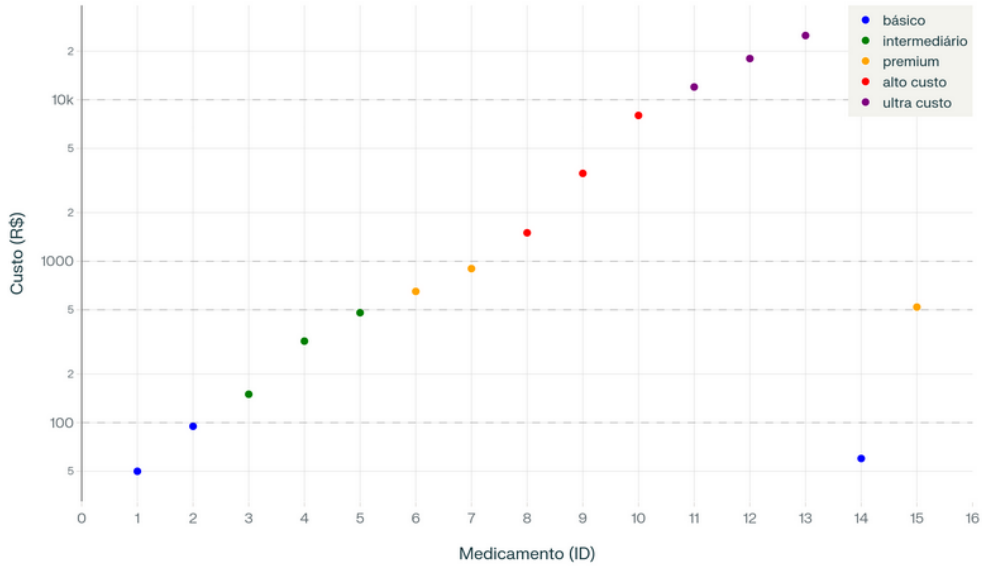
Fonte: Autoria própria (2025)

Os registros para o estado do Ceará abrangem os medicamentos Rituximabe, Infliximabe, Etanercepte e Ustequinumabe, sendo realizados de forma principal pelo Fundo Municipal de Saúde de Juazeiro do Norte e pela Secretaria da Saúde do Estado do Ceará. As dos medicamentos possuem uma variação entre os anos de 2021 e 2023.

4.5. Clusterização dos Dados por Medicamentos

Na etapa final foi feita a clusterização, no SUS não existe um "valor mínimo em reais" que define se um medicamento é de baixo, médio ou alto custo. Então foi feito análise fazendo a divisão de acordo de valores escolhidos, em cinco faixas de preço. A categoria básico com medicamentos com valores de 0 a 100,00, intermediário valores entre 101,00 e 500,00, premium inclui fármacos com valores de 501,00 a 1.000,00, alto custo valores de 1.001,00 a 10.000,00 e a por último ultra custo que engloba medicamentos na faixa de valores superiores a 10.001,00, a figura 10 mostra como ficou a divisão. Essa divisão foi utilizada para análise e visualização dos resultados de predição do modelo. Para cada cluster foram considerados, a descrição do medicamento, o valor, o período do preço e a quantidade de itens, igual ocorreu ao treinamento dos algoritmos.

Figura 10 - Categoria de custo de medicamento usados na pesquisa



Fonte: Autoria Própria (2025)

5. RESULTADOS E DISCUSSÃO

5.1. Casos Detectados

A análise dos dados epidemiológicos provenientes do DATASUS utilizando a biblioteca Pysus revelou o número de casos de doenças reumáticas registrados no Sistema Único de Saúde durante janeiro de 2020 e dezembro de 2024. A distribuição temporal desses casos evidencia uma tendência crescente significativa ao longo dos anos analisados, conforme é apresentado na Tabela 3.

Tabela 3 – Distribuição temporal de casos de doentes reumáticos no SUS (2020-2024)

ANO	CASOS
2020	13.297
2021	13.267
2022	14.901
2023	18.102
2024	22.730
TOTAL	82.297

Fonte: Autoria Própria (2025)

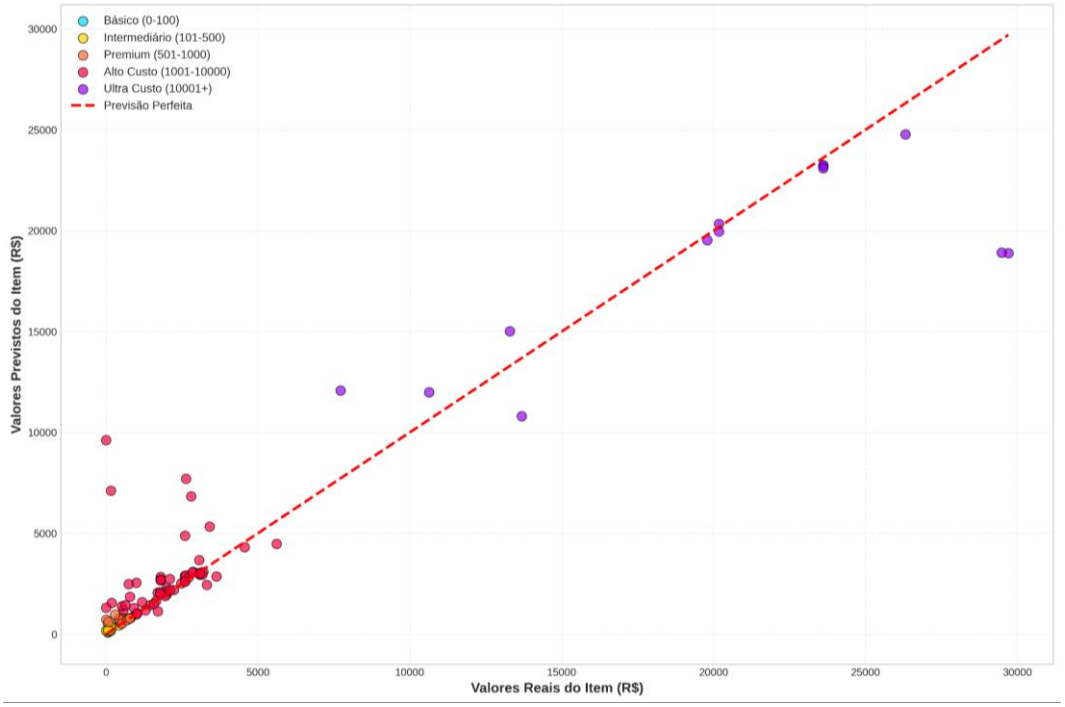
Na Tabela 3 é mostrado um total de 82.297 casos. No ano de 2020, foram registrados 13.297 casos no SUS, dessa forma foi possível perceber que o ano de 2021 se manteve estável referente ao ano anterior, com 13.267 casos. Em 2022 ocorreu um pequeno aumento, com 14.901 casos registrados.

Porém, esse crescimento foi mais significativo nos anos posteriores, com 18.102 casos em 2023 e 22.730 casos em 2024, o ano com maior registro.

5.2. Resultados - Random Forest

O RF apresentou poder explicativo em que a relação entre valores previstos e observados se mostrou numa crescente ao longo da faixa de preços, tendo uma tendência próxima à linha de identidade em diversos pontos, como é apresentado na Figura 11.

Figura 11 – Resultado do modelo Random Forest para medicamentos



Fonte: Autoria Própria (2025)

No gráfico o valor real (eixo x) e o valor previsto pelo modelo (eixo y) os pontos próximos à linha tracejada vermelha indicam boas previsões, enquanto desvios verticais de maior magnitude indicam erros mais altos em determinados pontos.

Na faixa de valores baixo e intermediário, existem pontos bem concentrados perto da origem, o modelo consegue capturar relativamente bem os itens mais baratos, com poucos desvios.

Os medicamentos Premium e Alto Custo ocupam a região entre 500,00 e 10.000,00 no eixo X, nessa faixa aparecem mais dispersão, para esses itens de preço médio e alto, o modelo ainda acerta a ordem de grandeza, mas com erros maiores em reais. Na faixa de ultra custo o modelo subestima.

Dessa forma, é possível perceber que os melhores resultados estão em até 5.000 mil reais. Além disso, é possível ver que a concentração de pontos ao redor da diagonal ocorreu uma dispersão a partir do que os valores aumentaram, esses algoritmos tendem a ter esse comportamento principalmente para previsões a regiões com menos exemplos ou maior variância.

As previsões do RF possui um padrão de variáveis numéricas e codificação de categorias, e portanto criou-se várias árvores para a amostras e também subconjuntos de atributos diferentes e agrega suas saídas para reduzir variância. Os erros são coerentes em

florestas aleatórias, principalmente quando a amostra é relativamente escassa.

Já referente às métricas de avaliação, o algoritmo no R^2 teve como resultado 0,89, isso significa que o modelo prediz de forma até excelente os preços, referente ao RMSE ficou R$ 2.200,07 e o MAE R$ 947,70. A tabela 4 apresenta a arquitetura proposta do modelo.

Tabela 4 – Métricas de desempenho do Random Forest

MÉTRICAS	RMSE	R^2	MAE
PERFEITO	0,0	1,0	0,0
RESULTADOS DO ALGORITMO	R$ 2.200,07	0,89	R$ 947,70

Fonte: Autoria Própria (2025)

Isso significa que o modelo explica 89.66% da variação dos preços em que quanto mais próximo de 1.0, melhor. Com uma boa base de dados, cada árvore da RF aprende caminhos mais estáveis e muitas árvores, o erro aleatório se cancela, isso resulta em pontos mais ao redor da diagonal e menor viés nos valores típicos do segmento.

Este explora as variáveis e quando a densidade de exemplos aumenta. Assim, as amostras geradas com reposição a partir do próprio conjunto de dados, fazem ser atributos mais confiáveis em múltiplas árvores. Esse ganho de cobertura do espaço de estados explica a boa calibração central e a redução de overfitting.

É possível perceber que ocorre uma dispersão de pontos muito acima da diagonal, isso deixa evidentes os erros e irregularidades em itens caros. Em particular, há um ponto de dados que fica distante das outras, isso faz mostrar uma distorção da escala e revela instabilidade com poucos exemplos.

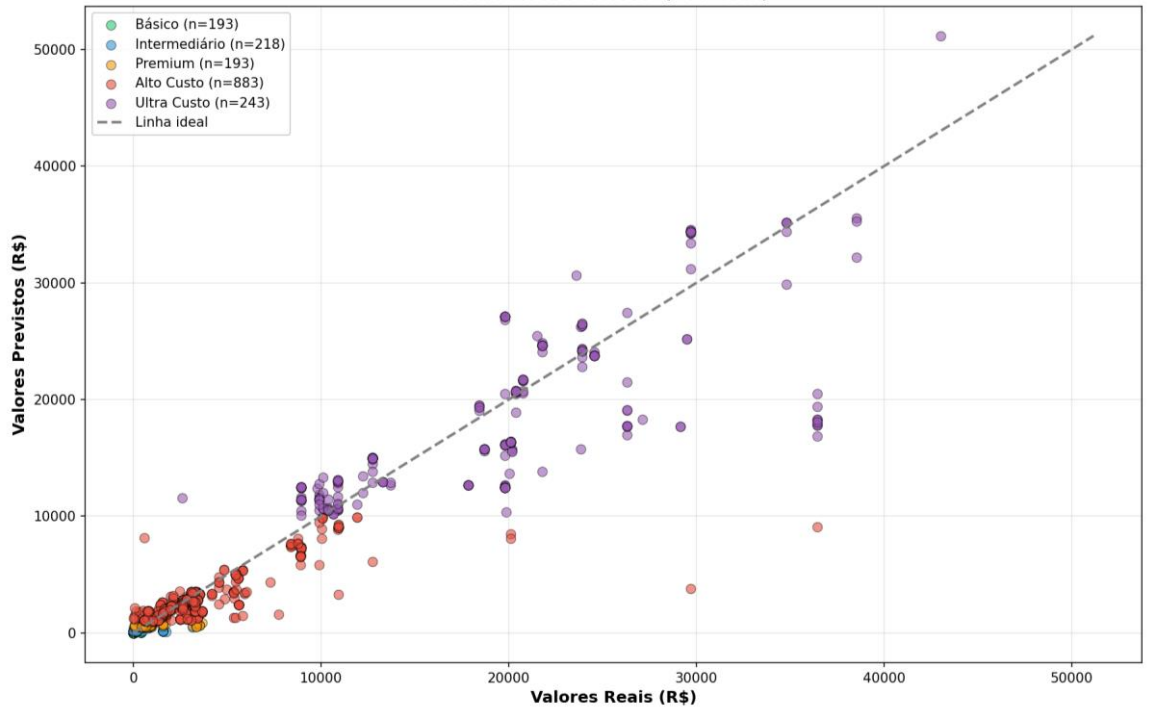
Devido a amostra pequena, isso faz que ocorra a redução da capacidade da RF de formar divisões de árvores, com menos observações por combinação de atributos, o agregador não consegue cancelar variâncias, isso aumenta o erro. Além disso, com os valores grandes podem ter especificações raras e o modelo tende a deixar mais suaves as pontas extremas e, portanto, quando acha dados padrões pouco frequentes, pode ocasionar o erro, como visto pelos pontos distantes da diagonal.

5.3. Resultados - Multi-Layer

Com relação ao MLP, adotou-se duas camadas ocultas, sendo a primeira é composta por 128 neurônios e a segunda por 64 neurônios. O treinamento tem um máximo de 200 épocas, e incorpora o Early Stopping com 20 épocas. Isso quer dizer que o processo acaba quando não ocorre melhora no desempenho do conjunto.

Os valores reais e os valores previstos pelo MLP são mostrados na Figura 12 abaixo, cuja a linha tracejada vermelha representa a predição ideal, onde os valores previstos correspondem exatamente aos valores reais.

Figura 12 – Resultado do modelo MLP para medicamentos



Fonte: Autoria Própria (2025)

O comportamento foi semelhante ao RF para valores até 5.000 mil reais. No entanto, para valores acima de 5.000 mil reais, os resultados simulados ficaram subestimados com relação aos reais.

Pela imagem, observa-se que, na faixa de valores básico, intermediário e premium, os pontos estão próximos da linha ideal. Isso indica que o modelo consegue prever razoavelmente bem os itens mais baratos. Há alguma dispersão, mas, no geral, a tendência é bem seguida.

A partir de valores acima de R$ 5.000,00, nas faixas alto custo e ultra custo, percebe-

se tanto subestimação quanto superestimação. Por fim, na faixa de ultra custo, a dispersão aumenta de forma mais acentuada, indicando maior variabilidade nos erros de previsão do modelo.

Já referente às métricas de avaliação o algoritmo no R^2 teve como resultado 0,6626, referente ao RMSE ficou R$ 4.083,44 e o MAE se obteve 1.660,86 fazendo o ser visto com desempenho mais elevado, como é mostrado na Tabela 5.

Tabela 5 – Métricas de desempenho do MLP

MÉTRICAS	RMSE	R^2	MAE
PERFEITO	0,0	1,0	0,0
RESULTADOS DO ALGORITMO	4.083,44	0,6626	1.660,86

Fonte: Autoria Própria (2025)

Isso mostra que a rede tem dificuldade em capturar extremos quando há menor densidade amostral e maior variabilidade interna.

Observa-se uma pouca concentração significativa de pontos nas regiões entre R$ 10.000,00 e R$ 30.000,00 indicando pouca capacidade de generalização do MLP nesta faixa.

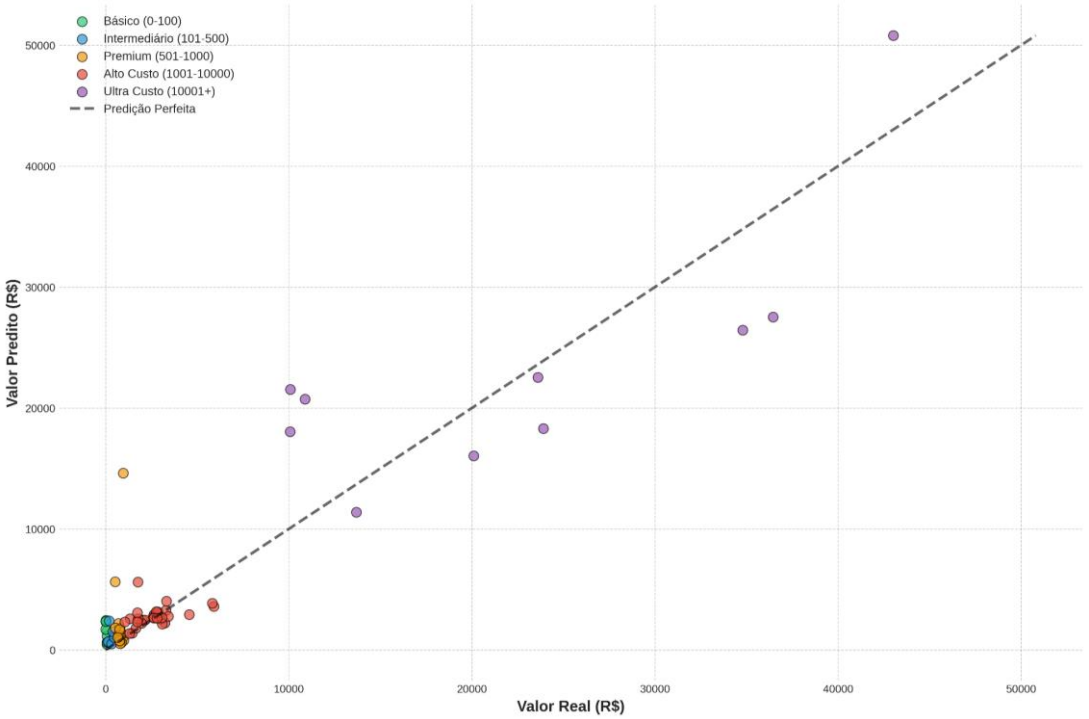
Percebe-se que à medida que os valores crescem, ocorre um maior espalhamento dos pontos em relação à linha inicial, isso evidenciou que o algoritmo apresenta enorme capacidade de submeter-se a variações nas previsões para itens de custo mais elevado. Isso pode ser atribuído a poucas amostras nas faixas de custo.

Os algoritmos acabam sendo insatisfatórios no próprio treinamento, pois não conseguem encontrar relações entre as variáveis, por insuficiência de exemplos na aprendizagem de padrões complexos em faixas de alto valor.

5.4. Resultados - KNN

Na Figura 13 são apresentados os valores reais e os valores previstos KNN para os custos dos medicamentos.

Figura 13 – Resultado da predição do modelo KNN para medicamentos



Fonte: Autoria Própria (2025)

Na imagem acima foi possível um mesmo comportamento que os algoritmos anteriores para valores até 5.000 mil reais e totalmente subestimados os valores previstos para acima desse valor, além disso nota-se que ocorre uma dispersão acentuada ao redor da linha, sem possuir um padrão.

Dá para ver que as categorias de menor valor básico e intermediário ficam próximas da linha, embora com alguns erros de sub e superestimação. À medida que os valores aumentam indo pra premium e alto custo, os pontos se espalham um pouco mais, mas ainda próximos da diagonal.

Já os itens de ultra custo, aparecem bem mais à direita do gráfico e mostram maior dispersão, alguns ficam abaixo da linha, indicando que o modelo subestima esses valores muito altos, e um ou outro fica acima, indicando superestimação. No geral, o KNN tende a se sair melhor nas faixas de custo mais baixas e médias, enquanto apresenta erros maiores e mais variáveis quando precisa prever itens muito caros.

Nota-se que os pontos concentram-se na região inferior do gráfico, mas os pontos também se localizam para baixo em relação à linha de referência. Pontos acima da linha indicavam previsão maior que o valor, enquanto pontos abaixo indicam previsão menor que o valor.

As métricas obtidas referentes a esse segmento são mostradas na Tabela 6,

comprovando o desempenho do KNN.

Tabela 6 – Métricas de desempenho da KNN

MÉTRICAS	RMSE	R^2	MAE
PERFEITO	0,0	1,0	0,0
RESULTADOS DO ALGORITMO	R$ 3.354,04	0,83	R$ 1.802,00

Fonte: Autoria Própria (2025)

O R^2 mostra que está obtendo uma boa previsão. Além disso, RMSE está muito alto para previsão de preços. O MAE também teve valor elevado, o que indica erros grandes nas previsões. Isso deixa claro que está de forma sistemática minimizando os custos reais, em que falha a captura de despesas com medicamentos. Um motivo provável deve ser o treinamento com poucos exemplos de valores tão altos.

5.5. Análise Comparativa

Com base no que foi apresentado, a análise comparativa dos três modelos revelou desempenhos bem diferentes, considerando as métricas utilizadas, é mostrado de forma visual os resultados obtidos nas Figuras 14, 15, 16.

Figura 14 – Gráfico comparativo com a métrica R^2

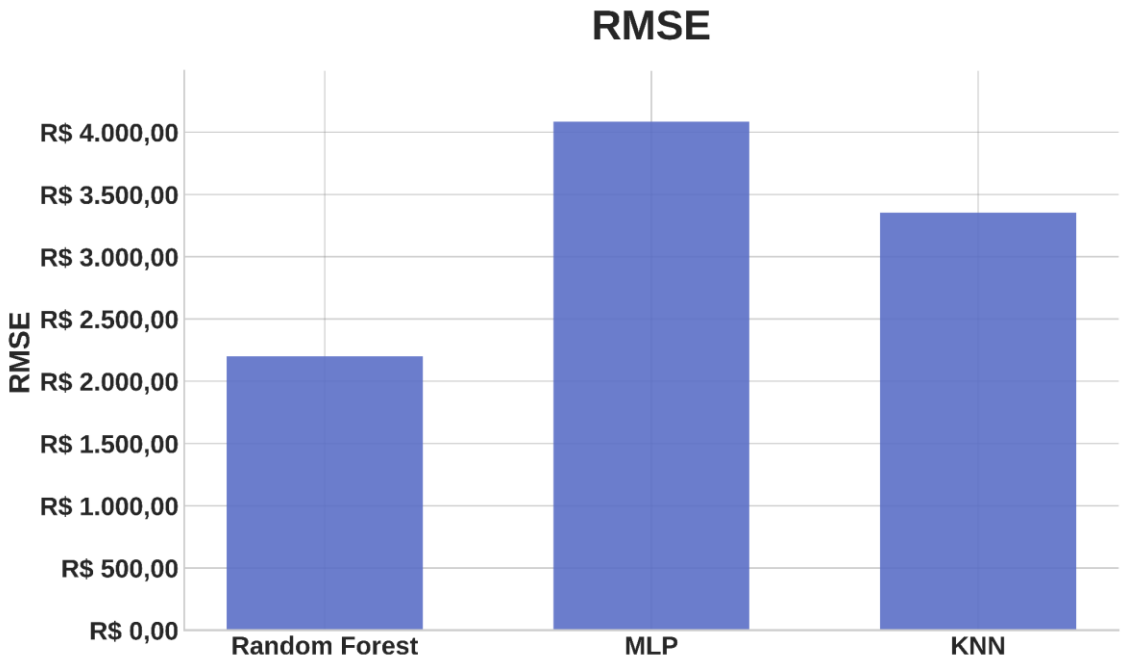
Fonte: Autoria Própria (2025)

O RF obteve o melhor desempenho, seguido pelo KNN, o comportamento observado pode ser justificado pelas características intrínsecas de cada algoritmo, ao agregar as previsões de muitas árvores, o RF reduz a variância do modelo final, tornando-o mais robusto e com melhor capacidade de generalização. Enquanto o KNN é muito flexível e pode modelar limites de decisão complexos, desde que os dados sejam bem agrupados no espaço de características.

Já o baixo resultado do MLP pode ser devido a sua arquitetura ou os dados que não eram ideais para um modelo de rede neural simples.

Por isso, o RF se mostrou o modelo mais adequado para o contexto específico de predição de custos farmacêuticos no SUS, enquanto o KNN apresentou desempenho intermediário e o MLP obteve o pior desempenho entre os três modelos.

Figura 15 – Gráfico comparativo com a métrica RMSE



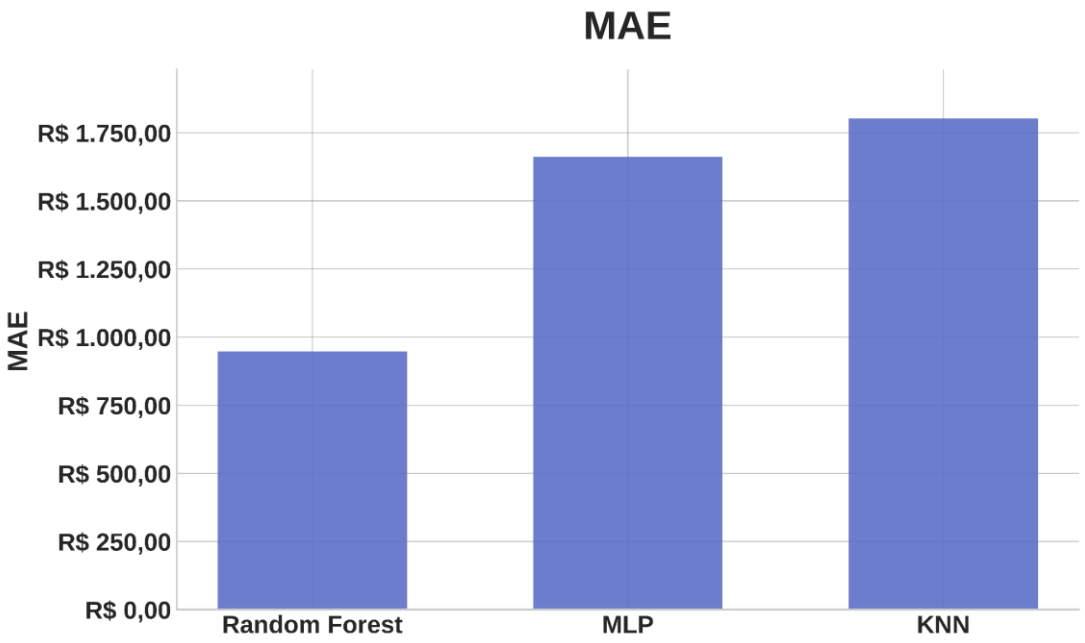
Fonte: Autoria Própria (2025)

O RMSE mostra erros moderados em relação a faixa de valores, dos três o RF se saiu com o melhor desempenho, enquanto o MLP teve o pior desempenho e já o KNN teve o desempenho intermediário. O desempenho como mencionado anteriormente é devido às características de cada algoritmo.

Um RMSE de R$ 2.200,07 para o Random Forest precisa ser avaliado considerando os custos reais de internações no SUS. Para um gestor do SUS, essa magnitude de erro pode ser aceitável para planejamento orçamentário agregado, mas problemática para decisões

individuais de alocação de recursos por paciente. O RMSE penaliza mais os erros grandes, indicando que o modelo ocasionalmente produz predições muito distantes do real - algo crítico quando os recursos são escassos e o subfinanciamento crônico do SUS exige precisão na gestão.

Figura 16 – Gráfico comparativo com a métrica MAE



Fonte: Autoria Própria (2025)

Em relação ao MAE, o KNN apresentou o pior desempenho. Ao analisar o MAE em relação ao RMSE do gráfico anterior, observamos dois aspectos importantes sobre o comportamento dos modelos, primeiro, confirma-se o bom desempenho e a robustez do RF. Em segundo, que o KNN foi o pior entre os três modelos, quando analisado nessa métrica.

Essa diferença está relacionada à forma como MAE e RMSE tratam os erros. O MAE atribui o mesmo peso a todos os erros, ao passo que o RMSE penaliza de forma mais intensa os erros grandes. Nesse contexto, destaca-se que o MLP comete erros elevados em algumas previsões, o que o torna menos confiável na prática. Já o KNN apresenta o maior MAE, indicando que suas previsões estão, de forma consistente, mais distante dos valores reais do que as dos outros modelos.

Pelo R^2, a ordem é RF, KNN, MLP, já pelo MAE, o RF continua melhor, mas o MLP fica levemente melhor que o KNN, por fim, pelo RMSE o RF também é superior, seguido do KNN e depois do MLP.

6. CONCLUSÃO

O estudo teve como principal objetivo fazer uma análise comparativa de algoritmos de aprendizagem de máquina para predição dos custos de medicamentos para o tratamento de doenças reumáticas no SUS. Cujo foi feito a busca, a coleta e o processamento dos dados de casos de doenças reumáticas, e logo após desenvolvido o estudo de três algoritmos diferentes, RF, MLP e KNN, e assim feito a avaliação do modelo apresentando qual obtém a maior capacidade de atingir o resultado esperado preditivo usando de métricas necessárias para comprovação e categorização nos valores dos medicamentos.

É necessário destacar que a pesquisa conseguiu alcançar um *pipeline* completo do DATASUS, que abrangeu casos que foram registrados entre os anos de 2020 e 2024. Dessa forma, os três algoritmos colocados em prática, treinados e avaliados utilizaram as métricas de desempenho R^2, RMSE e MAE, e isso permitiu fazer uma comparação sólida e valida.

O primeiro algoritmo analisado o RF foi o que mais demonstrou superioridade aos demais na predição de custos farmacêuticos, já que apresentou um R^2 de 0,89, um RMSE de R$ 2.200,07 e um MAE R$ 947,70 para os medicamentos. Enquanto o algoritmo MLP se mostrou perante ao anterior como desempenho intermediário, com um R^2 de 0,6626. O último algoritmo estudado foi a KNN, que em alguns aspectos apresentou comportamento semelhante ao RF e em outros ao MLP.

Nesse sentido, esse trabalho pode se mostrar como um grande contribuidor para a melhora na administração de recursos no SUS, a fim de fornecer incentivo técnico e científico na criação de ferramentas preditivas no planejamento orçamentário e na negociação de preços de medicamentos. Por isso, a metodologia se demonstrou fácil para a aplicação em situações reais de administração farmacêutica pública, uma vez que as políticas públicas são a que possui mais ações, programas e leis do governo na resolução de problemas e melhoria na qualidade de vida da população.

Por último, a principal dificuldade referente ao estudo é devido a pouca quantidade de dados disponíveis da área de estudo para ser feito o estudo com bastante precisão, onde alguns os modelos apresentaram-se com dificuldades bastante expressivas. A baixa quantidade de amostras de valores resultou com a grande maioria de métricas de desempenho, às vezes insatisfatórias.

Outra dificuldade encontrada se refere à qualidade dos registros no DATASUS, isso se mostra como um ponto que também deve ser tratado é o aperfeiçoamento nos sistemas de informação, a fim de evitar precariedade tecnológica nas unidades básicas de saúde e para isso

se torna necessário capacitação dos profissionais para utilização de ferramentas digitais.

E encerra-se como previsão futura a expansão do conjunto de dados para incorporar mais períodos tempos, essa aplicação em outras classes para prever custos de outros medicamentos, a exploração de outros algoritmos para investigação de outros modelos, também um desenvolvimento de interface para gestores, o que vai facilitar a tomada de decisão.

Em síntese, os resultados conquistados contribuem para a sustentabilidade financeira, economia, e vão garantir o acesso equitativo para os tratamentos para pacientes com doenças reumáticas.

7. REFERÊNCIAS

AL MAINI, M.; AL WESHAHI, Y.; FOSTER, H. E.; CHEHADE, M. J.; GABRIEL, S. E.; AL SALEH, J.; et al. **A global perspective on the challenges and opportunities in learning about rheumatic and musculoskeletal diseases in undergraduate medical education**. Clinical Rheumatology, v. 39, p. 627–642, 2020. DOI: https://10.1007/s10067-019-04544-y.

AL MAINI, M.; ADELOWO, F.; AL SALEH, J.; et al. **The global challenges and opportunities in the practice of rheumatology: White paper by the World Forum on Rheumatic and Musculoskeletal Diseases. Clinical Rheumatology**, 34:819–829, 2015. DOI: https://10.1007/s10067-014-2841-6.

QIFANG Bi, Katherine E. GOODMAN, Joshua KAMINSKY, Justin LESSLER, **What is Machine Learning? A Primer for the Epidemiologist**, *American Journal of Epidemiology*, Volume 188, Issue 12, December 2019, Pages 2222–2239, https://doi.org/10.1093/aje/kwz189

BOING, A. C.; BERTOLDI, A. D.; BOING, A. F.; BASTOS, J. L.; PERES, K. G. **Acesso a medicamentos no setor público: análise de usuários do Sistema Único de Saúde no Brasil**. Cadernos de Saúde Pública, Rio de Janeiro, v. 29, n. 4, p. 691-701, 2013.

BRASIL. Ministério da Saúde. Secretaria de Atenção à Saúde. **Doenças reumáticas [recurso eletrônico]**. Brasília, DF: Ministério da Saúde, 2013. Disponível em: https://bvsms.saude.gov.br/bvs/folder/doencas_reumaticas.pdf

Brasileiros gastaram mais de R$ 215 bilhões com medicamentos em 2024 - Medicina S/A. https://medicinasa.com.br/reajuste-medicamentos/ Acesso: 2025-11-08

CAMPOS, Gastão Wagner de Sousa. ***SUS: o que e como fazer?*** **Ciência & Saúde Coletiva**, v. 23, n. 6, p. 1707-1714, 2018. DOI: 10.1590/1413-81232018236.05582018.

CERRI, Ricardo; CARVALHO, André Carlos Ponce de Leon Ferreira de. *Aprendizado de máquina: breve introdução e aplicações*. **Cadernos de Ciência & Tecnologia**, Brasília, v. 34, n. 3, p. 297–313, set./dez. 2017.

CHAHDI, Ismail; ELMIAD, Aissa Kerkour; BADAOUI, Mohammed. **Data Preprocessing For Machine Learning Applications in Healthcare: A Review**. In: 14th International Conference on Intelligent Systems: Theories and Applications (SITA), 2023. IEEE. DOI: 10.1109/SITA60746.2023.10373591.

CHICCO, D., WARRENS, M. J., & Jurman, G. (2021). **The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation**. PeerJ Computer Science, 7, e623. https://doi.org/10.7717/peerj-cs.623

Doenças Reumáticas acometem mais de 15 milhões de brasileiros, de qualquer idade, causam limitações, aposentadoria precoce e sérios impactos no sistema de saúde no país - Sociedade Brasileira de Reumatologia. https://www.reumatologia.org.br/press-releases/doencas-reumaticas-acometem-mais-de-15-milhoes-de-brasileiros-de-qualquer-idade-causam-limitacoes-aposentadoria-precoce-e-serios-impactos-no-sistema-de-saude-no-

pais/ Acesso: 2025-11-11

FATEL, K. O.; ROVER, M. R. M.; MENDES, S. J.; LEITE, S. N.; STORPIRTIS, S. **Desafios na gestão de medicamentos de alto preço no SUS: avaliação da Assistência Farmacêutica em São Paulo, Brasil**. *Ciência & Saúde Coletiva*, **26**(11), 5481–5498, 2021. https://doi.org/10.1590/1413-812320212611.00842021.

FAZEKAS, M., VELJANOV, Z. & de Oliveira, A.B. **Predicting pharmaceutical prices. Advances based on purchase-level data and machine learning**. BMC Public Health 24, 1888 (2024). https://doi.org/10.1186/s12889-024-19171-9 .

FONTANA, É. **Introdução aos Algoritmos de Aprendizagem Supervisionada**. Departamento de Engenharia Química, Universidade Federal do Paraná (UFPR), 2020. Apostila.

GARDNER, M. W.; DORLING, S. R. **Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment**, 32(14–15), 2627–2636, 1998.

HECHT-NIELSEN, R. **Theory of the Backpropagation Neural Network**. In: International Joint Conference on Neural Networks (IJCNN). IEEE, 1989.

HE, X.; CHUA, T.-S. **Neural Factorization Machines for Sparse Predictive Analytics**. In: SIGIR 2017 — 40th International ACM Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, 7–11 Aug. 2017. New York: ACM, 2017. p. 355–364. DOI: https://10.1145/3077136.3080777.

HYNDMAN, R. J., & KOEHLER, A. B. (2006). **Another look at measures of forecast accuracy**. International Journal of Forecasting, 22(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

JORDAN, M. I.; MITCHELL, T. M. **Machine learning: Trends, perspectives, and prospects**. Science, v. 349, n. 6245, p. 255–260, 2015. DOI: https://10.1126/science.aaa8415.

KALE, Arati K.; PANDEY, Dev Ras. **Data Pre-Processing Technique for Enhancing Healthcare Data Quality Using Artificial Intelligence**. International Journal of Scientific Research in Science and Technology, v. 11, n. 1, p. 299–309, jan./fev. 2024. DOI: https://10.32628/IJSRST52411130.

KOIKE, M. (2025). **DataSUS: Uma Ferramenta Essencial para a Saúde Pública no Brasil**. Arquivos Brasileiros de Cardiologia, 122(2), e20250123. https://doi.org/10.36660/abc.20250123

KOHLER JC, MITSAKAKIS N, SAADAT F, BYNG D, MARTINEZ MG. **Does Pharmaceutical Pricing Transparency Matter? Examining Brazil's Public Procurement System. Global Health**. 2015 Aug 4;11:34. doi: 10.1186/s12992-015-0118-8. PMID: 26238110; PMCID: PMC4523918.

KOO, M.; LU, M.-C. **Rheumatic Diseases: New Progress in Clinical Research and**

Pathogenesis. Medicina, v. 59, p. 1581, 2023. DOI: https://10.3390/medicina59091581.

KOTSIANTIS, S. B. **Supervised machine learning: A review of classification techniques**. *Informatica*, **31**, 249–268, 2007.

LIMA, S. G. G.; BRITO, C.; ANDRADE, C. J. C. **O processo de incorporação de tecnologias em saúde no Brasil em uma perspectiva internacional**. *Ciência & Saúde Coletiva*, **24**(5), 1709–1722, 2019. https://doi.org/10.1590/1413-81232018245.17582017.

LINNÉR, L.; ERIKSSON, I.; PERSSON, M.; WETTERMARK, B. **Forecasting drug utilization and expenditure: ten years of experience in Stockholm**. *BMC Health Services Research*, 20:410, 2020. https://doi.org/10.1186/s12913-020-05170-0.

LUDERMIR, Teresa Bernarda. **Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. Estudos Avançados**, v. 35, n. 101, p. 85–94, 2021. DOI: 10.1590/s0103-4014.2021.35101.007.

MONARD, M. C.; BARANAUSKAS, J. A. **Conceitos sobre Aprendizado de Máquina**. In: *Sistemas Inteligentes para Engenharia*. [S.l.]: [s.n.], [s.d.]. cap. 4.

MUJICA, E. M. M.; BASTOS, J. L.; BOING, A. C. **Acesso a medicamentos, o Sistema Único de Saúde e as injustiças interseccionais**. Revista de Saúde Pública, 58:34, 2024. DOI: https://10.11606/s1518-8787.2024058005986.

MURTAGH, F. **Multilayer perceptrons for classification and regression**. Neurocomputing, 2, 183–197, 1990/1991. Elsevier.

NELSON AE, ARBEEVA L. **Narrative Review of Machine Learning in Rheumatic and Musculoskeletal Diseases for Clinicians and Researchers: Biases, Goals, and Future Directions**. J Rheumatol. 2022 Nov;49(11):1191-1200. doi: 10.3899/jrheum.220326. Epub 2022 Jul 15. PMID: 35840150; PMCID: PMC9633365.

OLIVEIRA, P.; MONTEIRO, P.; COUTINHO, M.; SALVADOR, M. J.; COSTA, M. E.; MALCATA, A. **Qualidade de vida e vivência da dor crónica nas doenças reumáticas**. Acta Reumatológica Portuguesa, 34, 511–519, 2009.

PAIM, J. S. **Sistema Único de Saúde (SUS) aos 30 anos**. Ciência & Saúde Coletiva, 23(6), 1723–1728, 2018. https://doi.org/10.1590/1413-81232018236.09172018

PONTES, A. P. M.; OLIVEIRA, D. C. de; GOMES, A. M. T. **The principles of the Brazilian Unified Health System, studied based on similitude analysis**. Revista Latino-Americana de Enfermagem, v. 22, n. 1, p. 59–67, 2014. DOI: 10.1590/0104-1169.2925.2395

RASCHKA, Sebastian; PATTERSON, Joshua; NOLET, Corey. **Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence**. Information, v. 11, n. 4, art. 193, 2020. DOI: 10.3390/info11040193.

RODRIGUES FILHO, F. J.; PEREIRA, M. C. **Decline in Public Spending on Biopharmaceutics in Brazil**. *Brazilian Journal of Pharmaceutical Sciences*, **58**, e20872,

2022. https://doi.org/10.1590/s2175-97902022e20872.

ROVER, M. R. M.; FARACO, E. B.; VARGAS-PELÁEZ, C. M.; COLUSSI, C. F.; STORPIRTIS, S.; FARIAS, M. R.; LEITE, S. N. **Acesso a medicamentos de alto preço: desigualdades na organização e resultados entre estados brasileiros**. *Ciência & Saúde Coletiva*, **26**(11), 5499–5508, 2021. https://doi.org/10.1590/1413-812320212611.27402020

SARKER IH. **Machine Learning: Algorithms, Real-World Applications And Research Directions**. SN Comput Sci. 2021;2(3):160. doi: 10.1007/s42979-021-00592-x. Epub 2021 Mar 22. PMID: 33778771; PMCID: PMC7983091.

SCHONLAU, M.; ZOU, R. Y. **The random forest algorithm for statistical learning**. *The Stata Journal*, **20**(1), 3–29, 2020. https://doi.org/10.1177/1536867X20909688

SCHMIDHUBER, J. **Deep Learning in Neural Networks: An Overview**. Neural Networks, 61, 85–117, 2015. Elsevier.

SINGH, K. A.; GUPTA, M. **The Role of Machine Learning in Various Sectors: A Comprehensive Review**. International Journal for Research in Applied Science and Engineering Technology (IJRASET), v. 13, 2025. DOI: 10.22214/ijraset.2025.70952.

Sistema Único de Saúde comemora 34 anos de democracia e cidadania — Agência Gov https://agenciagov.ebc.com.br/noticias/202409/sistema-unico-de-saude-comemora-34-anos-de-democracia-e-cidadania. Acesso: 2025-11-11

PARRY, R., JONES, W., STOKES, T. et al. **k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction**. Pharmacogenomics J 10, 292–309 (2010). https://doi.org/10.1038/tpj.2010.56

SMOLEN, J. S.; ALETAHA, D.; BARTON, A.; BURMESTER, G. R.; EMERY, P.; FIRESTEIN, G. S.; KAVANAUGH, A.; McINNES, I. B.; SOLOMON, D. H.; STRAND, V.; YAMAMOTO, K. **Rheumatoid arthritis. Nature Reviews Disease Prim**ers, v. 4, art. 18001, 2018. DOI: 10.1038/nrdp.2018.1.

TAVARES, N. U. L.; LUIZA, V. L.; OLIVEIRA, M. A.; COSTA, K. S.; MENGUE, S. S.; ARRAIS, P. S. D.; **et al.** *Free access to medicines for the treatment of chronic diseases in Brazil*. **Revista de Saúde Pública**, 50(Supl. 2):7s, 2016. DOI: 10.1590/S1518-8787.2016050006118.

UDDIN S, Haque I, LU H, Moni MA, Gide E. **Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction**. Sci Rep. 2022 Apr 15;12(1):6256. doi: 10.1038/s41598-022-10358-x. PMID: 35428863; PMCID: PMC9012855.

ZHANG Z. **Introduction to machine learning: k-nearest neighbors**. Ann Transl Med. 2016 Jun;4(11):218. doi: 10.21037/atm.2016.03.37. PMID: 27386492; PMCID: PMC4916348.