PGCOMP - Programa de Pós-Graduação em Ciência da Computação
Universidade Federal da Bahia (UFBA)
Av. Milton Santos, s/n - Ondina
Salvador, BA, Brasil, 40170-110

https://pgcomp.ufba.br
pgcomp@ufba.br

As técnicas de mineração de dados permitem extrair relações e características presentes nos dados que não são facilmente perceptíveis, especialmente em cenários envolvendo grandes volumes de dados. Estas técnicas permitem agrupar e classificar os dados, identificar padrões e exceções, bem como estabelecer relações de associação entre dados distintos. A análise preditiva é uma sub-área da mineração de dados que utiliza técnicas de aprendizado de máquina para a implementação de modelos analíticos de predição. Tais modelos baseiam-se no conhecimento prévio dos dados (análise descritiva) para projetar possíveis futuros acontecimentos acerca do contexto sendo analisado. Este trabalho é parte integrante de um projeto de pesquisa que objetiva integrar dados de notificação de malária presentes em diferentes fontes de informação e desenvolver modelos de análise preditiva sobre estes dados. Especificamente, este trabalho envolve o estudo de diferentes técnicas de predição e, a partir deste estudo, o desenvolvimento de modelos preditivos para o cenário epidemiológico da malária no Brasil. Quatro estudos de casos envolvendo dados epidemiológicos e climáticos são discutidos como prova de conceito dos modelos executados. Para a avaliação dos modelos foi utilizada a metodologia de validação evaluation on a rolling forecasting origin, a métrica utilizada foi o RMSE. Quatro algoritmos de aprendizado de máquina foram utilizados, suporte vector regression e random forest obtiveram os melhores resultados. Dentre os quatro estudos de caso, dois obtiveram maior precisão utilizando dados de clima em alguns cenários.

MSC | 084 | 2019

Desenvolvimento e Validação de Modelos Preditivos para Malária

Juracy Bertoldo Santos Junior

UFBA

Desenvolvimento e Validação de Modelos Preditivos para Malária

Juracy Bertoldo Santos Junior

Dissertação de Mestrado

Universidade Federal da Bahia

Programa de Pós-Graduação em
Ciência da Computação

Dezembro | 2019

Universidade Federal da Bahia
Instituto de Matemática e Estatística

Programa de Pós-Graduação em Ciência da Computação

DESENVOLVIMENTO E VALIDAÇÃO DE MODELOS PREDITIVOS PARA MALÁRIA

Juracy Bertoldo Santos Junior

DISSERTAÇÃO DE MESTRADO

Salvador
06 de dezembro de 2019

JURACY BERTOLDO SANTOS JUNIOR

DESENVOLVIMENTO E VALIDAÇÃO DE MODELOS PREDITIVOS PARA MALÁRIA

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: Prof. Marcos Ennes Barreto

Salvador
06 de dezembro de 2019

Sistema de Bibliotecas - UFBA

Santos Júnior, Juracy Bertoldo.
DESENVOLVIMENTO E VALIDAÇÃO DE MODELOS PREDITIVOS PARA MALÁRIA / Juracy Bertoldo Santos Junior – Salvador, 2019.
62p.: il.

Orientador: Prof. Dr. Prof. Marcos Ennes Barreto.
Dissertação (Mestrado) – Universidade Federal da Bahia, Instituto de Matemática e Estatística, 2019.

1. Computação. 2. Malária. 3. Controle Preditivo.. I. Barreto , Marcos Ennes. II. Universidade Federal da Bahia. Instituto de Matemática e Estatística. III Título.

CDD – S237
CDU – 004

"*DESENVOLVIMENTO E VALIDAÇÃO DE MODELOS PREDITIVOS PARA EPIDEMIA DE MALÁRIA*"

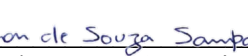
Juracy Bertoldo Santos Junior

Dissertação apresentada ao Colegiado do Programa de Pós-Graduação em Ciência da Computação na Universidade Federal da Bahia, como requisito parcial para obtenção do Título de Mestre em Ciência da Computação.

Banca Examinadora

Prof. Dr. Marcos Ennes Barreto (orientador / UFBA)

Profª. Drª. Tatiane Nogueira Rios (UFBA)



Prof. Dr. Vanderson de Souza Sampaio(UEA/FVS-AM)

RESUMO

As técnicas de mineração de dados permitem extrair relações e características presentes nos dados que não são facilmente perceptíveis, especialmente em cenários envolvendo grandes volumes de dados. Estas técnicas permitem agrupar e classificar os dados, identificar padrões e exceções, bem como estabelecer relações de associação entre dados distintos. A análise preditiva é uma sub-área da mineração de dados que utiliza técnicas de aprendizado de máquina para a implementação de modelos analíticos de predição. Tais modelos baseiam-se no conhecimento prévio dos dados (análise descritiva) para projetar possíveis futuros acontecimentos acerca do contexto sendo analisado. Este trabalho é parte integrante de um projeto de pesquisa que objetiva integrar dados de notificação de malária presentes em diferentes fontes de informação e desenvolver modelos de análise preditiva sobre estes dados. Especificamente, este trabalho envolve o estudo de diferentes técnicas de predição e, a partir deste estudo, o desenvolvimento de modelos preditivos para o cenário epidemiológico da malária no Brasil. Quatro estudos de casos envolvendo dados epidemiológicos e climáticos são discutidos como prova de conceito dos modelos executados. Para a avaliação dos modelos foi utilizada a metodologia de validação (*evaluation on a rolling forecasting origin*, a métrica utilizada foi o Raiz do Erro Médio Quadrático (RMSE). Quatro algoritmos de aprendizado de máquina foram utilizados, *suporte vector regression* e *random Forest* obtiveram os melhores resultados. Dentre os quatro estudos de caso, dois obtiveram maior precisão utilizando dados de clima em alguns cenários.

Palavras-chave: Mineração de dados. Análise preditiva. Predição baseada em dados. Malária.

ABSTRACT

Data mining techniques allow the extraction of relationships and characteristics from data which are not easily identifiable, specially in big data scenarios. These techniques provide routines to group and classify data, identify patterns and outliers, as well as derive associations among distinct data. Predictive analytics is a part of the data mining workflow that relies on machine learning to implement different prediction models. These models are based on previous knowledge about the data to forecast possible outcomes related to the domain being considered. This work belongs to a project focusing on the integration of different data sources from the Brazilian malaria ecosystem and on the design of different predictive analytics models to be used over these data sources. More specificallly to design a predictive analytics model applied to malaria surveillance in Brazil. Four case studies comprising epidemiologic and climate data are discussed as proof of concepts. For the evaluation of the models the methodology of cross-validation for time series evaluation on rolling forecasting origin was used, the metric used was ac RMSE. Four machine learning algorithms were used, support vector regression and random Forest obtained the best results. Within the four case studies, two achieved greater accuracy using weather data in some scenarios.

Keywords: Data mining. Predictive analytics. Malaria epidemics. Forecasting malária

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

Capítulo 1

INTRODUÇÃO

A Tecnologia da Informação (TI) é uma área do conhecimento a qual aplica a computação como forma de armazenar e transmitir dados, bem como, realizar o processamento destes para diversos tipos de aplicações. O uso de recursos de TI tornou-se essencial para a área de saúde, permitindo que dados coletados para fins administrativos passassem a ser usados para fins de prevenção, sistemas de apoio à decisão, sistemas de cuidado ao paciente e diversos outros propósitos (PINOCHET, 2011).

A Área da Saúde tem buscado, cada vez mais, integrar-se e utilizar-se de recursos computacionais a fim de i) automatizar seus processos, ii) promover a integração de diferentes tipos de dados (clínicos, laboratoriais, administrativos) para a melhoria e o embasamento de processos de tomada de decisões e iii) fomentar pesquisas para avanço tecnológico e metodológico.

Diversas organizações de saúde coletam, diariamente, uma grande quantidade de dados os quais, em conjunto com a TI, geram informação e colaboram para a melhoria da saúde populacional em nível de gestão. Esse aumento constante na quantidade de dados de saúde gerados diariamente, aliado às demandas que os diferentes tipos de usuários possuem, inviabilizam o manejo daqueles em tempo aceitável sem o uso de ferramentas computacionais.

A Mineração de Dados (MD) é uma ramo da Computação iniciada nos anos 80, quando os profissionais da área começaram a se preocupar com os grandes volumes de dados informáticos estocados e subutilizados. O uso das técnicas de MD permite extrair relações e características presentes nos dados e que não são diretamente visíveis. (AMO, 2004). A MD trata-se de uma área interdisciplinar que engloba técnicas de diversas matérias diferentes, essas matérias, também tiveram um aumento em seus estudos. O Aprendizado de Máquina (AM) foi uma dessas disciplinas.

O AM (*machine learning*) engloba um conjunto de técnicas usadas para apoiar aplicações de mineração de dados e tem, por objetivo, ensinar o computador a aprender determinadas tarefas a partir da experiência prévia ou do treinamento sobre determinados conjuntos de dados (FACELI et al., 2011a).

Uma sub-área associada à MD s e que utiliza técnicas deMD é a análise preditiva (FINLAY, 2014). Modelos analíticos para predição estão se tornando cada vez mais comuns devido aos avanços computacionais e à disponibilidade de grandes volumes de dados advindos de diferentes áreas do conhecimento.

1.1 MOTIVAÇÃO

A malária, também conhecida como paludismo, maleita ou sezão, ainda é um dos principais problemas de saúde pública do mundo. Os parasitas transmissores são do gênero *Plasmodium* e são conhecidas cerca de 150 espécies causadoras de malária em diferentes hospedeiros; destas, quatro espécies infectam o homem: *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malarie* e *Plasmodium ovale* (este último ocorre somente em regiões restritas do continente africano) (NEVES, 2005).

Aproximadamente 262 milhões e 214 milhões de casos de malária foram diagnosticados no mundo nos anos 2000 e 2015, respectivamente, caracterizando uma redução de 18%. Os óbitos nesses mesmos períodos foram de 839 mil e 438 mil, respectivamente, caracterizando uma redução de 48% (WHO, 2015).

No Brasil, houve uma significativa diminuição da morbidade e mortalidade desde o ano 2000, atingindo o objetivo 58.2 da World Health Organization (WHO) que é "reduzir a carga de malária em 75%". Entre 2000 e 2014, houve uma redução de 76,7% nos casos relatados, com uma média anual de 19% de redução neste período (PAHO, 2014). A partir de 2016, o número de casos reportados voltou a crescer, conforme ilustra a Figura 1.1.

Figura 1.1 Evolução dos casos de malária no Brasil.

A área da Floresta Amazônica é onde a malária é altamente prevalente, correspon-

dendo a cerca de 99,8% dos casos, sendo que destes, 15 municípios responsáveis por 57.3% em 2014. O *Plasmodium vivax* foi o parasita responsável por 82.9% dos casos em 2014, enquanto o *Plasmodium falciparum* e outros tipos causaram cerca de 16.3% dos casos. O principal vetor na região da Amazônia é o *Anopheles darlingi* (PAHO, 2014).

Em 2003, o governo brasileiro implantou o Programa Nacional de Prevenção e Controle de Malária (PNCM), conjunto de diretrizes para que os governos federais, estaduais e municipais possam prevenir e controlar efetivamente a doença. Um dos recursos estabelecidos pelo PNCM foi o SIVEP-malária, composto por módulos de notificação de casos, emissão de relatórios, cadastro de localidades, laboratórios, unidades e agentes notificantes. As fichas de notificação alimentam o sistema com uma série de informações, incluindo unidade notificante, agente notificante, local provável de ocorrência, resultado do exame laboratorial, esquema de tratamento, entre outros. No SIVEP-malária é possível verificar a distribuição dos exames de sangue realizados, se a notificação do caso foi registrada por busca ativa ou passiva, quais casos são importados e quais são autóctones, entre outras informações. Além disso, o sistema permite o cadastro com e sem internet, por conta da dificuldade de acesso das localidades que precisam utilizar o sistema.

Além do SIVEP, dados de outros sistemas de informação, tais como controle de vetores, Sistema de Informação sobre Mortalidade (SIM), Sistema de Informações Hospitalares (SIH) e Sistema de Informação de Agravo de Notificação (SINAN) podem ser integrados em atividades de monitoramento e pesquisa, uma vez que os recursos de tecnologia da informação permitem a integração destas diferentes bases de dados como uma abordagem viável, apesar de ainda não totalmente explorada, para o combate da malária no Brasil.

1.1.1 Projeto de pesquisa – edital GCE

Em 2016, a Fundação Bill & Melinda Gates lançou a chamada Grand Challenges Explorations (GCE) com o objetivo de acelerar o processo de erradicação da malária no mundo[1]. O desafio concentrava-se na busca de soluções inovadoras as quais possibilitassem um progresso referente à disponibilidade e ao uso dos dados para auxiliar na tomada de decisão no contexto do combate à malária no mundo.

O projeto *"Integrando dados socioeconômicos e de saúde para combate à malária"*[2] foi um dos 96 projetos selecionados dentre as 1.400 propostas recebidas e teve, como objetivo principal, a criação de uma plataforma de software para integração de dados de bases governamentais ligados ao sistemas brasileiro de monitoramento de malária e, também, propor métodos de análise preditiva e mineração visual de dados. Este trabalho é parte integrante deste projeto, especificamente no tema referente ao desenvolvimento de modelos de análise preditiva.

O projeto esteve vigente entre outubro de 2016 e abril de 2019, e propôs uma série de objetivos específicos:

1. Captação de dados dos sistemas de informações de malária (transmissão, monito-

[1] <https://gcgh.grandchallenges.org/>

[2] <https://gcgh.grandchallenges.org/grant/integrating-socioeconomic-and-health-data-combat-malaria>

ramento e tratamento);

2. Análise da qualidade e cobertura dos dados;

3. Emprego de rotinas de integração de dados para a construção de uma base única de notificações e demais informações relevantes para a análise, considerando também dados climáticos.

4. Implementação de estruturas de indexação e anotação dos dados.

5. Desenvolvimento e validação de métodos de análise preditiva.

6. Desenvolvimento e validação de metáforas para mineração visual de dados.

7. Geração de portal de acesso, relatórios técnicos e artigos para divulgação dos métodos e resultados produzidos.

Para alcançar os objetivos do projeto, em consonância com os objetivos da chamada GCE, vislumbrou-se a utilização da vigilância em saúde em sinergia com os avanços tecnológicos a fim de acelerar os esforços da eliminação da malária em níveis regional e nacional. Os eventos de malária ocorrem de forma dinâmica e complexa; entretanto, com o uso de soluções originais que agregam informações e permitem uma análise detalhada de dados, acredita-se poder apoiar na eliminação da doença.

A análise preditiva permite aos algoritmos encontrar diferentes padrões e comportamentos expressos nos dados, os quais não são visivelmente ou tão facilmente identificáveis por outros métodos ou ferramentas. Com o emprego da análise preditiva, busca-se reduzir custos, melhor alocar os recursos existentes, evitar e melhor controlar as doenças e possibilitar uma variedade de estudos.

1.2 OBJETIVOS GERAIS E ESPECÍFICOS

No contexto deste trabalho, a análise preditiva foi usada para o desenvolvimento e validação de modelos preditivos em relação ao número de casos de malária, tendo por base dados dos sistemas de notificação e dados climáticos.

Para alcançar o objetivo principal, alguns objetivos específicos foram considerados:

1. Integração de dados de malária disponíveis em diferentes sistemas de informação: SIVEP, SINAN e SIM.

2. Execução de análises descritivas dos dados, de modo a construir uma base de conhecimento a ser usada durante o desenvolvimento dos modelos preditivos.

3. Integração com outras fontes de dados , como dados climáticos, de modo que os modelos preditivos considerem também dados referentes a fatores que colaboram para a propagação de vetores de transmissão (incidência da doença) ou impactam no seu tratamento.

4. Definição e desenvolvimento de modelos analíticos usando algoritmos de aprendizado de máquina.

5. Realização de testes usando as bases de dados do cenário epidemiológico da malária no Brasil e posterior validação dos resultados.

1.3 METODOLOGIA

A metodologia empregada neste trabalho para a concepção dos modelos preditivos propostos tomou por base a revisão sistemática da literatura acerca de modelos preditivos para doenças transmissíveis (em uma abordagem mais geral) e, posteriormente, para o caso específico de malária.

Procurou-se identificar um conjunto mínimo e comum de atributos usados em diferentes modelos preditivos e se este conjunto poderia ser mapeado para os dados presentes nos bancos de dados disponíveis.

1.4 ESTRUTURA DO TEXTO

Este texto está organizado da seguinte forma: o capítulo 2 apresenta uma breve revisão da literatura, descrevendo os principais conceitos e características relacionados aos tema da dissertação. O capítulo 3 detalha os recursos utilizados em termos de softwares e bases de dados, como também a criação da Base Nacional de Notificações de Malária (BNNM). O capítulo 4 apresenta os experimentos realizados e os resultados. O capítulo 5 temos a discussão,conclusão e trabalhos futuros.

Capítulo

2

Nesta seção são descritos os principais conceitos diretamente relacionados ao trabalho desenvolvido .

FUNDAMENTAÇÃO TEÓRICA

2.1 DESCOBERTA DE CONHECIMENTO

A Descoberta de Conhecimento (*Knowledge Discovery in Databases* – KDD) é um processo empregado para a identificação de padrões e de associações presentes em conjuntos de dados, os quais não são trivialmente percebidos sem o auxílio de técnicas computacionais e estatísticas (KLOSGEN; ZYTKOW, 2002).

Esse processo consiste em várias etapas, as quais vão desde a obtenção e armazenamento dos dados, passando pela aplicação de diferentes técnicas para a execução de tarefas específicas e chega à visualização e interpretação de tais dados, conforme sintetizado na Figura 2.1.

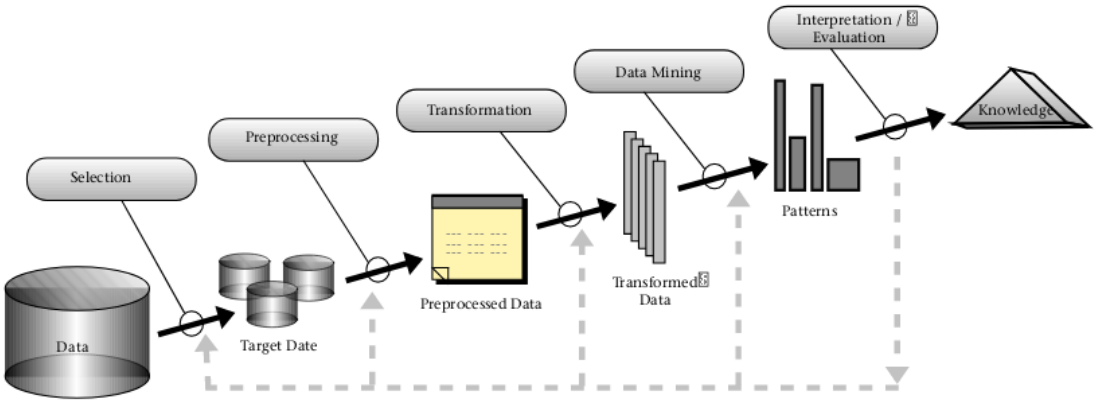


Figura 2.1 Etapas do processo de descoberta de conhecimento (Fonte:(FAYYAD et al., 1996)).

Uma das primeiras formalizações do KDD foi proposta por (FAYYAD et al., 1996). O processo consistia em várias etapas sequenciais, mas com possibilidade de iteração, conforme sumarizado a seguir:

1. Desenvolvimento e compreensão do domínio da aplicação: adquirir conhecimento relevante do domínio em questão e os objetivos do usuário final.

2. Criação do conjunto de dados alvo: selecionar atributos que serão usados na tarefa de descoberta.

3. Pré-processamento dos dados: remover inconsistências, padronizar e empregar estratégias para tratamento de dados ausentes.

4. Redução e projeção de dados: encontrar atributos úteis para representação dos dados, por meio de redução de dimensionalidade ou outros métodos a fim de reduzir o número de atributos.

5. Escolha das tarefas de mineração de dados: escolher a categoria de algoritmos de mineração, tais como sumarização, classificação, regressão e agrupamento, e decidir quais parâmetros são mais adequados.

6. Mineração de dados: gerar padrões a partir dos dados analisados, tais como regras de classificação, modelos de regressão, tendências, agrupamentos etc.

7. Interpretação: interpretar os padrões encontrados e retornar, caso necessário, às etapas anteriores para otimizar o processo.

8. Usar o conhecimento descoberto: definir ações baseadas no novo conhecimento, registrar e reportar aos interessados.

O processo não é trivial; logo, as etapas devem ser realizadas na ordem, com a possibilidade de repetição. No primeiro momento, é preciso selecionar os dados que serão explorados. Posteriormente, é necessário realizar o pré-processamento para que a qualidade dos dados não interfira no desempenho dos algoritmos na etapa de mineração, ao eliminar aqueles que são redundantes e inconsistentes, entre outros tipos de limpeza. É preciso normalizar os dados para garantir que os mesmos estejam na mesma escala. Após todas as etapas referidas anteriormente, é possível executar a mineração dos dados por meio de um conjunto de diferentes de algoritmos.

2.2 APRENDIZADO DE MÁQUINA

Aprendizado de máquina é um dos segmentos da inteligência artificial que tem por objetivo desenvolver técnicas para ensinar o computador a aprender determinadas atividades a partir de exemplos, dados e experiência (SOCIETY, 2017). Outra definição, proposta em (MITCHELL, 1997), considera o AM como "a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência". Ao longo dos últimos

anos essas técnicas vêm sendo adotadas em larga escala em campos complexos e intensivos, tais como Medicina, Astronomia e Biologia, fornecendo a possibilidade de extrair conhecimento oculto ou não-trivialmente perceptível.

Segundo (FACELI et al., 2011a), os algoritmos de AM podem ser classificados, de acordo com o tipo de tarefa que implementam, em supervisionados e não-supervisionados, conforme ilustrado na Figura 2.2.

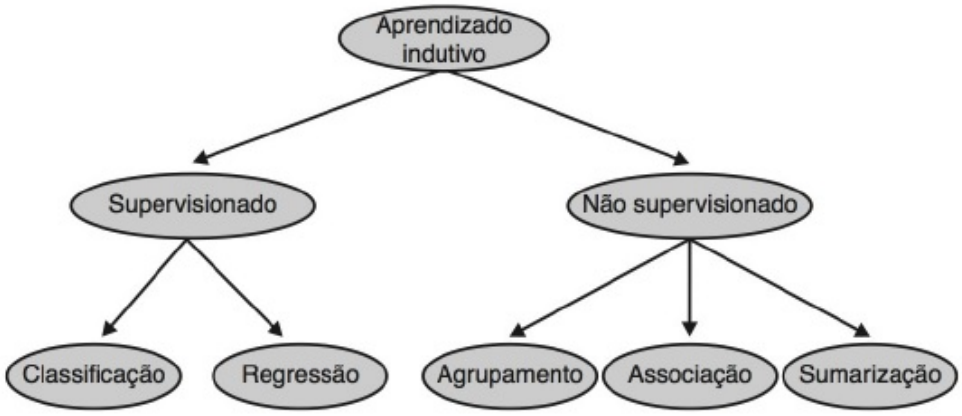


Figura 2.2 Hierarquia do aprendizado de máquina (Fonte: (FACELI et al., 2011a)).

A figura 2.3 detalha exemplos do emprego das diferentes técnicas de AM, com os respectivos algoritmos (ou métodos) mais representativos em cada categoria.

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set.	Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors	Assigning voters into known buckets by political parties eg: soccer moms. Bucketing new customers into one of known customer groups.
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from known data set.	Linear regression, Logistic regression	Predicting unemployment rate for next year. Estimating insurance premium.
Anomaly detection	Predict if a data point is an outlier compared to other data points in the data set.	Distance based, Density based, LOF	Fraud transaction detection in credit cards. Network intrusion detection.
Time series	Predict if the value of the target variable for future time frame based on history values.	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	K means, density based clustering - DBSCAN	Finding customer segments in a company based on transaction, web and customer call data.
Association analysis	Identify relationships within an itemset based on transaction data.	FP Growth, Apriori	Find cross selling opportunities for a retailor based on transaction purchase history.

Figura 2.3 Categorias e algoritmos de AM (Fonte: (KOTU; DESHPANDE, 2015)).

2.2.1 Aprendizado de Máquina Supervisionado

As tarefas preditivas ou supervisionadas têm, como processo de aprendizado, um conjunto de regras a partir de instâncias, ou seja, generalizar uma função que possa classificar novas instâncias. Neste contexto, cada instância possui um atributo resultante da classificação, o qual é considerado o rótulo daquela instância. Assim, um conjunto de instâncias referentes a um desfecho médico pode ter um rótulo com valor “presente” ou “ausente”, indicando se uma determinada doença foi ou não confirmada.

Se um especialista estiver disponível, ele pode apontar quais atributos são mais informativos; do contrário, pode-se utilizar técnicas de força bruta, embora este método não possa ser utilizado diretamente, sendo necessário um maior esforço na etapa de pré-processamento (MAGLOGIANNIS, 2007). A Figura 2.4 ilustra a abordagem supervisionada.

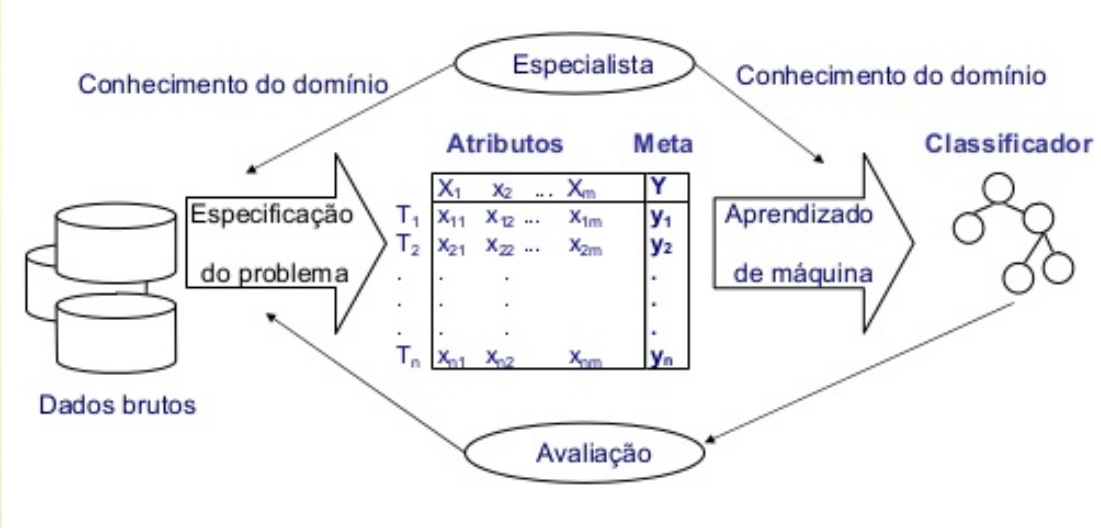


Figura 2.4 Aprendizado de máquina Supervisionado (Fonte: (OLIVEIRA, 2017).)

Uma técnica de AM preditiva é uma função que constrói um estimador a partir do conjunto de dados. O rótulo recebe valores em domínio conhecido: se os rótulos forem nominais, trata-se de um problema de classificação; se o domínio for um conjunto infinito e ordenado de valores, trata-se de um problema de regressão (FACELI et al., 2011a), conforme representado na Figura 2.5

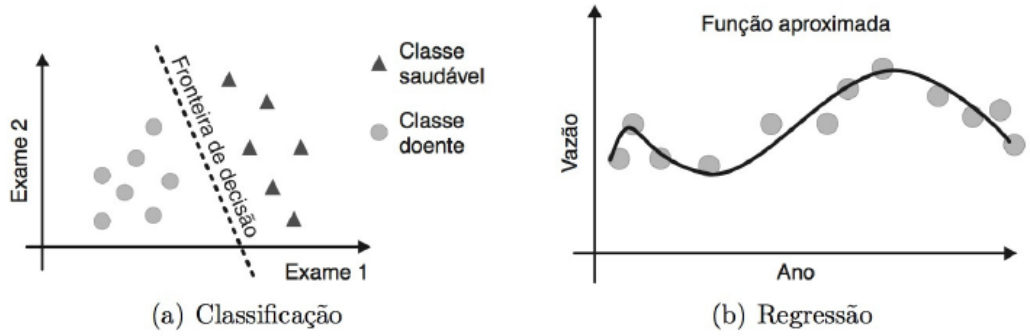


Figura 2.5 Tarefas de classificação e regressão. (Fonte: (FACELI et al., 2011a).)

As tarefas supervisionadas são classificadas, de acordo com o tipo de rótulo, em discretas (caso da classificação) ou contínuas (caso da regressão) (FACELI et al., 2011a). As figuras 2.6 e 2.7 são exemplos de bases de dados usadas para classificação e regressão, respectivamente.

sepal_length	sepal_width	petal_length	petal_width	Iris_class
5	2	3.5	1	versicolor
6	2.2	4	1	versicolor
6.2	2.2	4.5	1.5	versicolor
6	2.2	5	1.5	virginica
4.5	2.3	1.3	0.3	setosa
5.5	2.3	4	1.3	versicolor
6.3	2.3	4.4	1.3	versicolor
5	2.3	3.3	1	versicolor
4.9	2.4	3.3	1	versicolor
5.5	2.4	3.8	1.1	versicolor
5.5	2.4	3.7	1	versicolor
5.6	2.5	3.9	1.1	versicolor
6.3	2.5	4.9	1.5	versicolor
5.5	2.5	4	1.3	versicolor
5.1	2.5	3	1.1	versicolor
4.9	2.5	4.5	1.7	virginica
6.7	2.5	5.8	1.8	virginica
5.7	2.5	5	2	virginica
6.3	2.5	5	1.9	virginica
5.7	2.6	3.5	1	versicolor
5.5	2.6	4.4	1.2	versicolor
5.8	2.6	4	1.2	versicolor

Categorical value

Figura 2.6 Base de dados Íris para tarefas de classificação. (Fonte: (LICHMAN, 2013).)

	Cycle time (ns) MYCT	Main memory (KB) Min. MMIN	Main memory (KB) Max. MMAX	Cache (KB) CACH	Channels Min. CHMIN	Channels Max. CHMAX	Performance PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
3	29	8000	32000	32	8	32	220
4	29	8000	32000	32	8	32	172
5	29	8000	16000	32	8	16	132
...							
207	125	2000	8000	0	2	14	52
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

Figura 2.7 Base de dados CPUxPerformance para tarefas de regressão. (Fonte: (LICHMAN, 2013).)

2.2.2 Algoritmos de Aprendizado de Máquina Supervisionado

2.2.3 Análise Preditiva

A análise de dados (*data analytics*) é uma categoria de aplicações que se utiliza de técnicas de aprendizado de máquina e de métodos estatísticos para o manejo e a interpretação de diferentes tipos de dados (estruturados e não-estruturados) (GARTNER...,). Esta análise pode ser feita de forma descritiva, preditiva, prospectiva ou diagnóstica.

A análise descritiva é a forma mais básica de análise, pois informa o que já aconteceu (por exemplo, quantos casos de uma determinada doença ocorreram em determinado período). Esse tipo de análise sumariza os eventos que já ocorreram de forma humanamente legível. É o primeiro passo para entender o conjunto de dados para posterior transformação. Normalmente os resultados são apresentados por meio de relatórios e gráficos. Esse tipo de análise ajuda em tarefas de gerenciamento de saúde, como por exemplo, quantos pacientes vivem com um determinada doença, resultados de referência contra expectativas do governo ou área para melhorar medidas de qualidade clínica, entre outros aspectos. Esse nível elementar ainda permanece fora do alcance de muitas organizações, segundo (HEALTHCARE..., 2015).

A análise preditiva é a capacidade de utilizar os dados para prever o que pode acontecer no futuro. Assim, busca-se evidências para reduzir custos desnecessários e evitar penalidades por não controlar doenças crônicas ou evitar eventos adversos que podem ser prevenidos. Esse tipo de análise está longe da maioria das organizações, pois exige muito mais que somente ler eventos históricos, uma vez que é preciso acesso, em tempo real, infraestrutura adequada e pessoal técnico qualificado.

Apesar do interesse de grande parte da comunidade, ainda existem muitos desafios tecnológicos. As organizações que conseguem eliminar os obstáculos e utilizar esse tipo de análise em grande escala conseguem trazer benefícios para o paciente e para a gestão da saúde, como pontuações de risco preditivo para prevenção de suicídios, aumento de vigilância em UTIs, risco de readmissão ou aumento de uma doença em determinada região. (HEALTHCARE..., 2015).

Segundo (BIG..., 2013), um modelo prescritivo também é preditivo, pois se considera que um modelo prescritivo representa vários modelos preditivos em paralelo, um para cada ação de entrada possível. É o tipo de análise utilizada para chegar a um resultado-alvo. O modelo é capaz de prever as consequências com base em diferentes escolhas de entrada. Então, devem existir pelo menos dois componentes para ir de preditivo a prescritivo:

1. Os gestores dos dados devem conseguir escolher entre quais ações tomar com base no resultado previsto do modelo.

2. O sistema de realimentação (*feedback*) deve ser capaz de rastrear os resultados ajustados nas ações tomadas. O modelo deve ser inteligente o suficiente para assimilar a relação complexa entre a ação do gestor e o resultado através dos dados de realimentação.

A figura 2.8 mostra uma visão geral das três principais técnicas de análise de dados.

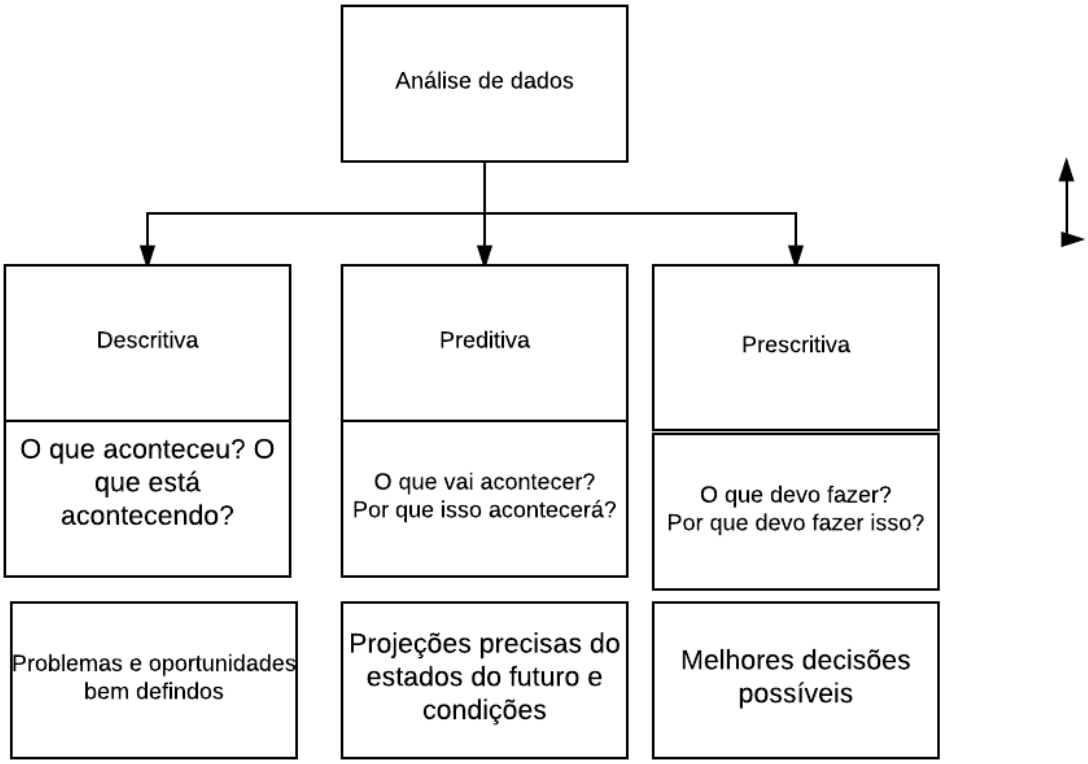


Figura 2.8 Principais categorias da análise de dados. (Fonte: (DELEN; DEMIRKAN, 2013).)

2.3 MODELOS PREDITIVOS

Com o enorme aumento da quantidade de dados subutilizados, os cuidados com saúde precisam de pesquisas e ferramentas mais eficientes para conseguir extrair informação de valor dos dados. Através dessa rica fonte, é possível maior entendimento e melhores cuidados com a saúde; embora o paradigma de dados ainda apresente muitos desafios (CHAWLA; DAVIS, 2013).

Existem dois paradigmas para criação de modelos preditivos: baseado em modelos (*model-driven*) e baseado em dados (*data-driven*). O *model-driven* é o tipo mais comum, pois a partir de uma ideia inicial bem formulada, com parâmetros bem definidos de como o sistema interage e como ele não funciona, gera-se uma hipótese do resultado ao qual se deseja chegar. Tenta-se, então, chegar a esse resultado a partir de experimentos. Esse tipo de abordagem tem como limitação a complexidade, mas por outro lado, é interessante já que permite um maior entendimento e compreensão do sistema e utiliza-se de relações já conhecidas cientificamente.

O *data-driven* é mais recente e está em alta por conta da disponibilidade de grandes bases de dados e de modelos (algoritmos) de AM que os suportam. Nessa abordagem, o modelo procura relações que não foram imaginadas *a priori*, procurando identificar estruturas e os relacionamentos automaticamente.

A primeira etapa para a construção de um modelo preditivo é a seleção das variáveis preditoras relevantes para o modelo. Na literatura não há consenso sobre como realizar essa etapa da melhor forma. A seleção inadequada de variáveis é um problema importante e comum que pode resultar em modelos ineficientes. Com o emprego de técnicas de AM, o problema da seleção de variáveis é menor pois não são utilizadas hipóteses e conhecimento prévio (WALJEE; HIGGINS; SINGAL, 2014).

2.3.1 Predição de Séries Temporais utilizando aprendizado de máquina

Uma série temporal pode ser definida como uma sequência S de medições históricas de uma variável observável y em intervalos de tempo t iguais. O estudo das séries temporais tem diversos objetivos, dentre eles a previsão do futuro com base no conhecimento do passado. A predição de valores futuros tem um papel importante em muitos campos da ciência (BONTEMPI; TAIEB; BORGNE, 2013).

É possível utilizar a série histórica de um problema de séries temporais para um problema de aprendizado supervisionado. Este tipo de aprendizado consiste em um conjunto finito de observações e a relação entre um conjunto de variáveis de entrada e uma ou mais variáveis de saída, que são consideradas as variáveis dependentes. Pode-se utilizar n valores passados como treino e os valores posteriores como teste (BONTEMPI; TAIEB; BORGNE, 2013).

2.3.2 Validação de modelos preditivos

Durante a aplicação dos algoritmos de AM a problemas do mundo real, normalmente os cientistas de dados não têm conhecimento de certos aspectos do problema. Dessa forma, para mensurar a qualidade da predição, usa-se unicamente um conjunto de exemplos rotulados e bem conhecidos. Essa abordagem não permite que se afirme, a priori, qual modelo (algoritmo) irá gerar a melhor solução para um determinado problema. Como estratégia exploratória, é normal se utilizar das características do problema e dos algoritmos para melhor escolher uma lista de modelos candidatos. Para cada algoritmo candidato, é possível realizar ajustes de parâmetros a fim de se testar diferentes configurações e observar os respectivos resultados. Dessa forma, fica evidente a natureza experimental do domínio de AM (FACELI et al., 2011b).

É necessário avaliar o resultado da predição usando o que foi predito de forma genuína. Normalmente, os dados são separados em duas amostras: treinamento e teste. Enquanto a amostra de treinamento é utilizada para a estimação dos parâmetros do modelo, a amostra de teste é utilizada para avaliar a precisão (acurácia/erro da predição) do modelos. Como a amostra de teste não é utilizada durante o treinamento do modelo, ela representa uma boa indicação de como o modelo processa e toma decisões sobre novas amostras de dados. Esse método de validação pode ser visto na figura 2.9, e é chamado normalmente de *hold-out set*. (HYNDMAN; ATHANASOPOULOS, 2014).

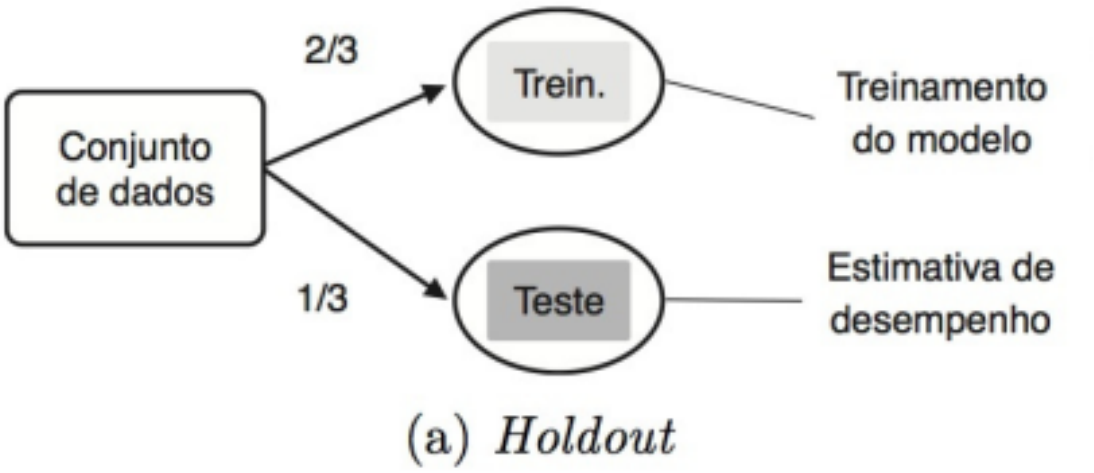
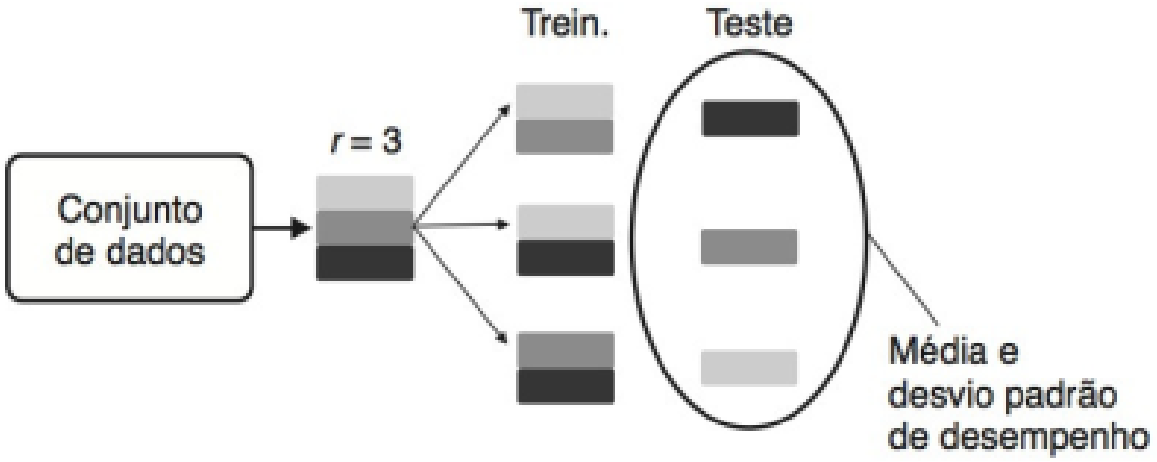


Figura 2.9 Técnica de validação *hold-out* (Fonte: (FACELI et al., 2011a)).

Um método mais sofisticado é a validação cruzada (*cross validation*), na qual o conjunto de dados é subdividido em r subconjuntos de tamanho aproximadamente igual, dos quais *r-1* são utilizados como amostras de treinamento, e o subconjunto restante é usado como amostra de teste. A cada ciclo de validação, um subconjunto fica como teste e o restante como treino. Ao final, o desempenho é calculado através da média de desempenhos observados em cada rodada de validação com diferentes subconjuntos de teste (FACELI et al., 2011b). A validação por validação cruzada pode ser vista na figura 2.10



(c) Validação cruzada

Figura 2.10 Técnica de validação cruzada (Fonte: (FACELI et al., 2011a)).

Entretanto, um método mais adequado para validar séries temporais é chamado de validação cruzada para séries temporais (ou *evaluation on a rolling forecasting origin*). Neste tipo de validação, o conjunto de testes pode ser uma única observação ou mais de uma, enquanto o conjunto de treinamento consiste, apenas, em observações que ocorreram antes das observações de teste. Dessa forma, nenhuma observação futura é utilizada para a construção do modelo. Na figura 2.11, os pontos azuis representam as observações do conjunto de treinamento, e o ponto vermelho representa a observação de teste. Em cada linha, tem-se uma nova previsão que avança no tempo, sendo que a cada avanço, a observação de teste se torna parte do conjunto de treinamento e uma nova observação é então alocada como o novo conjunto de teste (HYNDMAN; ATHANASOPOULOS, 2014).

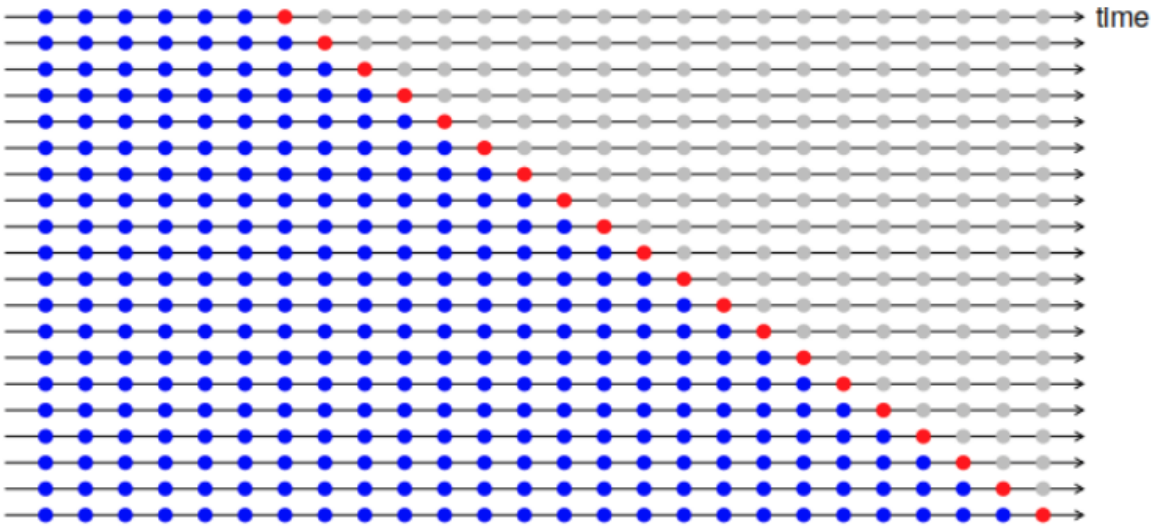


Figura 2.11 Validação cruzada para séries temporais (*rolling origin*) (Fonte: (HYNDMAN; ATHANASOPOULOS, 2014)).

A validação cruzada para séries temporais , pode-se subdividir em duas formas:

1. Com janelas deslizantes ou fixas, quando o conjunto de treino tem tamanho fixo e avança no tempo, conforme ilustra a Figura 2.12, na parte superior. Neste modo pode-se observar que o conjunto de treino , os quadrados em vermelho, se mantém sempre com o mesmo tamanho. Enquanto o conjunto de teste, em cinza, sempre estará a frente no tempo em relação ao dados de treino. Por exemplo, de um a cinco como treino e seis como teste, numa segunda iteração, os dados de treino seriam de dois a seis, e o dado de teste seria sete, e assim por diante, utilizando o horizonte de teste igual a um.

2. Com janela de expansão, quando o conjunto de treino tem tamanho dinâmico e acumulativo, e a cada iteração do modelo, novas observações são adicionadas. Por exemplo, a primeira iteração, utilizando o horizonte de teste igual a um, conforme a Figura 2.12 na parte inferior a esquerda, tem-se de um a cinco como dados de treino e o seis como dado de teste. A cada iteração a base de treino avança em um passo no tempo, agrega o ponto que anteriormente era base de teste, e o novo ponto de teste passa a ser o sete, e assim por diante.

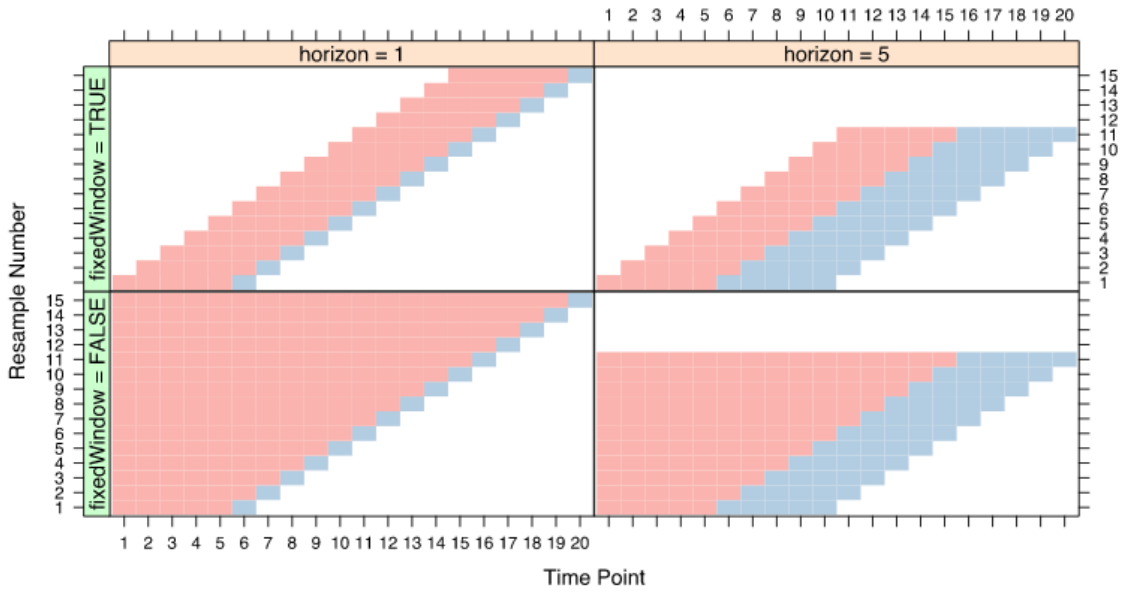


Figura 2.12 Técnica de validação cruzada (Fonte: (KUHN, 2009)).

A predição é então avaliada com base na média do erro sobre o conjunto de teste.

2.3.3 Predição Multi-Passos

Diante de um problema de predição multi-passos, é preciso escolher qual estratégia de previsão utilizar. As duas principais abordagens são a recursiva e a direta. A estratégia recursiva utiliza somente um modelo, sendo essa sua vantagem. Entretanto, esse tipo de abordagem utiliza o valor predito como parâmetro para as próximas predições; dessa forma essa abordagem é sensível a acumulação de erros. A estratégia direta difere por não utilizar o valor predito como parâmetro do modelo, não ocorrendo dessa forma acumulação de erros. A desvantagem é que cada predição para um determinado horizonte tem como entrada dados reais, ou seja, para cada horizonte diferente é preciso um modelo, sendo assim aumenta-se a necessidade de poder de computacional (BONTEMPI; TAIEB; BORGNE, 2013).

2.4 TRABALHOS RELACIONADOS

Existem na literatura diversos estudos que objetivam realizar predição com o escopo em malária , utilizando número de casos, para fornecer aos serviços de saúde informações para melhor planejamento estratégico , a partir de diversas dimensões de tempo, e usando diferentes técnicas de modelagem , normalmente AM e/ou estatísticas.

Em (GIROND et al., 2017), é proposta a modelagem estatística por meio do método SARIMA e utilizadas variáveis climáticas/ambientais como temperatura, precipitação e Índice de Vegetação da Diferença Normalizada (NVDI). Os dados sobre a malária de Madagascar foram obtidos junto ao programa de controle da malária deste país. Dados sobre controle de vetores como pulverização residual interna e distribuição de redes de

longa duração com inseticidas também foram adicionados ao modelo. A granularidade dos dados foi semanal, e foram utilizadas as métricas RMSE e MASE para avaliar o modelo preditivo.

(MODU et al., 2017) aplicou vários algoritmos de AM como *Support Vector Machines* (SVM),*K-Nearest Neighbours* (KNN) e *Decision tree* para predição da incidência da malária no distrito de Ejisu-Juaben em Gana. Os preditores foram variáveis climáticas como umidade relativa, precipitação, temperaturas máxima e mínima e velocidade do vento. Os dados foram discretizados para se trabalhar com classificação. Portanto, foram utilizadas técnicas de agrupamento para escolher as classes. Para validar o modelo, foi utilizado *10-fold cross validation*. SVM foi o algoritmo com a taxa de acerto mais alta.

(ADIMI et al., 2010) utilizou dados de vinte e três províncias do Afeganistão de 2003 a 2004. Dados meteorológicos e ambientais foram utilizados como precipitação, temperatura e índice de vegetação. Além disso, foi utilizada a modelagem estatística por meio da regressão linear. Os últimos seis meses da série temporal foram utilizados para validação e; o restante da série, para treinamento. Como resultado, obteve-se que o NVDI é um forte preditor,seguido pela temperatura e, por último, a precipitação com um baixo poder preditivo. A granularidade dos dados foi mensal e, para validação, a métrica RMSE foi utilizada.Como resultado, foi encontrada uma forte relação de dependência das infecções dos meses anteriores ao predito e defasagens maiores, como dois e três meses.O autor enfatiza que, além de fatores ambientais; outros como, socioeconômicos, serviço de saúde e movimento de refugiados, podem influenciar na transmissão da doença. Entretanto, esses fatores normalmente não mudam com frequência, o que possibilita a predição da malária com base em parâmetros ambientais.

(GOMEZ-ELIPE et al., 2007) utilizou o método ARIMA para realizar predição em dados de granularidade mensal de notificação de malária. As variáveis preditoras incluem precipitação, temperatura e NVDI. Um ponto relevante o qual o autor menciona é em relação ao fatores que podem influenciar pela variabilidade da malária em razão da resistência do parasita aos antimaláricos comuns e a inseticidas, a movimentos populacionais e a infecções subjacentes como por exemplo ao Vírus da Imunodeficiência Humana (HIV)

No estudo realizado por (CUNHA et al., 2010), foram utilizadas redes neurais para realizar predição da incidência da malária para o município de Cantá-Roraima. Os dados foram provenientes do Sistema de Informação de Vigilância Epidemiológica (SIVEP) do período de 2003 a 2009, e o resultado foi comparado com um modelo de regressão logística, sendo que a métrica de validação foi o RMSE. O estudo constatou que as redes neurais obtiveram menor erro em relação ao modelos de regressão, e o autor, no texto, não menciona quais atributos foram utilizados como preditores.

No trabalho realizado por (KALIPE; GAUTHAM; BEHERA, 2018), foram utilizados algoritmos de classificação popular e recente, como regressão linear, regressão logística, máquina de vetores de suporte, vizinhos mais próximos, *eXtreme Gradient Boosting*, floresta aleatória e Redes Neurais. A base de dados utilizada de 2005 a 2011 foi classificada em três níveis de risco denominados alto, médio e baixo. Os modelos foram validados, utilizando o método de amostragem *holdout*. Os algoritmos foram comparados utilizando diversas métricas como *accuracy*, *precision*, *recall* e *Matthews Correlation*

Coefficient. O algoritmo *eXtreme Gradient Boosting* foi o mais eficiente.

No contexto da predição em malária, é importante evidenciar a revisão realizada por (SILVA et al., 2019b) o qual mostra quais ações devem ser executadas no âmbito do controle de vetores no Brasil, pois se trata de componente essencial para prevenção da doença e útil do ponto de vista de modelos preditivos.

A partir da leitura dos trabalhos relacionados acima, além de outros, é possível verificar que a malária é uma doença afetada por múltiplos fatores (climáticos e ambientais), e esse processo complexo dificulta a predição acurada e precisa. Entretanto, os sistemas de suporte à decisão podem ajudar a entender e direcionar recursos de saúde pública para programas de eliminação de forma mais rápida, ajudando a reduzir a transmissão e no melhor manejo dos recursos.

Capítulo 3

DESENVOLVIMENTO DOS MODELOS PREDITIVOS

A metodologia proposta está baseada no fluxo padrão do Knowledge Discovery Process (KDP), ilustrado na figura 2.1. A primeira etapa foi a aquisição das bases de dados necessárias ao trabalho, as quais são descritas na seção 3.2.

A segunda etapa consistiu na análise descritiva dos dados presentes nestas bases para o levantamento de indicadores de qualidade (ou completude), bem como na necessidade de tarefas de padronização e limpeza dos dados.

Uma vez pré-processadas, as bases foram integradas e usadas para atividades de análise descritiva, dos pontos de vista estatístico e epidemiológico, para gerar subsídios ao modelo preditivo a ser desenvolvido.

As análises descritivas, tanto das bases de saúde, como das bases climáticas, ajudaram no desenvolvimento do modelo preditivo, por meio da análise da qualidade, relevância das variáveis e entendimento do domínio.

Para a concepção do modelo preditivo, foram utilizados algoritmos de aprendizado de máquina com o intuito de prever o número de casos de malária em alguns municípios específicos.

3.1 FERRAMENTAS UTILIZADAS

Para o desenvolvimento deste trabalho, foram usadas as seguintes ferramentas:

3.1.1 Apache Spark

É um *framework* para processamento de dados em paralelo em larga escala (APACHE. . . , 2017). Pode ser utilizado com as linguagens Java, Python e Scala, permitindo execução local (computador isolado) ou em um agregado (conjunto de computadores) gerenciado. O Spark provê suporte para linguagens de consulta de dados (SQL), bem como rotinas de aprendizado de máquina e manipulação de dados representados na forma de grafos.

3.1.2 R

O R é uma linguagem multiplataforma para implementação de soluções analíticas, o qual tem a capacidade de criar gráficos e pode ser executado de forma paralela e em conjunto com o Apache Spark. Permite a criação de visualizações interativas e há a disponibilidade de uso de diversos pacotes para a implementação do processo de KDD, como por exemplo, pré-processamento de bases de dados, análise descritivas, transformação e execução de modelos preditivos.

3.1.3 Bibliotecas de aprendizado de máquina

- scikit-learn: Biblioteca *open-source* para mineração de dados e análise de dados. Suporta diversos algoritmos de classificação, regressão, agrupamento, pré-processamento, redução de dimensionalidade, entre outros.

- caret: Biblioteca de aprendizado de máquina disponível no ambiente R. (KUHN, 2009).

3.2 BASE DE DADOS

Para realizar a análise proposta neste trabalho foi necessário obter um conjunto de bases de dados relacionadas ao Sistema Brasileiro de Monitoramento de Malária. As bases de dados não-identificadas concedidas ao projeto foram as seguintes: Sistema de Informação de Vigilância Epidemiológica - malária (SIVEP-malária), Sistema de Informação de Agravos de Notificação (SINAN-malária) e Sistema de Informação sobre Mortalidade (SIM-malária), conforme sumarizado na Tabela 3.1.

Tabela 3.1 Sistemas de informações de monitoramento de malária.

Sistema de informação	Período disponível
SINAN	2003 a 2015
SIVEP	2003 a 2018
SIM	2003 a 2015

3.2.1 Sistema de Informação de Vigilância Epidemiológica

O SIVEP-malária é o sistema oficial para armazenamento das notificações de malária da Amazônia Legal, desde a implantação do PNCM em 2003, abrangendo ao todo nove estados (Acre, Amapá, Amazonas, Pará, Rondônia, Roraima, Tocantins, Mato Grosso e Maranhão). A base de dados não-identificada do SIVEP-malária é referente ao período de 2003 a 2018 e contém 5.490.603 registros com 43 atributos. Todo registro é uma notificação com resultado positivo.

3.2.2 Sistema de Informação de Agravo de Notificação

O SINAN-malária é o sistema oficial para armazenamento das notificações de malária na região fora da Amazônia Legal. A base de dados não-identificada contém 43 e 79, respectivamente, para o período de 2003 a 2016 e de 2007 a 2015. Todo o período do SINAN contém 42.670 registros.

3.2.3 Sistema de Informação sobre Mortalidade

O SIM é o sistema de informação responsável por armazenar dados sobre mortalidade no Brasil. Para a presente pesquisa, foram extraídos do SIM todos os óbitos registrados,cerca de 1004, no período de 2003 a 2015, cuja causa do óbito está associada à doença.

3.2.4 Dados Climáticos

Os dados climáticos foram obtidos no NOOA, foram extraídos, por dia, para todos os municípios do Brasil, um arquivo para cada município. A primeira etapa foi agregar todos os arquivos em uma única base de dados por ano ,resultado em arquivos com todos os municípios do Brasil por ano.

As variáveis escolhidas foram:

- Precipitação
- Temperatura
- Umidade relativa do ar
- Escorrência superficial

Todas as variáveis de clima foram, então, sumarizadas para médias mensais ou o acumulado do mês a depender da variável, por município.

3.2.5 Municípios para validação do modelo preditivo

Para validar os modelos preditivos, foi selecionado um município para cada estrato do Índice Parasitário Anual (IPA), utilizando com referência os dados de 2018. A capital Manaus também foi escolhida por sua importância na região, uma vez que a cidade tem um dos maiores números de casos notificados positivos do Brasil, contudo, tem seu IPA classificado como baixo por conta da população total do município. O IPA segue a seguinte classificação, conforme a tabela 3.2

Índice Parasitário Anual	**Risco**
$\leq$9	Baixo
10 – 49,9	Médio
$\geq$50	Alto

Tabela 3.2 Classificação do Índice Parasitário Anual.

Os municípios da tabela 3.3 foram selecionados conforme o estrato do IPA.

Município	IPA
Manaus	Alto
São Gabriel da Cachoeira	Alto
Humaitá	Médio
Boca do Acre	Baixo

Tabela 3.3 Lista de municípios para validação dos modelos preditivos.

3.3 SISTEMAS DE NOTIFICAÇÃO DE MALÁRIA

As ações do PNCM relativas à vigilância da malária no Brasil podem ser prejudicadas por conta do uso de sistemas de informações heterogêneos como o SIVEP e SINAN. A diferença entre a especialização e a infraestrutura dos agentes de saúde, tanto na Amazônia Legal, quanto fora dela, como também a falta de planejamento para algumas regiões em relação a locais de produção, é fator de impacto no combate à malária. Porém, a falta de uma visão central em relação aos dados é uma questão desafiadora para o combate da doença, mesmo levando em consideração que a maioria dos casos são registrados no SIVEP, e os casos do SINAN levaram a um número significativo de mortes por atraso no tratamento.

No âmbito da pesquisa, a utilização dos sistemas de forma isolada não oferece uma visão geral da malária no brasil. Assim, é necessário que os pesquisadores escolham quais amostras usar e como lidar com problemas de pré-processamento, vinculação e harmonização dos conjuntos de dados. Dessa forma, uma base de dados integrada de notificações da malária pode ajudar a resolver alguns problemas mencionados, como também, disponibilizar outros dados relevantes como os bancos de dados climáticos, controle de vetores e de mortalidade.

A BNNM foi criada para unificar os registros do SIVEP e do SINAN. Os critérios de pré-processamento e harmonização dos dados são elencados a seguir.

Primeiramente foram trabalhados nos dados dos dois sistemas de informação de forma isolada, para posterior união dos dados. Foram selecionadas as seguintes variáveis das bases de dados individuais de notificação:

- Data de notificação
- Código do município de Notificação
- Código do município de Residência
- Código do município de Infecção
- Sexo
- Raça
- Idade
- Zona
- Tipo de Lâmina
- Sintomas
- Resultado do exame
- Data do exame
- Data dos primeiros Sintomas
- Data do tratamento
- ID LVC
- Gestante

- Semana epidemiológica
- País de infecção

Para o SIVEP e/ou SINAN, foram aplicadas as seguintes transformações:

- Correção das variáveis sexo e raça que estavam invertidas para os anos de 2003 a 2016 e para o ano de 2018.
- Correção de todas as variáveis selecionadas em relação ao dicionário de dados. Os valores não mapeados foram transformados para ignorado; o valor 999 foi escolhido para representar os ignorados e o valor 999999 foi usado para variáveis de código de município inexistentes.
- Criação da variáveis ano e mês a partir da data de notificação
- Alteração da variável categórica raça para string (ex: 1 para branca e 4 para parda).
- A variável resultado do exame, que continha onze categorias, foi recategorizada segundo consulta a especialista da área.
- As variáveis sintomas, zona e gestante foram transformadas de binárias para strings.
- Utilizando a variável idade, foi criada uma nova variável com as faixas etárias de 0 até 64 anos, contadas de 5 em 5, e uma faixa para 65 anos ou mais. Valores estranhos foram transformadas para 999.
- Foi criada a variável oportunidade de diagnóstico e a oportunidade de tratamento segundo o estudo de (BRAZ et al., 2016).
- Foi criada a variável autóctone-importado.
- Foi criado um algoritmo para gerar uma nova variável referente à semana epidemiológica. O banco original do SIVEP contém, originalmente, uma variável que mostra a semana epidemiológica da notificação, entretanto cerca de 2,5% dos dados têm inconsistências e não possuem a semana epidemiológica correta.
- Foram adicionadas, para todas as variáveis com código do IBGE, as versões de 6 e 7 dígitos.
- Foi criada uma variável para contabilizar todos os registros do banco com o valor numérico 1 (um).

Após o pré-processamento, as seguintes variáveis foram selecionadas:

- Código do município de notificação
- Código do município de residência
- Código do município de infecção
- Sexo
- Raça
- Faixa etária

- Zona
- Gestante
- Tipo de lâmina
- Sintomas
- Resultado do exame
- Oportunidade do tratamento
- Oportunidade do diagnóstico
- Autóctone/Importado
- Sistema de notificação
- Semana epidemiológica
- Ano de infecção
- Mês de infecção
- País de infecção

Realizado o pré-processamento e seleção das variáveis das duas bases de dados, foi então criado um algoritmo para compatibilizar os nomes para cada variável correspondente. Posteriormente, foi realizada uma operação de união a qual resultou em uma base de dados com todas as notificações individuais do Brasil, para o período de 2003 a 2018, com cerca de 5.987.002 registros, com os dados do ano de 2003 até o de 2015 para o SINAN e 2003 a 2018 para o SIVEP.

Foi preciso refletir sobre como disponibilizar o acesso de forma flexível à base de dados gerada. Primeiramente, a ideia foi criar variáveis para cada informação necessária (por exemplo, uma variável que teria a quantidade de gestantes sem sintomas da faixa-etária 20-24). Com esse tipo de especificidade, não haveria flexibilidade e seria preciso criar muitas variáveis e re-execuções da base de dados sempre que uma determinada informação fosse necessária.

Para solucionar esse problema, uma segunda ideia foi implementada: agrupamento por todas as variáveis do banco de dados e somatório dentro do agrupamento, resultando na variável com o total de casos para aquele conjunto de atributos. Com esta abordagem foram geradas, por exemplo, i) a quantidade de gestantes, sem sintomas, da faixa etária 20-24 e ii) a quantidade de gestantes com sintomas ou da faixa etária de 30-34. Em resumo, esta abordagem permite o cruzamento de quaisquer variáveis do banco de dados e a contagem das notificações. Na figura 3.1, pode-se observar, para o ano de 2018, diversos recortes para a variável gestante com diversas faixas etárias contabilizando o número de casos.

Base

Show 10 entries Search:

	ano_infec	sexo	resexame	faixa_etaria	sintomas	gestante	n_casos
36	2018	F	VIVAX	FAIXA_15-19	SIM	SIM	935
37	2018	F	VIVAX	FAIXA_20-24	SIM	SIM	859
53	2018	F	VIVAX	FAIXA_25-29	SIM	SIM	539
71	2018	F	VIVAX	FAIXA_30-34	SIM	SIM	319
83	2018	F	VIVAX	FAIXA_35-39	SIM	SIM	202
94	2018	F	VIVAX	FAIXA_10-14	SIM	SIM	128
104	2018	F	FALCIPARUM	FAIXA_15-19	SIM	SIM	107
108	2018	F	FALCIPARUM	FAIXA_20-24	SIM	SIM	97

Figura 3.1 Exemplo de cruzamento de variável da BNNM (Fonte: Autoria Própria).

3.4 ANÁLISE DESCRITIVA

Os dados para todos os municípios serão descritos a partir da visualização da decomposição da série temporal para a variável total de casos. No eixo x tem-se os anos e no eixo y tem-se o total de casos. Os gráficos seguintes ilustram a a série temporal,a tendência, a sazonalidade e por último o ruído. Por conta do tamanho da série os dados foram divididos em 2003 a 2010 e 2011 a 2018.

3.4.1 Análise descritiva – Manaus

A decomposição da série temporal para os casos positivos de Manaus para o período de 2003 a 2010 e 2011 a 2018 podem ser vistos, respectivamente, nas figuras 3.2 e 3.3. Percebe-se visualmente uma tendência de decrescimento ao longo de 2003 até 2010 e também entre 2011 a 2018, entretanto nesse último período tem-se picos mais destacados entre 2015 e 2018. Em *season year* é possível verificar a sazonalidade, onde pode-se perceber dois padrões: um entre 2003 e 2010 e outro a partir de 2010. Em *remainder*, tem-se o ruído da série.

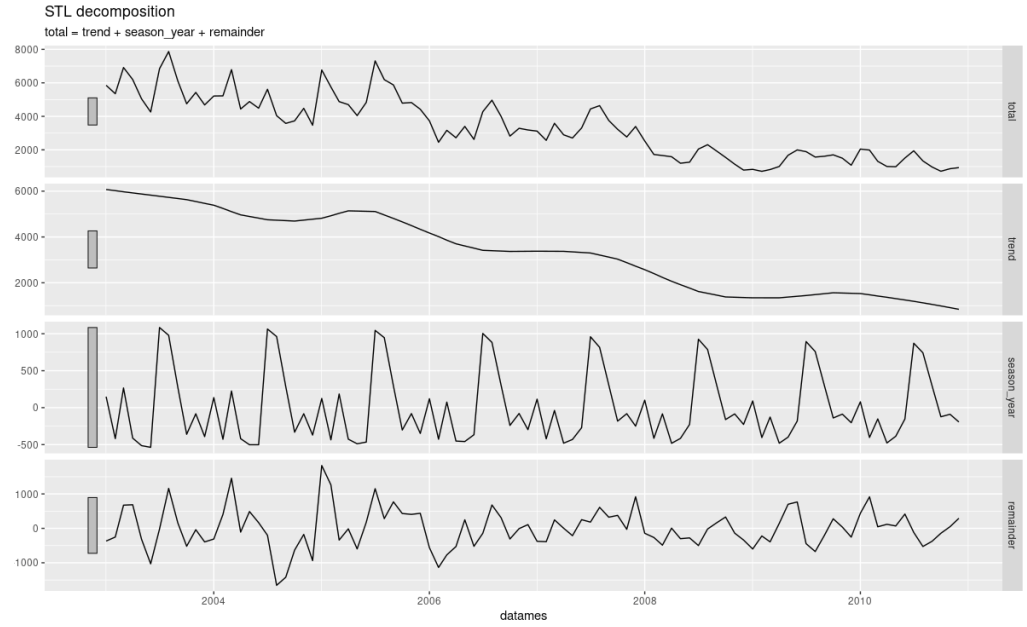


Figura 3.2 Decomposição da série temporal para o município de Manaus de 2003 a 2010.

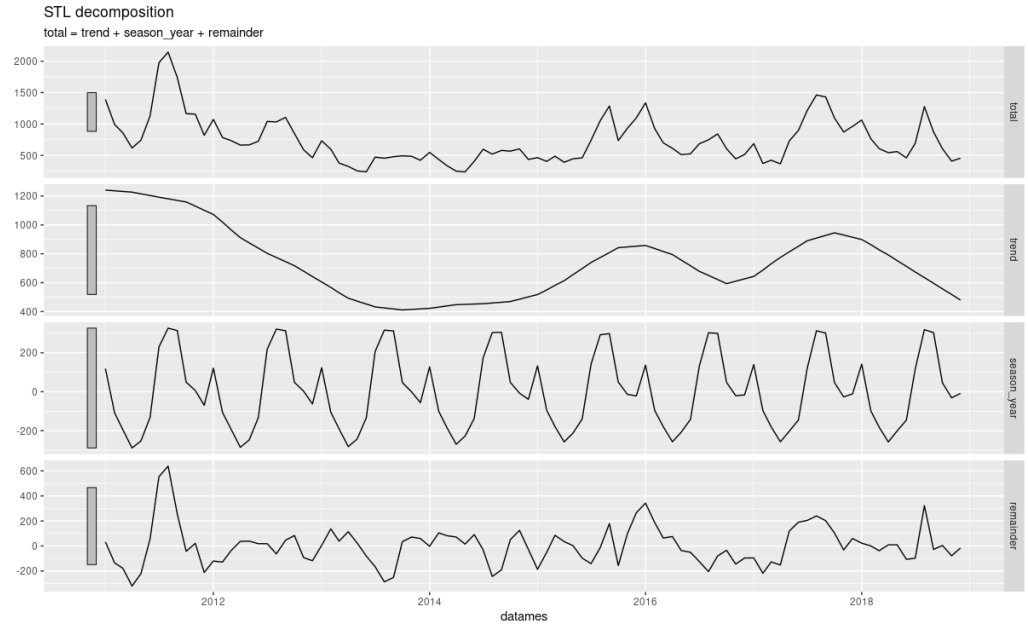


Figura 3.3 Decomposição da série temporal para o município de Manaus de 2011 a 2018.

Nas figuras 3.4 e 3.5 tem-se um gráfico que permite a identificação do período sazonal dos dados. O eixo x traz os diversos anos da série e o eixo y mostra o total de casos por mês. Esse tipo de gráfico permite visualizar o padrão subjacente com maior clareza. A partir das figuras de Manaus, pode-se perceber um aumento dos casos sempre nos meses

de junho a agosto e um aumento mais moderado entre fevereiro e março – esse último presente somente de 2003 a 2010.

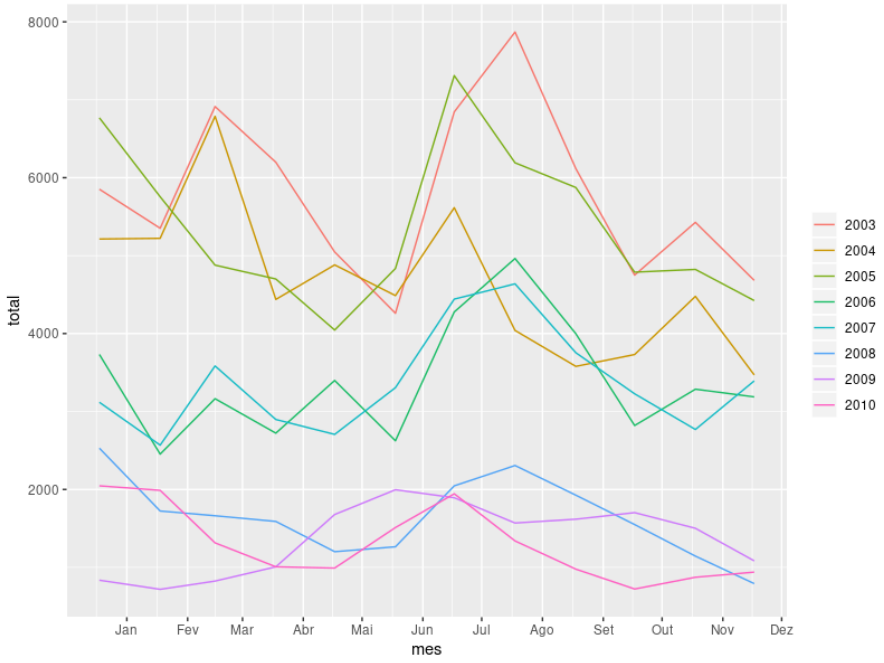


Figura 3.4 Gráfico sazonal para o município de Manaus de 2003 a 2010.

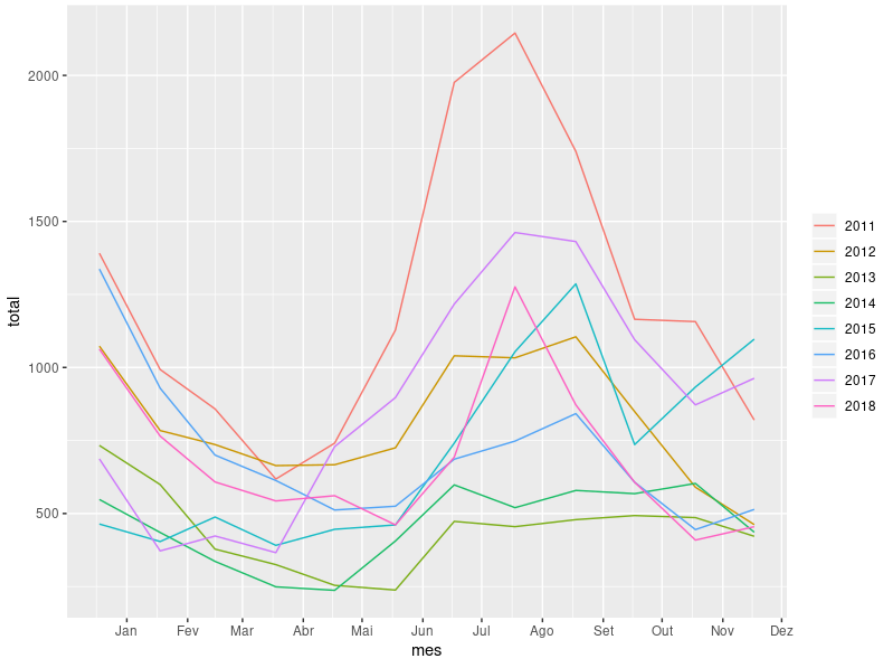


Figura 3.5 Gráfico sazonal para o município de Manaus de 2011 a 2018.

3.4.2 Aṅalise Descritiva – São Gabriel da Cachoeira

A decomposição da série temporal para os casos positivos de São Gabriel da Cachoeira para o período de 2003 a 2010 e 2011 a 2018 podem ser vistos respectivamente nas figuras 3.6 e 3.7. Percebe-se visualmente uma tendência- de crescimento ao longo de 2003 até 2010 e também entre 2011 a 2018. Em *season year* é possível verificar a sazonalidade, onde pode-se perceber dois padrões: um entre 2003 e 2010 e outro a partir de 2011. Em *remainder* tem-se o ruído da série.

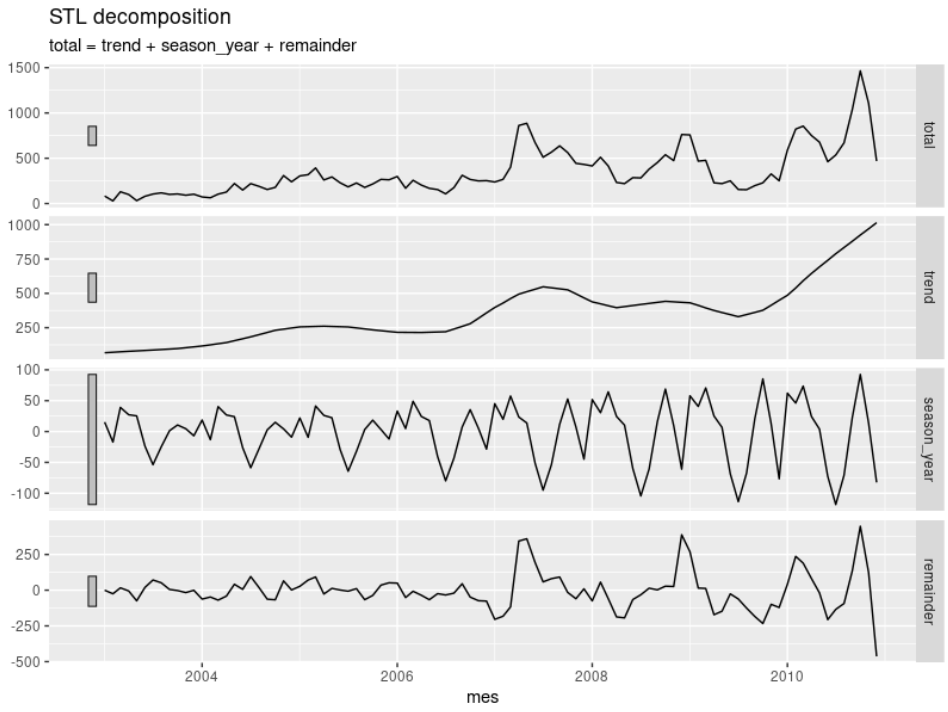


Figura 3.6 Decomposição da série temporal para o município de São Gabriel da Cachoeira de 2003 a 2010.

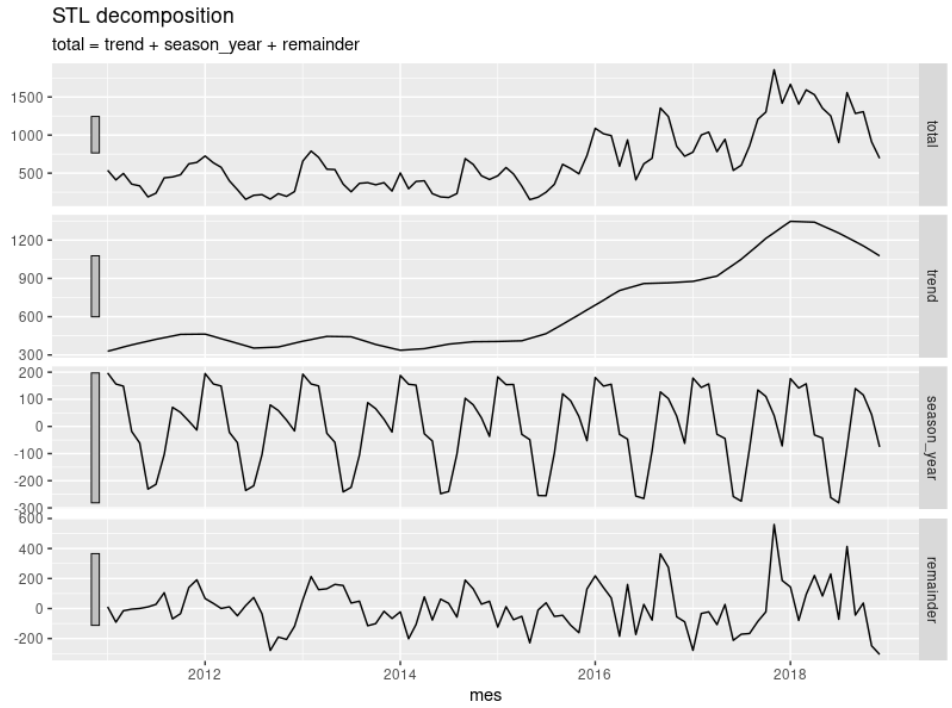


Figura 3.7 Decomposição da série temporal para o município de São Gabriel da Cachoeira de 2011 a 2018.

As figuras 3.8 e 3.9 são respectivamente do período de 2003 a 2010 e 2011 a 2018. É possível perceber visualmente um padrão de aumento de casos nos meses de julho a setembro para todo o período.

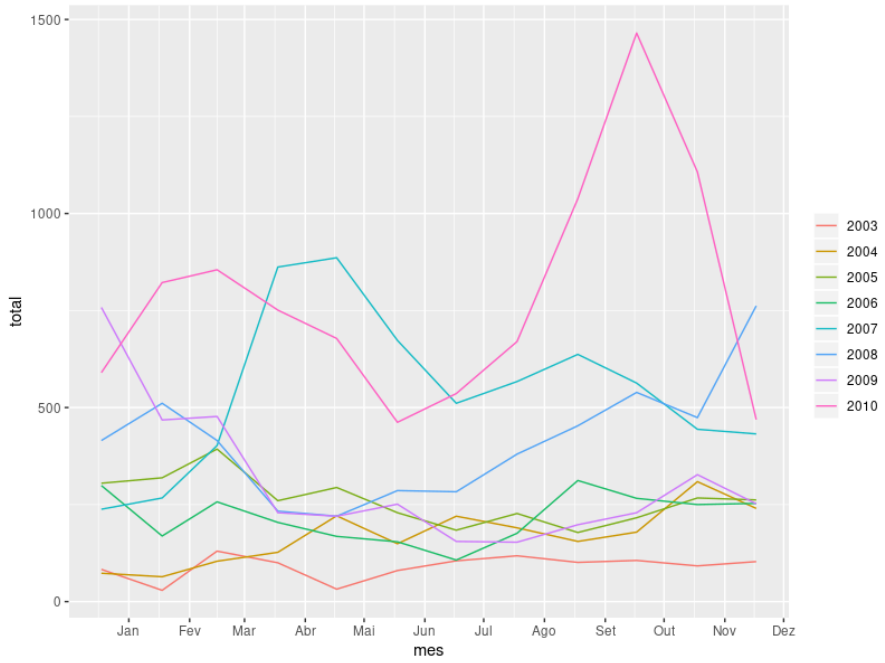


Figura 3.8 Gráfico sazonal para o município de São Gabriel da Cachoeira de 2003 a 2010.

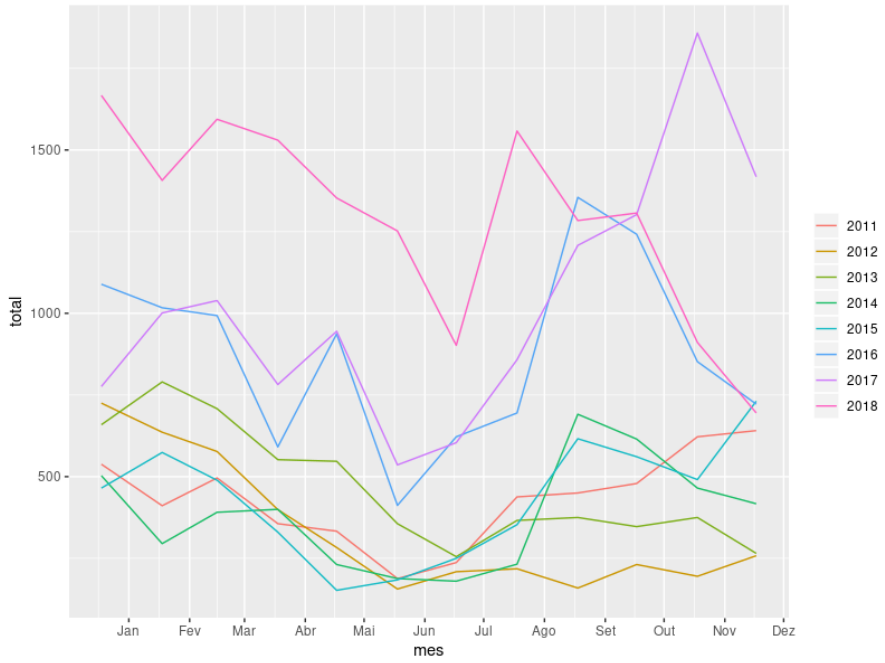


Figura 3.9 Gráfico sazonal para o município de São Gabriel da Cachoeira de 2011 a 2018.

3.4.3 Análise Descritiva – Boca do Acre

As figuras 3.10 e 3.11 são, respectivamente, do período de 2003 a 2010 e 2011 a 2018 para Boca do Acre. É possível perceber uma tendência decrescente do número de casos entre 2003 a 2010 com alguns picos em 2004 e 2007. Para o período de 2011 a 2018 a tendência é crescente, com aumento em 2012, decréscimo em 2017 e aumento novamente a partir de 2018. Em *season year* é possível perceber uma mudança do padrão sazonal quando comparado os dois períodos.

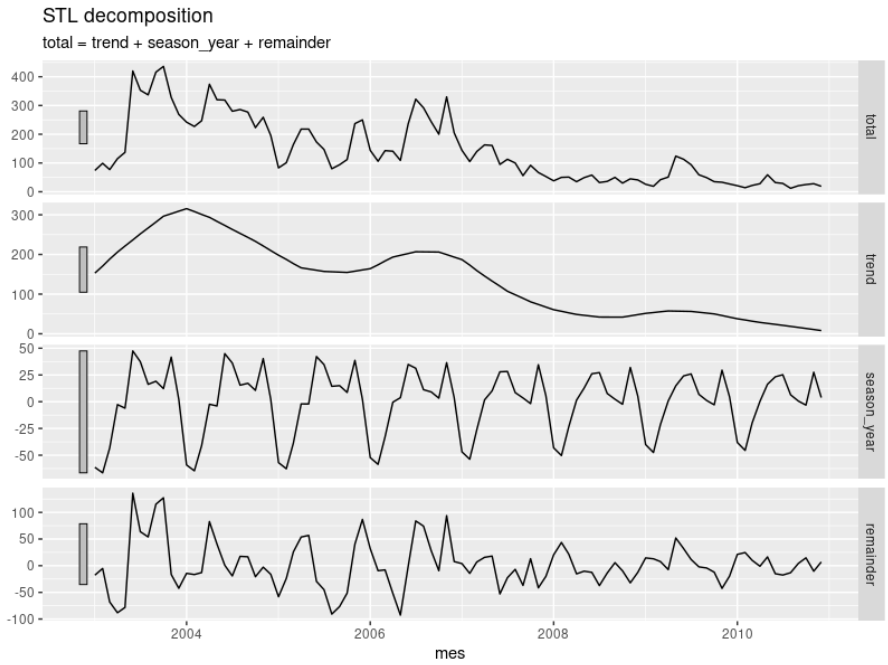


Figura 3.10 Decomposição da série temporal para o município de Boca do Acre de 2003 a 2010.

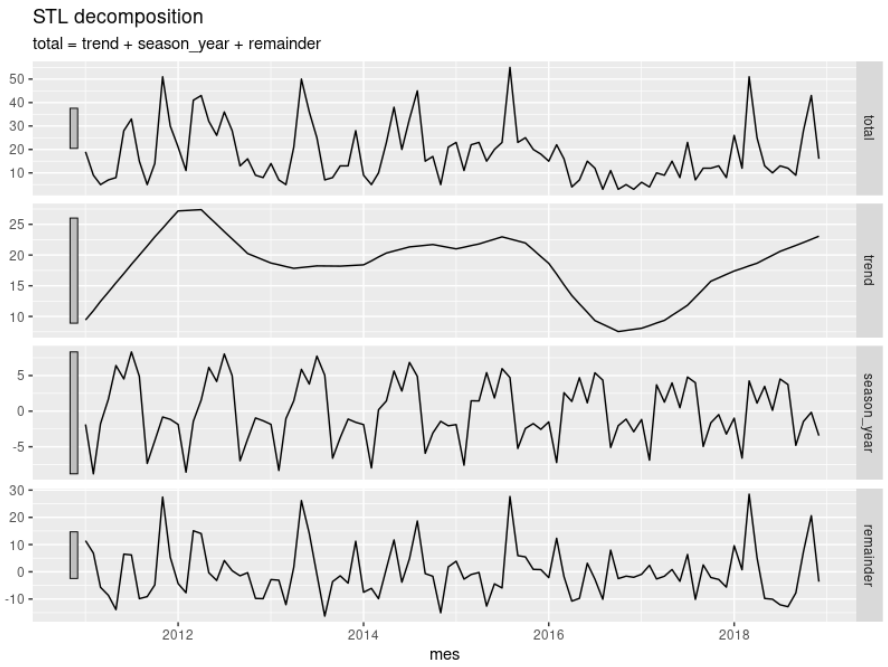


Figura 3.11 Decomposição da série temporal para o município de Boca do Acre de 2011 a 2018.

Nas figuras 3.12 e 3.13 são descritos os dados, respectivamente, para os períodos 2003 a 2010 e 2011 a 2018. Neste caso, não há um padrão visual que permita a fácil identificação da sazonalidade no município.

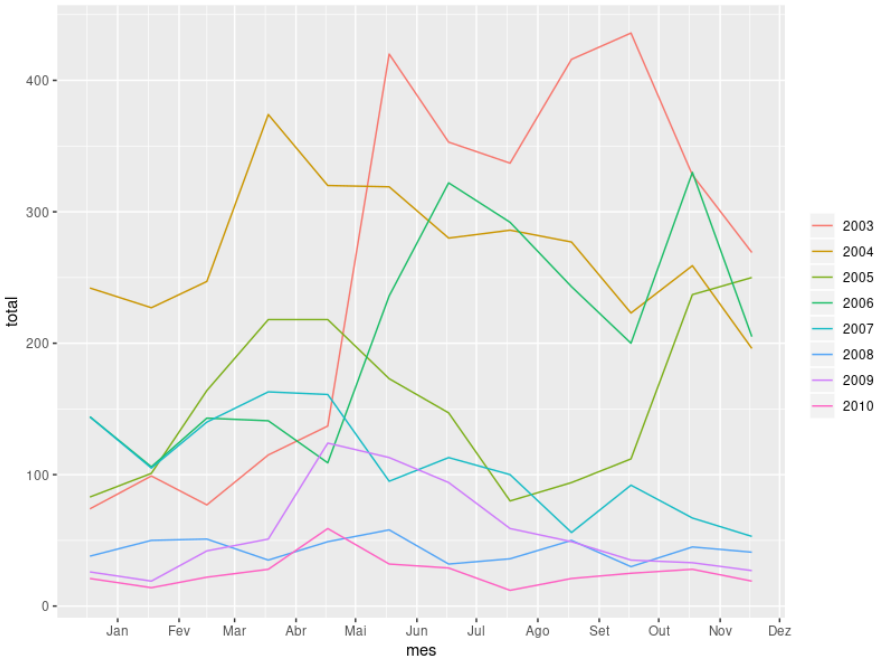


Figura 3.12 Gráfico sazonal para o município de Boca do Acre de 2003 a 2010.

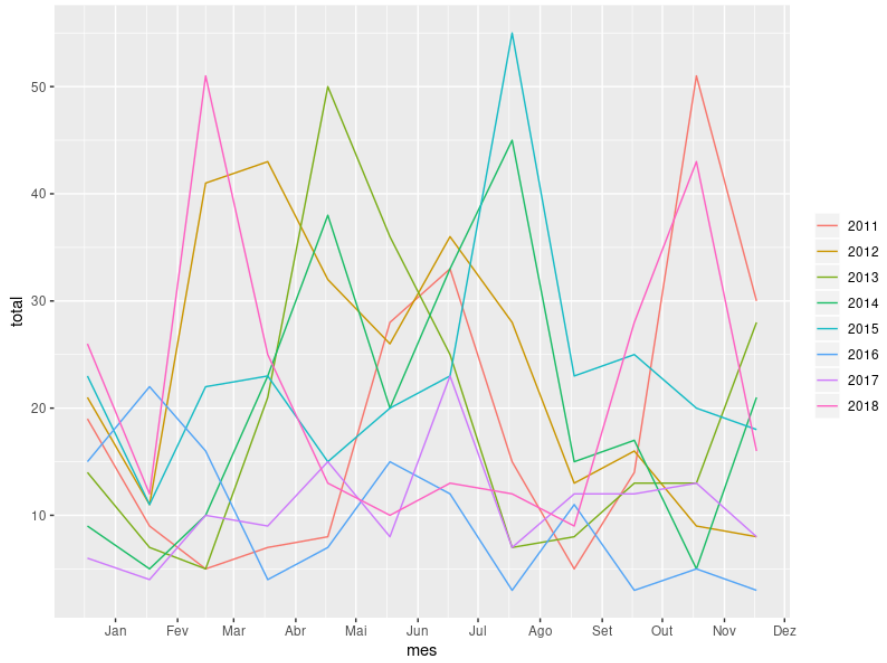


Figura 3.13 Gráfico sazonal para o município de Boca do Acre de 2011 a 2018.

3.4.4 Análise Descritiva - Humaitá

As figuras 3.14 e 3.15 são, respectivamente, para os períodos de 2003 a 2010 e 2011 a 2018 para o município de Humaitá. É possível perceber visualmente um tendência crescente do número de casos entre 2003 a 2010, e a redução a partir de 2008. Para o período de 2011 a 2018 a tendência é decrescente, com picos em 2014 e 2017. Em *season year* é possível perceber uma mudança do padrão sazonal quando comparadas as duas figuras.

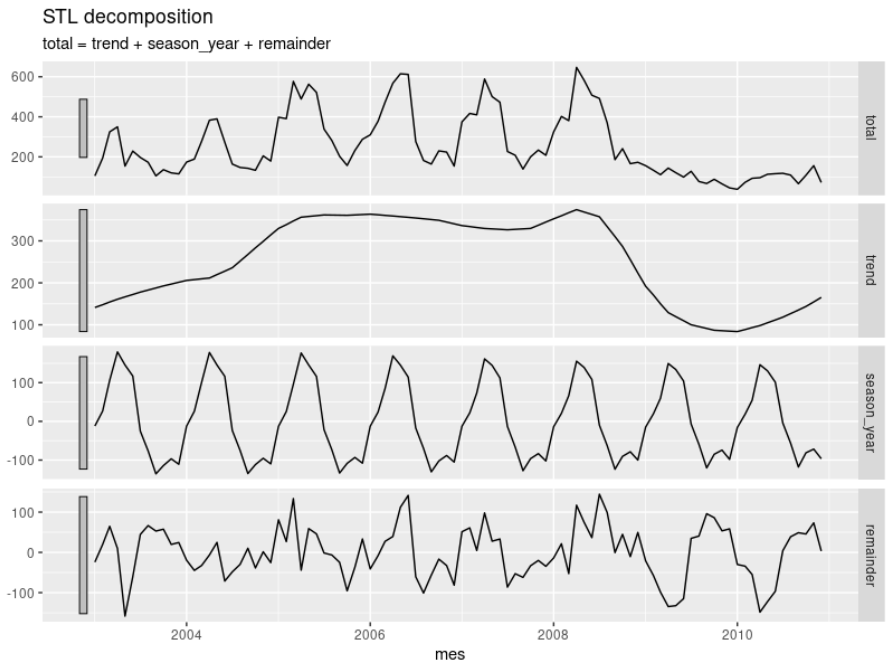


Figura 3.14 Decomposição da série temporal para o município de Humaitá de 2003 a 2010.

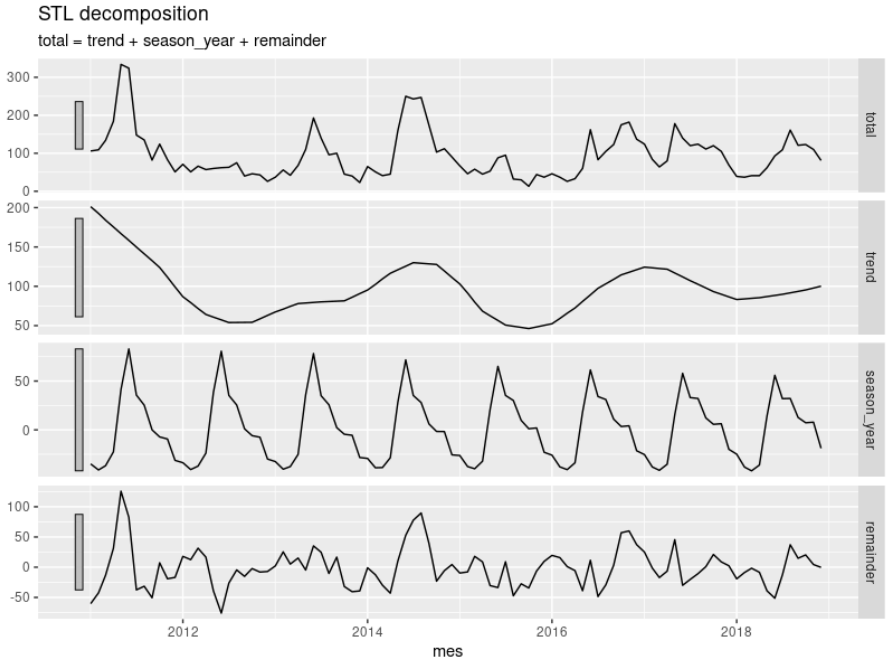


Figura 3.15 Decomposição da série temporal para o município de Humaitá de 2011 a 2018.

Nas figuras 3.16 e 3.17, respectivamente para os períodos 2003 a 2010 e 2011 a 2018, também não há um padrão visual que permita a fácil identificação da sazonalidade no município.

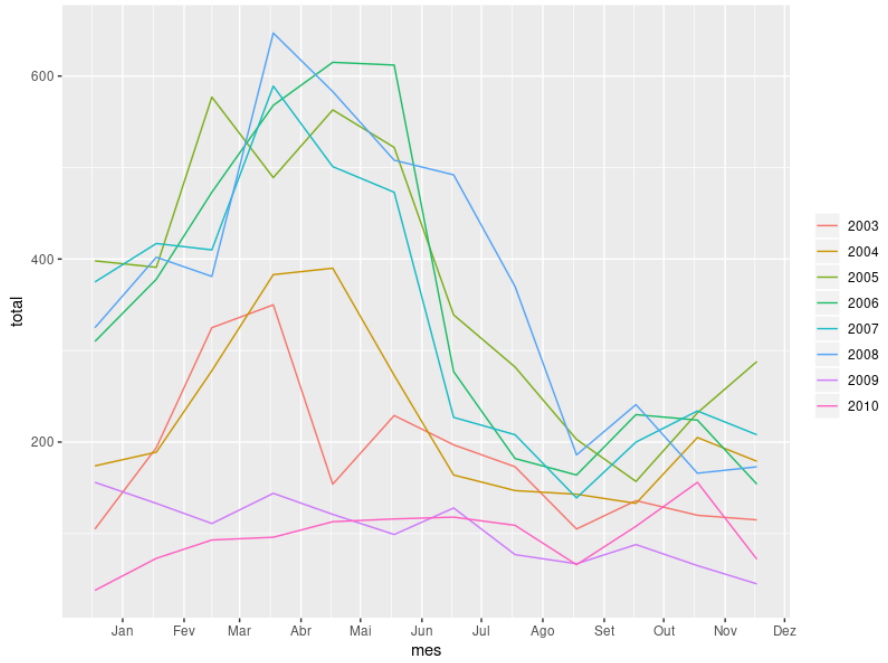


Figura 3.16 Gráfico sazonal para o município de Humaitá de 2003 a 2010.

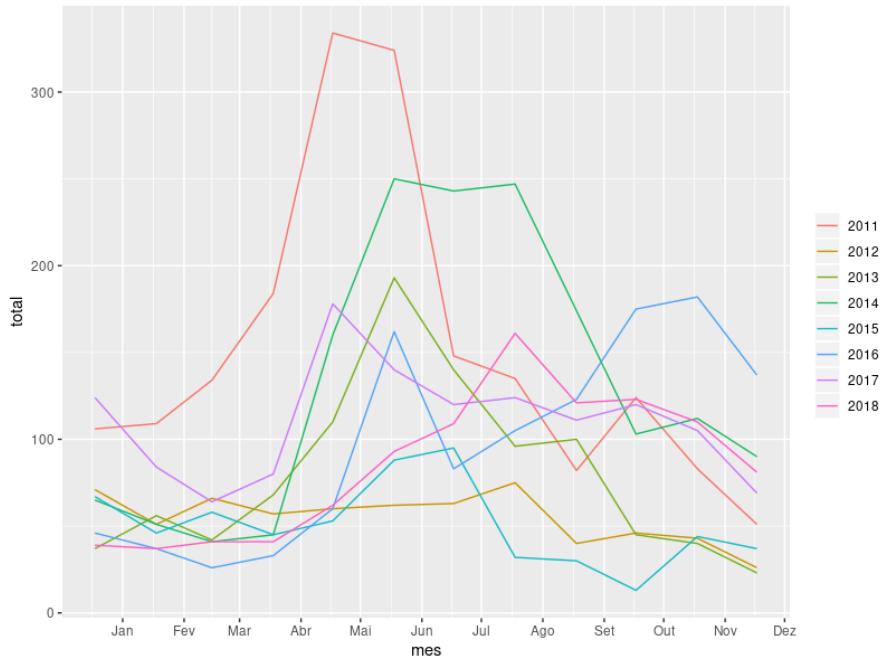


Figura 3.17 Gráfico sazonal para o município de Humaitá de 2011 a 2018.

3.4.5 Métodos utilizados

Para desenvolver os modelos preditivos foram utilizados os seguintes algoritmos: K-Nearest Neighbours (KNN), Máquina de vetores de suporte para regressão (SRV), Floresta Aleatória (RF) e Modelo Linear Generalizado (GLMNET).

3.4.5.1 ***K-Nearest Neighbours*** é um método baseado em distância, o qual classifica um novo ponto a partir de exemplos do conjunto de dados que estão próximos a ele. É um algoritmo preguiçoso, pois não aprende um modelo compacto, já que apenas armazena os objetos. Funciona tanto para problemas de classificação, como para regressão, e pode ser aplicado até mesmo em problemas complexos. Por ser um algoritmo preguiçoso, a fase de treinamento requer pouco esforço computacional, entretanto, classificar um objeto de teste, requer calcular a distância desse novo ponto em relação a todos os outros objetos de treinamento. Esse método é afetado pela presença de anomalias (*outliers*) e dados redundantes, como todo os algoritmos baseados em distância; além disso, tem problemas relacionados a dimensionalidade dos dados, pois dois dados próximos podem não estar próximos em casos de alta dimensionalidade. (FACELI et al., 2011a)

3.4.5.2 ***Random Forest*** é baseado em árvores de decisão. Trabalhar bem com alta dimensionalidade e multicolinearidade. Também é utilizado para estudo da importância e seleção de atributos e detecção de *outliers*. Esse tipo de modelo combina o resultado de um conjunto de árvores de decisão para obter uma resposta única, que tende a apresentar resultados melhores que cada árvore separadamente. Esse algoritmo é utilizado por (BUCZAK et al., 2015) para predição de malária.

3.4.5.3 ***Elastic-Net Regularized Generalized Linear Models*** é um algoritmo que usa regressão linear utilizando regularização através de *lasso* ou *elasticnet*. É um algoritmo considerado rápido, faz otimização de parâmetros a cada iteração por meio da descida cíclica de coordenadas.

3.4.5.4 ***Support Vector Regression*** é um método criado a partir do *Support Vector Machine*(SVM), método originalmente desenvolvido para classificação, sendo posteriormente estendido para problemas de regressão, denominado *Support Vector Regression*(SVR). Tem como base o aprendizado estatístico e tem como parâmetros C e o tipo de função kernel. O algoritmo busca encontrar um preditor que aproxime bem os dados de amostra. Tem excelente generalizações e alto poder de acerto, além de não depender da dimensionalidade do espaço de entrada do dados (AWAD; KHANNA, 2015; FACELI et al., 2011a).

Capítulo

4

EXPERIMENTAÇÃO E AVALIAÇÃO

Neste capítulo serão apresentados os experimentos realizados juntamente com a engenharia de atributos. Os atributos escolhidos para os experimentos foram criados a partir da base de dados BNNM e de clima. Uma nova base de dados foi entrão criada, para cada município, sendo que cada observação se refere a um mês. O período foi de 01/2003 a 05/2018. A variável meta (ou y) é sempre o total de casos. Os atributos de cada experimento estão representados a seguir:

1. **Experimento um:** Número de indivíduos assintomáticos, número de indivíduos do sexo masculino, número de indivíduos do sexo feminino, número de gestantes e o número de casos notificados positivos de malária.

2. **Experimento dois:** Todos os atributos da etapa um mais variáveis climáticas: temperatura, precipitação, umidade, potencial de evaporação e escorrência superficial.

Para avaliar um algoritmo de AM supervisionado, frequentemente é utilizado a análise de desempenho do preditor gerado pelo mesmo na rotulação de novos objetos que não foram utilizados durante a etapa de treinamento. Para problemas de regressão, o erro pode ser calculado utilizando a distância entre o valor predito e o valor conhecido. Neste trabalho é utilizada a métrica RMSE, definida como $\sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2}$; onde: y_i é o valor real, x_i é o valor predito e n é o total de observações.

Para cada um dos algoritmos utilizados, foram testados diversos parâmetros de entrada. A busca por parâmetros utilizou $N = 100$ parâmetros diferentes para cada algoritmo, com o objetivo de que, ao final, cada um seja executado com os parâmetros que melhor se adaptem aos dados RMSE.

Os experimentos foram executados de um até três meses a frente. Para cada um dos horizontes, foram utilizadas variáveis defasadas e a abordagem de predição multi-passos direta.

4.1 EXPERIMENTOS

4.1.1 Município de Boca do Acre

Predição para um mês a frente - com janela de expansão		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	24,09	38,44
GLMNET	25,39	29,25
Random Forest	23,50	27,05
SVR	**20,41**	24,52

Tabela 4.1 RMSE para Boca do Acre: janela de expansão e horizonte de um mês.

Predição para um mês a frente - com janela deslizantes		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	24,49	28,18
GLMNET	22,29	26,10
Random Forest	22,45	25,82
SVR	**21,70**	28,43

Tabela 4.2 RMSE para Boca do Acre: janela deslizante e horizonte de um mês.

Para Boca do Acre, observando as tabelas 4.1 e 4.2, pode-se avaliar que para o horizonte de predição de um mês a frente a melhor abordagem foi com janela de expansão. Dentro desta abordagem, o melhor algoritmo foi o SVR com RMSE de 20,41, utilizando os atributos do experimento um. Pode-se observar que o uso dos dados climáticos não melhorou o modelo em nenhum algoritmo nem nas diferentes abordagens de janela deslizante e de expansão, conforme a Figura 4.1.

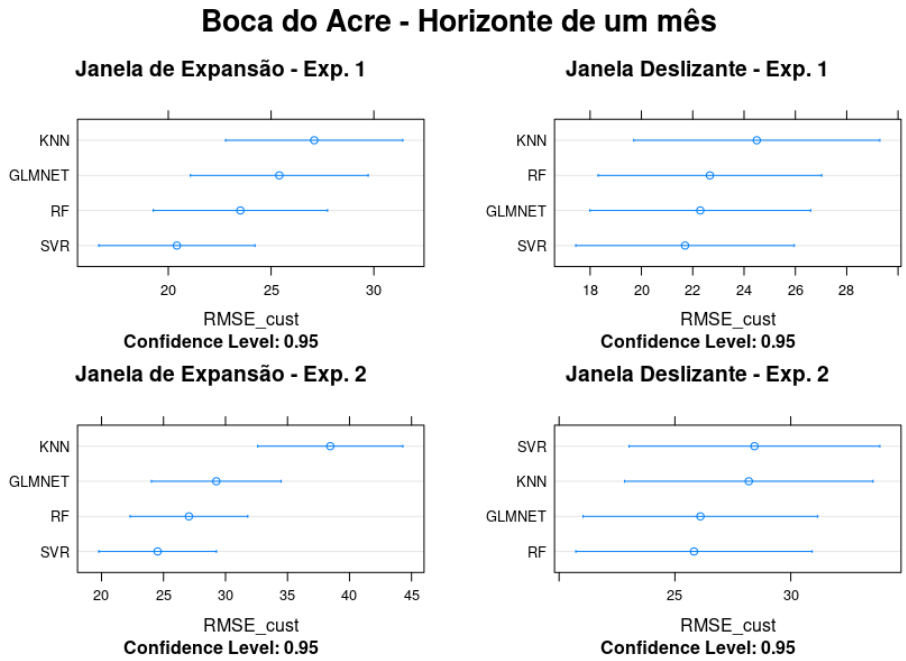


Figura 4.1 Comparação entre os modelos para horizonte de um mês - Boca do Acre.

Predição para dois meses a frente - com janela de expansão		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	26,28	46,27
GLMNET	26,88	44,32
Random Forest	**26,07**	37.02
SVR	27,23	32,88

Tabela 4.3 RMSE para Boca do Acre: janela de expansão para o horizonte de dois meses.

Predição para dois meses a frente - com janela deslizante		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	26,10	28,13
GLMNET	27,06	28,72
Random Forest	**25,70**	28,24
SVR	26,52	31,48

Tabela 4.4 RMSE Boca do Acre: janela deslizante para o horizonte de dois meses.

Observando as tabelas 4.3 e 4.4, a predição com horizonte de dois meses a frente teve como melhor abordagem o uso da janelas deslizante, utilizando o algoritmo *Randon Forest* com RMSE de 26,52. Não houve melhora nos modelos com o uso das variáveis do experimento dois, como é mostrado na Figura 4.2.

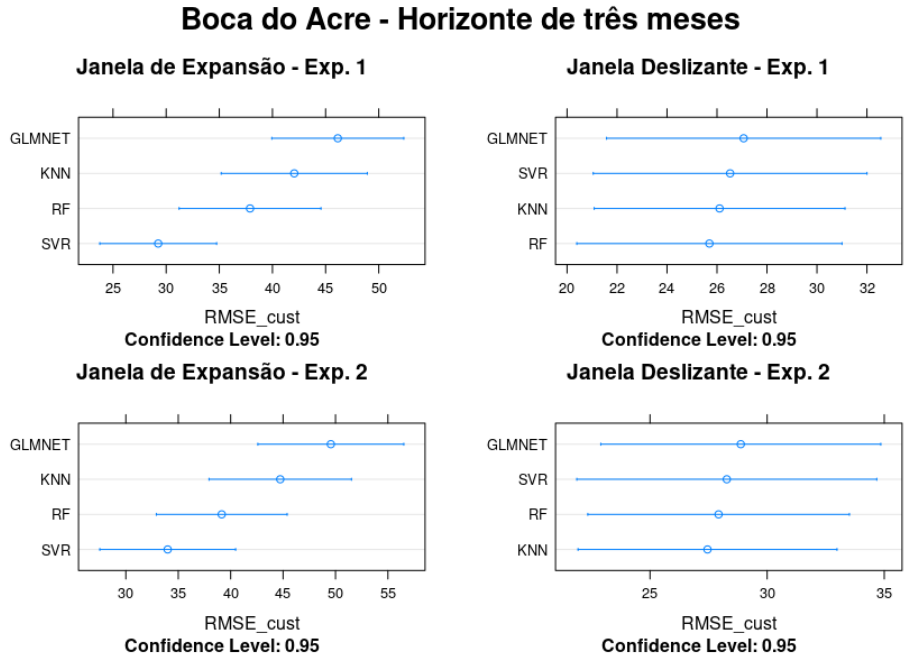


Figura 4.2 Comparação entre os modelos para horizonte de dois meses - Boca do Acre.

Predição para três meses a frente - com janela de expansão		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	42,04	44,72
GLMNET	46,14	49,55
Random Forest	37,89	39,14
SVR	**29,25**	33,99

Tabela 4.5 RMSE para Boca do Acre: janela de expansão para o horizonte de três meses.

Predição para três meses a frente - com janela deslizante		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	26,10	27,45
GLMNET	27,06	28,87
Random Forest	**25,69**	27,92
SVR	26,52	28,27

Tabela 4.6 RMSE para Boca do Acre: janela deslizante para o horizonte de três meses.

Conforme as tabelas 4.5 e 4.6, a predição com horizonte de três meses a frente teve como melhor abordagem o uso da janela deslizante, utilizando o algoritmo *Randon Forest* com RMSE de 26,69. Não houve melhora nos modelos com o uso das variáveis do experimento dois. A comparação dos resultados apresenta-se na Figura 4.3.

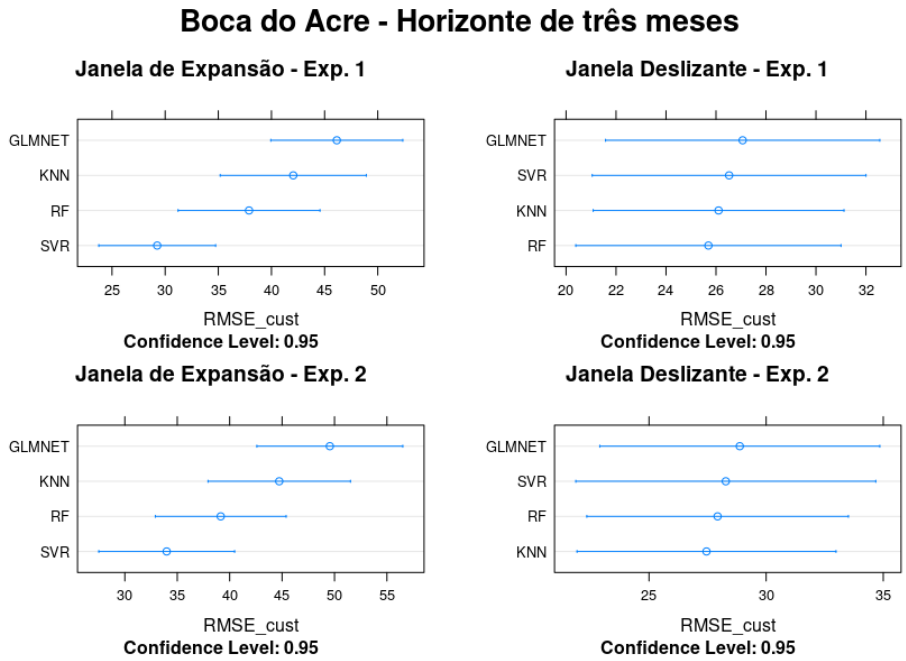


Figura 4.3 Comparação entre os modelos para horizonte de três meses - Boca do Acre.

4.1.2 Município de Manaus

Predição para um mês a frente - com janela de expansão		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	453,52	603,80
GLMNET	458,64	544,96
Random Forest	435,33	480,20
SVR	**389,13**	428,68

Tabela 4.7 RMSE para Manaus: janela de expansão para o horizonte de um mês.

Predição para um mês a frente - com janela deslizante		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	437,20	469,70
GLMNET	**391.84**	439,61
Random Forest	408,99	444,20
SVR	395,80	462,16

Tabela 4.8 RMSE para Manaus: janela deslizante para o horizonte de um mês.

A melhor abordagem para Manaus, conforme as tabelas 4.7 e 4.8, foi utilizando janelas de expansão e o melhor algoritmo foi o SVR com RMSE de 389,13. Pode-se observar

que não houve melhora com o uso das variáveis do experimento dois e também com o uso de janelas deslizante. A Figura 4.4 apresenta a comparação entre os modelos para essa abordagem.

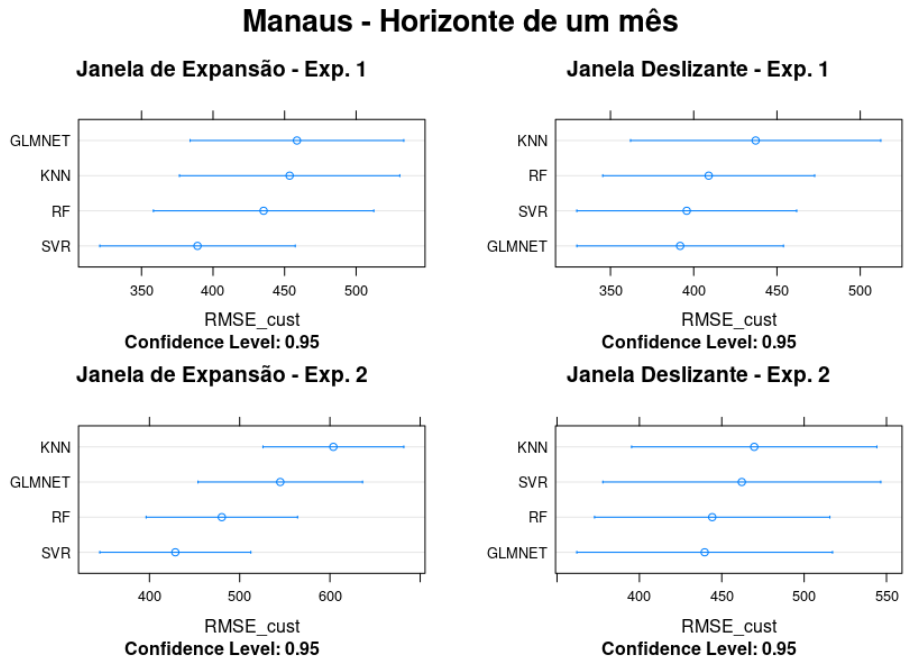


Figura 4.4 Comparação entre os modelos para horizonte de um mês - Manaus

Predição para dois meses a frente - com janela de expansão		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	588,80	671,52
GLMNET	654,97	689,28
Random Forest	588.74	571,79
SVR	**500,20**	555,80

Tabela 4.9 RMSE para Manaus: janelas de expansão para o horizonte de dois meses.

Predição para dois meses a frente - com janela deslizante		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	**487,32**	522,11
GLMNET	493,57	515,19
Random Forest	508,64	506,58
SVR	488,33	522,98

Tabela 4.10 RMSE para Manaus: janela deslizante para o horizonte de dois meses.

Para a predição com o horizonte de dois meses a frente para Manaus, conforme as tabelas 4.10 e 4.9, observa-se que a melhor abordagem foi a abordagem de janelas deslizante com o algoritmo KNN. A comparação entre os modelos apresenta-se na Figura 4.5.

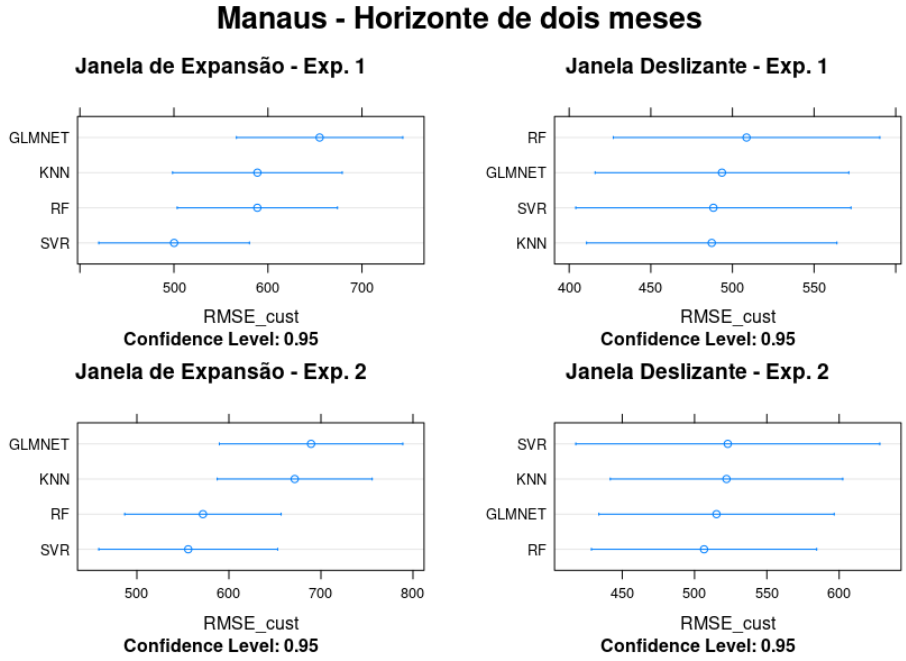


Figura 4.5 Comparação entre os modelos para horizonte de dois meses - Manaus.

Predição para três meses a frente - com janela de expansão		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	646,32	730,43
GLMNET	714,88	762,26
Random Forest	638,76	599,30
SVR	**557,24**	609,38

Tabela 4.11 RMSE para Manaus: janela de expansão para o horizonte de dois meses.

Predição para três meses a frente - com janela deslizante		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	**498,25**	521,80
GLMNET	500,08	521,75
Random Forest	521,47	503,51
SVR	503,72	564,39

Tabela 4.12 RMSE para Manaus: janela deslizante para o horizonte de três meses.

Já para a predição com horizonte de três meses a frente, a melhor abordagem foi com janela deslizante e o algoritmo KNN. Conforme a Figura 4.6 e as Tabelas 4.11 e 4.12.

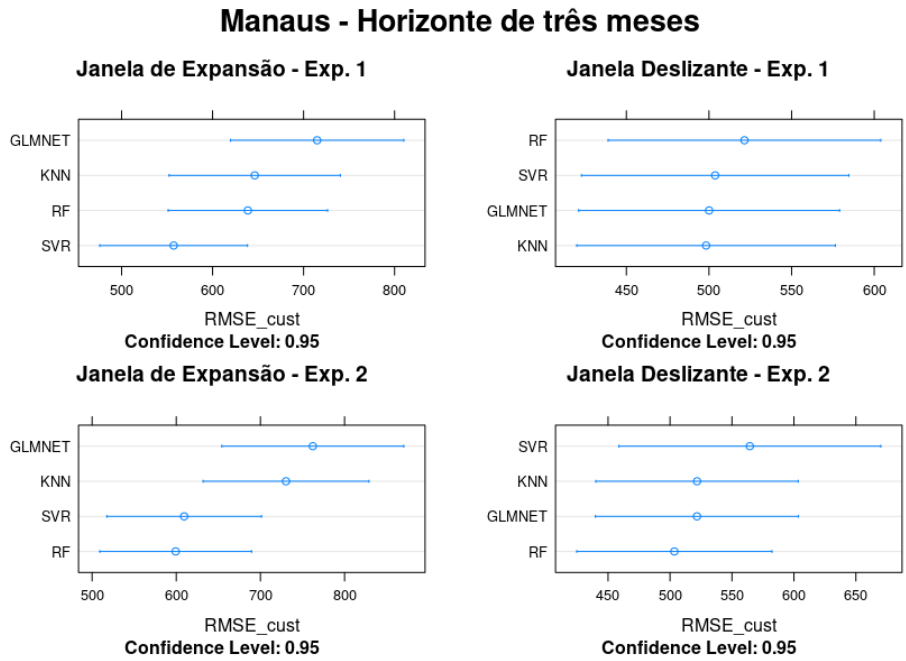


Figura 4.6 Comparação entre os modelos para horizonte de três meses - Manaus.

4.1.3 Município de São Gabriel da Cachoeira

Predição para um mês a frente - com janela de expansão		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	143,47	146,86
GLMNET	125,31	126,21
Random Forest	**123,25**	126,30
SVR	126,33	132,36

Tabela 4.13 RMSE para São Gabriel da Cachoeira: janela de expansão para o horizonte de um mês.

Predição para um mês a frente - com janela deslizante		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	161,69	163,38
GLMNET	145,82	151,64
Random Forest	**139,98**	152,11
SVR	152,66	188,73

Tabela 4.14 RMSE para São Gabriel da Cachoeira: janela deslizante para o horizonte de um mês.

Para a predição com o horizonte de um mês a frente para São Gabriel da Cachoeira, conforme as tabelas 4.13 e 4.14, observa-se que a melhor abordagem foi com janelas de expansão e com o algoritmo *Random Forest*. Não houve melhora utilizando os atributos do experimento dois. A comparação entre os modelos apresenta-se na Figura 4.7.

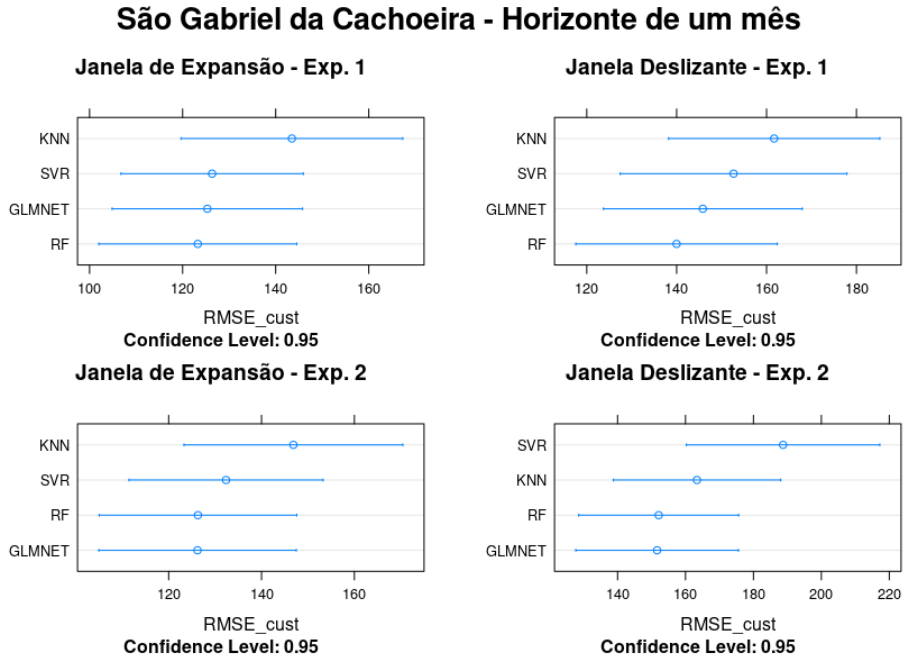


Figura 4.7 Comparação entre os modelos para horizonte de um mês - São Gabriel da Cachoeira.

Predição para dois meses a frente - com janela de expansão		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	180,2	174,25
GLMNET	169,29	169,04
Random Forest	171,47	**166,38**
SVR	170,17	171,53

Tabela 4.15 RMSE para São Gabriel da Cachoeira: janela de expansão para o horizonte de dois meses.

Predição para dois meses a frente - com janela deslizante		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	183,15	181,28
GLMNET	180,89	171,90
Random Forest	**170,68**	173,27
SVR	189,15	200,60

Tabela 4.16 RMSE para São Gabriel da Cachoeira: janela deslizante para o horizonte de dois meses.

Para a predição com o horizonte de dois meses a frente para São Gabriel da Cachoeira, conforme as tabelas 4.15 e 4.16, observa-se que a melhor abordagem foi com janela de expansão e com o algoritmo *Random Forest*. Houve melhora do modelo utilizando dados climáticos, como mostrado na Figura 4.8

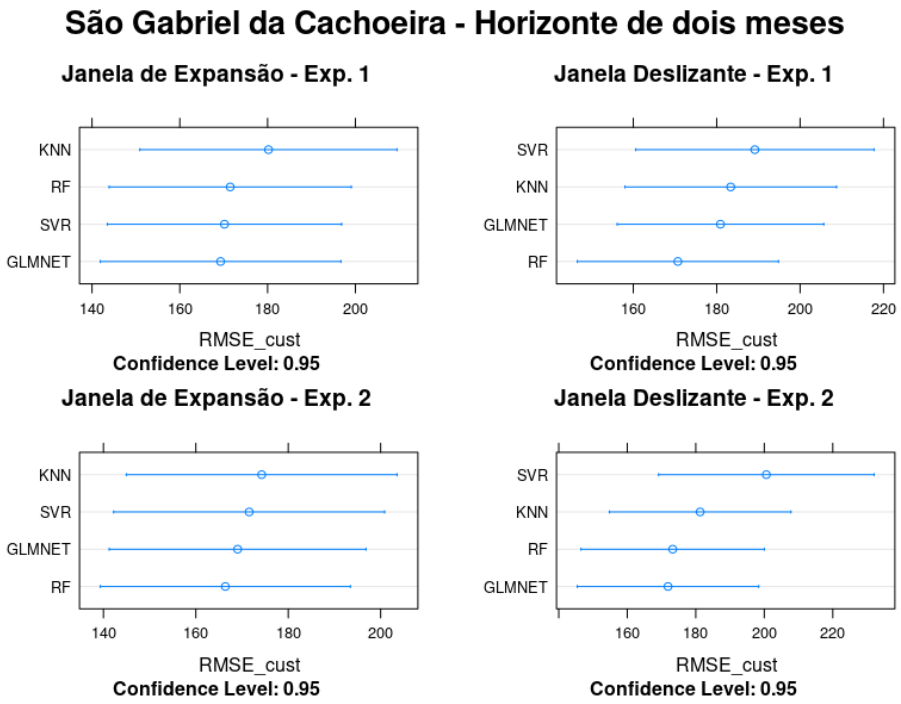


Figura 4.8 Comparação entre os modelos para horizonte de dois meses - São Gabriel da Cachoeira.

Predição para três meses a frente - com janela de expansão		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	201,16	196.73
GLMNET	199,01	194.83
Random Forest	191,28	**181.81**
SVR	200,88	197.45

Tabela 4.17 RMSE para São Gabriel da Cachoeira: janela de expansão para o horizonte de três meses.

Predição para três meses a frente - com janela deslizante		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	193,71	186,69
GLMNET	192,42	185,25
Random Forest	186,21	**176,15**
SRV	224,03	219,14

Tabela 4.18 RMSE para São Gabriel da Cachoeira: janela deslizante para o horizonte de três meses.

Para a predição com o horizonte de três meses a frente para São Gabriel da Cachoeira, conforme as tabelas 4.17 e 4.17, observa-se que a melhor abordagem foi com janelas

deslizante e com o algoritmo *Random Forest*. Houve melhora do modelo utilizando dados climáticos conforme a Figura 4.9

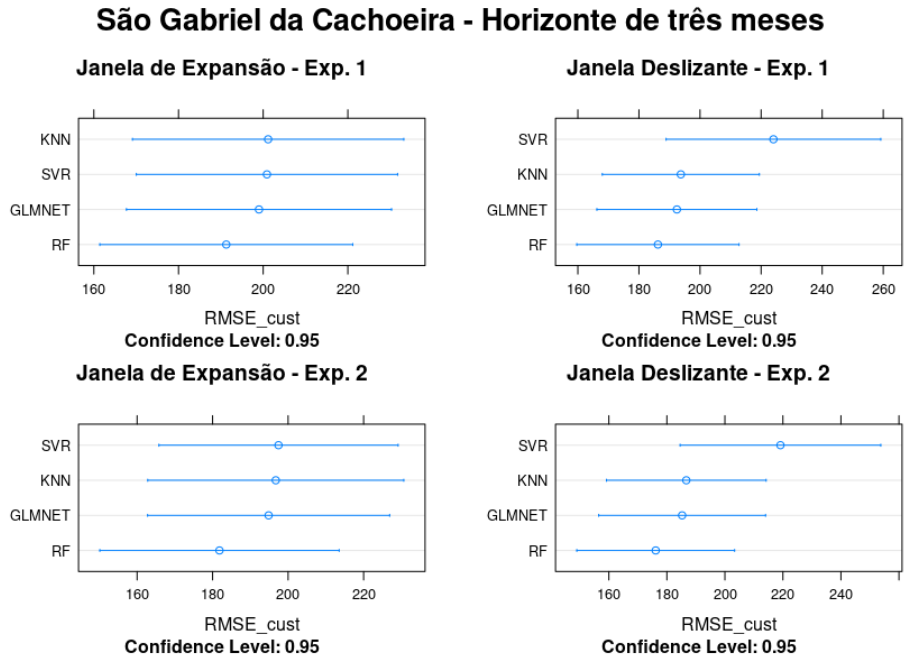


Figura 4.9 Comparação entre os modelos para horizonte de três meses - São Gabriel da Cachoeira.

4.1.4 Município de Humaitá

Predição para um mês a frente - com janela de expansão		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	55,70	63,04
GLMNET	53,23	55,25
Random Forest	53,63	53,97
SVR	**51,01**	54,65

Tabela 4.19 RMSE para Humaitá: janelas de expansão para o horizonte de um mês.

Predição para um mês a frente - com janela deslizante		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	55,39	**53,12**
GLMNET	56,24	56,50
Random Forest	54,42	55,21
SVR	54,98	62,53

Tabela 4.20 RMSE para Humaitá: janela deslizante para o horizonte de um mês.

Para a predição com o horizonte de um mês a frente para Humaitá, conforme as tabelas 4.19 e 4.20, observa-se que a melhor abordagem foi com janela deslizante e com o algoritmo SVR com dados do experimento um, conforme a Figura 4.10.

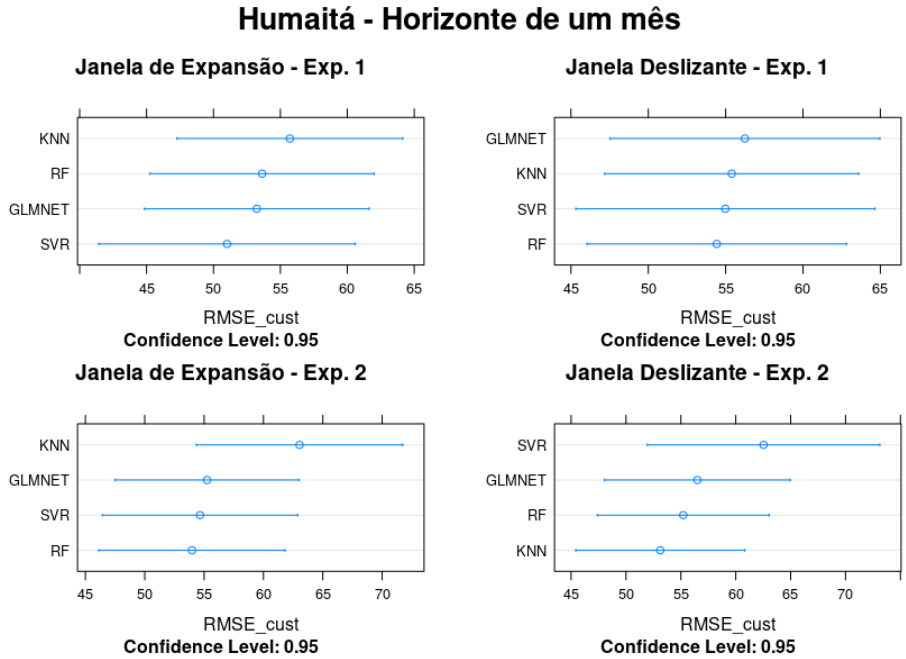


Figura 4.10 Comparação entre os modelos para horizonte de um mês - Humaitá.

Predição para dois meses a frente - com janela de expansão		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	78,85	78,10
GLMNET	80,85	74,89
Random Forest	80,53	**66,03**
SVR	74,10	75,65

Tabela 4.21 RMSE para Humaitá: janela de expansão para o horizonte de dois meses.

Predição para dois meses a frente - com janela deslizante		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	72,14	63,38
GLMNET	75,82	65,26
Random Forest	73,27	**58,66**
SVR	72,78	73,39

Tabela 4.22 RMSE para Humaitá: janela deslizante para o horizonte de dois meses.

Para a predição com o horizonte de dois meses a frente para Humaitá, conforme as tabelas 4.21 e 4.22, observa-se que a melhor abordagem foi com janela deslizante e com o algoritmo *Random Forest* com dados do experimento dois, conforme a Figura 4.11.

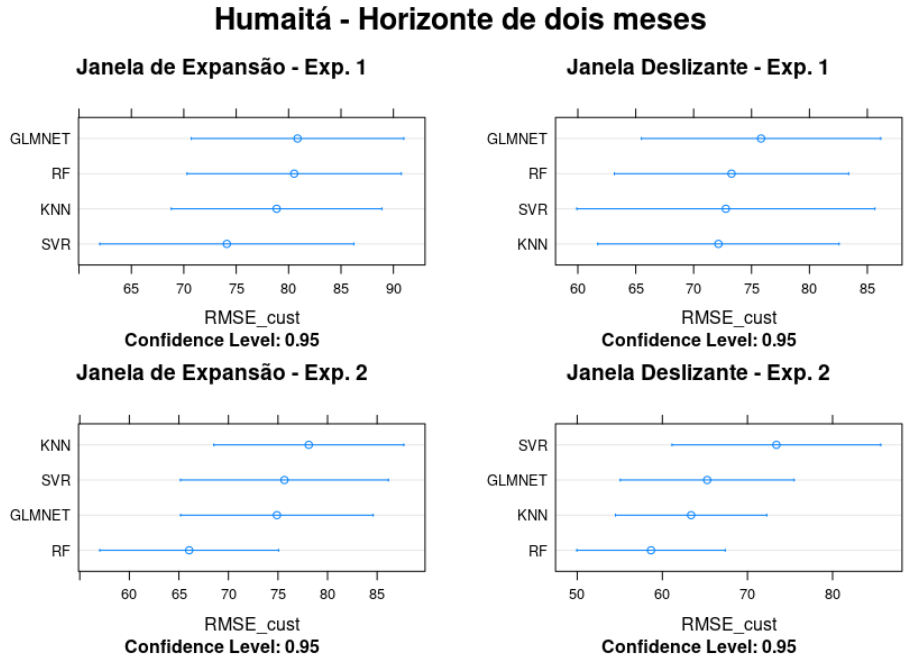


Figura 4.11 Comparação entre os modelos para horizonte de dois meses - Humaitá.

Predição para três meses a frente - com janela de expansão		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	93,47	86,86
GLMNET	93,79	85,21
Random Forest	94,89	**80,34**
SVR	84,85	84,31

Tabela 4.23 RMSE para Humaitá: janelas de expansão para o horizonte de três meses.

Predição para três meses a frente - com janela deslizante		
Modelo	RMSE Exp. Um	RMSE Exp. Dois
KNN	80,33	69,87
GLMNET	81,73	**60,98**
Random Forest	78,93	63,67
SVR	80,24	64,69

Tabela 4.24 RMSE para Humaitá: janela deslizante para o horizonte de três meses.

Para a predição com o horizonte de três meses a frente para Humaitá, conforme as tabelas 4.23 e 4.24, observa-se que a melhor abordagem foi com janela deslizante e com o algoritmo GLMNET com dados do experimento dois. A comparação também apresenta-se na Figura 4.12

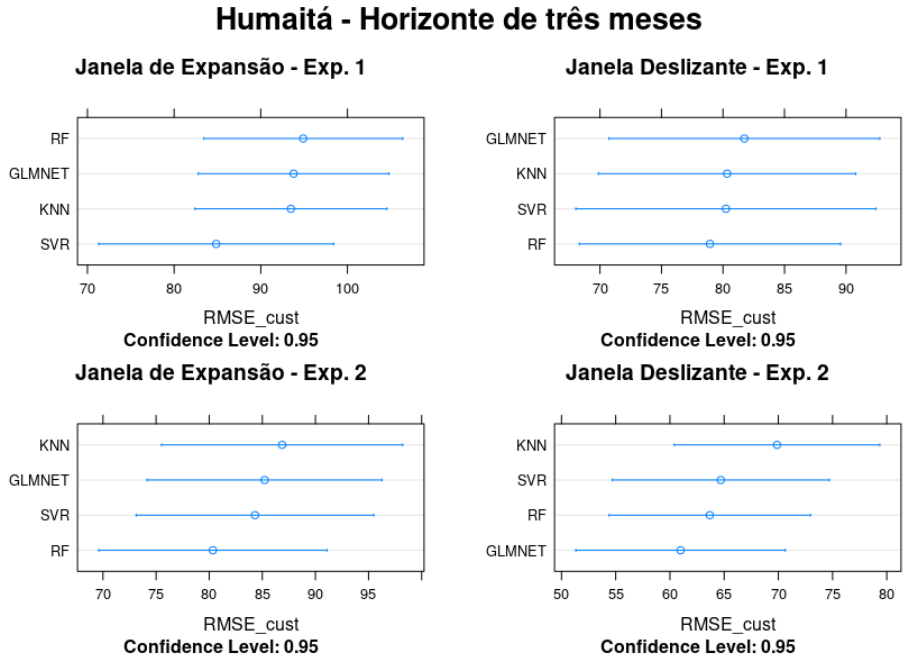


Figura 4.12 Comparação entre os modelos para horizonte de três meses - Humaitá.

4.2 RESULTADOS

Para o município de Boca do Acre o melhor modelo foi sempre utilizando o experimento um, e o algoritmo que obteve os melhores resultados, em geral, foi o SVR, conforme as tabelass 4.1 a 4.6 e sumarizado na tabela 4.25.

Para Manaus os melhores modelos foram utilizando o experimento um, conforme as tabelas de 4.7 a 4.12. O melhor algoritmo no geral foi o SVR, que obteve em três experimentos o menor erro, conforme a tabela 4.25.

Para o município São Gabriel da Cachoeira, no experimento dois com horizonte de dois e três meses a frente, os dados climáticos melhoraram o modelo reduzindo o erro. Em todos os experimentos o algoritmo RF foi o melhor.

Para o município Humaitá, os atributos climáticos influenciaram de forma positiva, reduzindo o erro em experimentos com horizonte de predição de um, dois e três meses a frente. No total dos experimentos, o melhor algoritmo foi o SVR.

Município	**KNN**	**GLMNET**	**SVR**	**RF**
Boca do Acre	0	0	3	3
São Gabriel da Cachoeira	0	0	0	6
Humaitá	1	1	1	3
Manaus	2	1	3	0

Tabela 4.25 Sumário do desempenho dos modelos preditivos desenvolvidos.

Capítulo 5

CONCLUSÕES E TRABALHOS FUTUROS

5.1 DISCUSSÃO

A predição pode ser uma ferramenta importante no combate à malária, particularmente em conjunto com dados de clima e ecológicos. Diversos estudos neste âmbito utilizam tais dados. Entretanto alguns evidenciam que em certos casos, pode ocorrer insensibilidade a mudanças climáticas por fatores como, taxa de sobrevivência do vetor, taxa de recuperação dos indivíduos entre outros (HAY et al., 2001).

Nos resultados apresentados, Manaus e Boca do Acre não apresentaram melhora no modelo com o uso dos dados de clima, enquanto São Gabriel da Cachoeira e Humaitá houve em algum dos experimentos a redução do RMSE. Portanto, confirma-se a característica complexa da doença, com diferentes fatores que impactam na oscilação da mesma.

Dessa forma, é importante notar que a maioria dos estudos de predição de malária se concentram em poucos preditores, apesar de existirem outros potenciais como, dados do uso da terra, distribuição de mosquiteiros, pulverização residual interna e dados de controle de vetores. Portanto, a precisão da predição pode ser influenciada se os dados de intervenções não forem considerados nos modelos preditivos (ZINSZER et al., 2012).

Neste sentido, é importante evidenciar a necessidade da coleta de forma estruturada dos dados de controle de vetores. (SILVA et al., 2019a) evidenciam a importância do preenchimento dos dados no SIVEP-malária sobre o controle de vetores, pois além da importância da avaliação do efeito das ações, esses dados podem ser utilizados em modelos preditivos. O estudo considera também que o controle vetorial é componente essencial no combate a malária, especialmente no contexto para eliminar a transmissão da malária no Brasil.

5.2 CONCLUSÕES

O presente trabalho utilizou modelos de AM na análise preditiva do total de casos de malária com o intuito de auxiliar a gestão de recursos e esforços. Os modelos utilizam

dados históricos para estimar o total de casos de malária em diferentes horizontes e com diferentes abordagens e métodos.

Foram realizados dois experimentos, sendo cada um destes subdivididos em dois tipos de avaliação, com janelas deslizantes fixas e com janelas deslizantes não fixas. Avaliou-se os modelos para identificar aqueles com o menor erro médio.

Dessa forma as principais contribuições deste trabalho foram:

1. Integração dos dados de notificação de malária para o Brasil.
2. Identificação de potenciais algoritmos e modelos de predição utilizando AM.
3. Confirmação que os dados de climáticos são potencialmente relevantes para os modelos preditivos.

5.2.1 Trabalhos Futuros

Pretende-se investigar algoritmos de redes neurais como *multilayer perceptron*, *LSTM – long short-term memory* e outros algoritmos de aprendizagem profunda, de forma separada por municípios e também utilizando agrupamento de municípios similares. Pretende-se também utilizar técnicas de discretização no atributo *meta* para avaliar os modelos como classificação. Pretende-se avaliar os resultados obtidos com especialistas do domínio com o intuito de refinamento dos mesmos. Além disso, pretende-se desenvolver um algoritmo para automatizar a atualização da BNNM em conjunto com os dados de clima. Por fim, pretende-se integrar os modelos e os dados de notificações atualizados à ferramenta desenvolvida, no âmbito do projeto GCE, para a geração de alertas.

REFERÊNCIAS BIBLIOGRÁFICAS

ADIMI, F. et al. Towards malaria risk prediction in afghanistan using remote sensing. *Malaria Journal*, v. 9, n. 1, p. 125, May 2010. ISSN 1475-2875. Disponível em: <https://doi.org/10.1186/1475-2875-9-125>.

AMO, S. D. Técnicas de mineração de dados. 2004.

APACHE Spark™ - Lightning-Fast Cluster Computing. 2017. <https://spark.apache.org/>. (Accessed on 11/03/2017).

AWAD, M.; KHANNA, R. Support vector regression. In: _____. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Berkeley, CA: Apress, 2015. p. 67–80. ISBN 978-1-4302-5990-9. Disponível em: <https://doi.org/10.1007/978-1-4302-5990-9_4>.

BIG Data Reduction 3: From Descriptive to Prescriptive. 2013. <https://community.lithium.com/t5/Science-of-Social-Blog/Big-Data-Reduction-3-From-Descriptive-to-Prescriptive/ba-p/81556>. Accessed: 2017-11-01.

BONTEMPI, G.; TAIEB, S. B.; BORGNE, Y.-A. L. Machine learning strategies for time series forecasting. In: _____. *Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 62–77. ISBN 978-3-642-36318-4. Disponível em: <https://doi.org/10.1007/978-3-642-36318-4_3>.

BRAZ, R. M. et al. Avaliação da completude e da oportunidade das notificações de malária na Amazônia Brasileira, 2003-2012. *Epidemiologia e Serviços de Saúde*, scielo, v. 25, p. 21 – 32, 03 2016. ISSN 1679-4974. Disponível em: <http://scielo.iec.pa.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742016000100003&nrm=iso>.

BUCZAK, A. L. et al. Fuzzy association rule mining and classification for the prediction of malaria in South Korea. *BMC Medical Informatics and Decision Making*, BMC Medical Informatics and Decision Making, p. 1–17, 2015. ISSN 14726947.

CHAWLA, N. V.; DAVIS, D. A. Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine*, v. 28, n. 3, p. 660–665, Sep 2013. ISSN 1525-1497. Disponível em: <https://doi.org/10.1007/s11606-013-2455-8>.

CUNHA, G. B. da et al. Use of an artificial neural network to predict the incidence of malaria in the city of Cantá, state of Roraima. *Revista da Sociedade Brasileira de Medicina Tropical*, v. 43, n. 5, p. 567–70, 2010. ISSN 1678-9849. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/21085871>.

DELEN, D.; DEMIRKAN, H. Data, information and analytics as services. *Decision Support Systems*, Elsevier B.V., v. 55, n. 1, p. 359–363, 2013. ISSN 01679236. Disponível em: <http://dx.doi.org/10.1016/j.dss.2012.05.044>.

FACELI, K. et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, v. 2, p. 192, 2011.

FACELI, K. et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. [S.l.]: LTC, 2011.

FAYYAD et al. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, v. 17, n. 3, p. 37, 1996. ISSN 0738-4602.

FINLAY, S. *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods*. Palgrave Macmillan UK, 2014. (Business in the Digital Economy). ISBN 9781137379283. Disponível em: <https://books.google.com.br/books?id=_em2AwAAQBAJ>.

GARTNER IT Glossary - Analytics. <https://www.gartner.com/it-glossary/analytics/>. Accessed: 2017-11-09.

GIROND, F. et al. Analysing trends and forecasting malaria epidemics in Madagascar using a sentinel surveillance network: a web-based application. *Malaria Journal*, BioMed Central, v. 16, n. 1, p. 1–11, 2017. ISSN 14752875.

GOMEZ-ELIPE, A. et al. Forecasting malaria incidence based on monthly case reports and environmental factors in karuzi, burundi, 1997–2003. *Malaria Journal*, v. 6, n. 1, p. 129, Sep 2007. ISSN 1475-2875. Disponível em: <https://doi.org/10.1186/1475-2875-6-129>.

HAY, S. I. et al. Malaria early warning in kenya. *Trends in Parasitology*, v. 17, n. 2, p. 95 – 99, 2001. ISSN 1471-4922. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1471492200017633>.

HEALTHCARE Big Data Analytics: From Description to Prescription. 2015. <https://healthitanalytics.com/news/healthcare-big-data-analytics-from-description-to-prescription>. Accessed: 2017-11-01.

HYNDMAN, R.; ATHANASOPOULOS, G. *Forecasting: principles and practice*. OTexts, 2014. ISBN 9780987507105. Disponível em: <https://books.google.com.br/books?id=gDuRBAAAQBAJ>.

KALIPE, G.; GAUTHAM, V.; BEHERA, R. K. Predicting Malarial Outbreak using Machine Learning and Deep Learning Approach: A Review and Analysis. *Proceedings - 2018 International Conference on Information Technology, ICIT 2018*, IEEE, p. 33–38, 2018.

KLOSGEN, W.; ZYTKOW, J. The knowledge discovery process. *Handbook of data mining and knowledge discovery*, 2002.

KOTU, V.; DESHPANDE, B. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Amsterdam: Morgan Kaufmann, 2015. ISBN 978-0-12-801460-8. Disponível em: <http://www.sciencedirect.com/science/book/9780128014608>.

KUHN, M. *The caret Package*. 2009.

LICHMAN, M. *UCI Machine Learning Repository*. 2013. Disponível em: <http://archive.ics.uci.edu/ml>.

MAGLOGIANNIS, I. *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press, 2007. (Frontiers in artificial intelligence and applications). ISBN 9781586037802. Disponível em: <https://books.google.com.br/books?id=vLiTXDHr_sYC>.

MITCHELL, T. M. *Machine Learning*. [S.l.]: WCB McGraw-Hill, 1997.

MODU, B. et al. Towards a Predictive Analytics-Based Intelligent Malaria Outbreak Warning System. *Applied Sciences*, v. 7, n. 8, p. 836, 2017. ISSN 2076-3417. Disponível em: <http://www.mdpi.com/2076-3417/7/8/836>.

NEVES, D. *Parasitologia Humana*. [S.l.]: EDITORA ATHENEU, 2005. ISBN 9788538807155.

OLIVEIRA, S. R. M. Introdução à Aprendizagem de Máquina (Parte II) Resumo da Aula. n. Parte II, 2017.

PAHO. Malaria report situation: Brazil. n. 23, p. 0–3, 2014.

PINOCHET, L. H. C. Tendências de tecnologia de informação na gestão da saúde. *Mundo saúde*, v. 35, n. 4, p. 382–94, 2011.

SILVA, D. C. B. da et al. Current vector control challenges in the fight against malaria in brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, v. 52, p. e20180542, 2019.

SILVA, D. Clarys Baia-da et al. Current vector control challenges in the fight against malaria in brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, v. 52, 03 2019.

SOCIETY, T. R. *Machine learning : the power and promise of computers that learn by example*. [s.n.], 2017. 125 p. ISBN 9781782522591. Disponível em: <https://royalsociety.org/{~}/media/policy/projects/machine-learning/publications/machine-learning-report.p>.

WALJEE, A. K.; HIGGINS, P. D.; SINGAL, A. G. A primer on predictive models. *Clinical and translational gastroenterology*, Nature Publishing Group, v. 5, n. 1, p. e44, 2014.

WHO. WORLD MALARIA REPORT 2015 Summary. 2015.

ZINSZER, K. et al. A scoping review of malaria forecasting: past work and future directions. *BMJ Open*, British Medical Journal Publishing Group, v. 2, n. 6, 2012. ISSN 2044-6055. Disponível em: <https://bmjopen.bmj.com/content/2/6/e001992>.