

Otávio Carvalho

**Predição da cintilação ionosférica de amplitude
pela aplicação de métodos de aprendizagem de
máquina**

São Luís – MA

2021

Otávio Carvalho

Predição da cintilação ionosférica de amplitude pela aplicação de métodos de aprendizagem de máquina

Trabalho apresentado ao Programa de Pós-graduação em Engenharia Aeroespacial (PP-GAERO), como requisito parcial à obtenção do título de Mestre em Engenharia Aeroespacial pela Universidade Federal do Maranhão (UFMA)

Universidade Federal do Maranhão

Centro de Ciências Exatas e Tecnológicas – CCET

Programa de Pós-Graduação em Engenharia Aeroespacial

Orientador: Prof. Dr. Alex Oliveira Barradas Filho

São Luís – MA

2021

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Carvalho, Otávio.

Predição da cintilação ionosférica de amplitude pela
aplicação de métodos de aprendizagem de máquina / Otávio
Carvalho. - 2021.

89 p.

Orientador(a): Alex Oliveira Barradas Filho.

Dissertação (Mestrado) - Programa de Pós-graduação em
Engenharia Aeroespacial, Universidade Federal do Maranhão,
São Luís, 2021.

1. Aprendizagem de máquina. 2. Cintilação
ionosférica. 3. Comunicação e navegação via satélite. 4.
Modelos preditivos. I. Filho, Alex Oliveira Barradas. II.
Título.

Otávio Carvalho

Predição da cintilação ionosférica de amplitude pela aplicação de métodos de aprendizagem de máquina

Trabalho apresentado ao Programa de Pós-graduação em Engenharia Aeroespacial (PP-GAERO), como requisito parcial à obtenção do título de Mestre em Engenharia Aeroespacial pela Universidade Federal do Maranhão (UFMA)

Trabalho aprovado. São Luís – MA, 21 de dezembro de 2021:

Prof. Dr. Alex Oliveira Barradas Filho
Orientador

**Prof. Dr. André Carlos Ponce de Leon
Ferreira de Carvalho**

**Prof. Dr. Francisco das Chagas de
Souza**

Prof. Dr. João Viana da Fonseca Neto

São Luís – MA
2021

Agradecimentos

Agradeço primeiramente à minha família que sempre me apoiou nesse projeto. Agradeço ao meu orientador, professor Alex, que sempre me incentivou e sem o qual este trabalho seria muito mais difícil. Agradeço à Universidade Federal do Maranhão por todo o suporte, em especial à coordenação do Programa de Pós-graduação em Engenharia Aeroespacial. Agradeço aos meus amigos pelo apoio, em especial a Mardiani, Marlon e Weslly. Agradeço ao professor Ricardo Yvan pelas sugestões e pelo apoio. Agradeço aos professores Eurico Rodrigues de Paula e Marcelo Banik de Pádua pelo auxílio fornecido na etapa de coleta de dados do trabalho.

Os autores agradecem ao Programa EMBRACE/INPE do MCTI, pelo fornecimento de dados para este trabalho.

Reconhecemos o uso do serviço OMNIWeb da Space Physics Data Facility da NASA/GSFC e dados OMNI.

O autor agradece o uso de dados provenientes dos projetos CIGALA e CALIBRA. As estações de monitoramento da rede foram implantadas no contexto dos Projetos CIGALA e CALIBRA, ambos financiados pela Comissão Europeia (CE) no âmbito do FP7-GALILEO-2009-GSA e FP7-GALILEO-2011-GSA-1a, respectivamente, e Projeto FAPESP Número 06/04008-2.

Resumo

A sociedade atual está cada vez mais dependente de produtos e serviços baseados no uso de satélites, dentre os quais destacam-se os sistemas de comunicação e navegação via satélite, tecnologias afetadas diretamente pelo fenômeno da cintilação ionosférica, que pode comprometer ou até mesmo inviabilizar a utilização de tais sistemas. Nesse contexto, torna-se importante desenvolver ferramentas capazes de prever a ocorrência da cintilação ionosférica, no entanto, a modelagem deste fenômeno é bastante complexa devido à influência de uma série de outros aspectos. Portanto, o principal objetivo deste trabalho é desenvolver modelos preditivos de curto prazo, tanto de caráter quantitativo quanto qualitativo, acerca da cintilação ionosférica de amplitude. Para isso serão usadas técnicas de aprendizagem de máquina, considerando informações relativas à atividade geomagnética, à dependência temporal e geográfica do fenômeno, às atividades solar e interplanetária e ao estado da ionosfera, avaliando a influência de diferentes atributos no desempenho dos modelos obtidos. A metodologia usada baseou-se no emprego de técnicas de aprendizagem de máquina, tanto de regressão quanto de classificação, destacando os algoritmos *Random Forest*, regressão logística e regressão linear múltipla. Os principais resultados obtidos relacionam-se à predição da cintilação ionosférica com 30 minutos de antecedência, além dos aspectos relativos à análise dos atributos utilizados. Por fim, são indicados os direcionamentos futuros da pesquisa realizada.

Palavras-chaves: Comunicação e navegação via satélite. Cintilação ionosférica. Modelos preditivos. Aprendizagem de máquina.

Abstract

Today's society is increasingly dependent on products and services based on the use of satellites, among which satellite communication and navigation systems stand out, technologies directly affected by the phenomenon of ionospheric scintillation, which can compromise or even make unfeasible the use of such systems. In this context, it becomes important to develop tools capable of predicting the occurrence of ionospheric scintillation, however, the modeling of this phenomenon is quite complex due to the influence of several other aspects. Therefore, the main objective of this work is to develop short-term predictive models, both quantitative and qualitative, about amplitude ionospheric scintillation. For this, machine learning techniques will be used, considering information related to geomagnetic activity, temporal and geographic dependence of the phenomenon, solar and interplanetary activities and the state of the ionosphere, evaluating the influence of different attributes on the performance of the models obtained. The methodology used was based on the use of machine learning techniques, both regression and classification, highlighting the algorithms Random Forest, logistic regression, and multiple linear regression. The main results obtained are related to the prediction of ionospheric scintillation 30 minutes in advance, in addition to aspects related to the analysis of the attributes used. Finally, the future directions of the research carried out are indicated.

Keywords: Satellite communication and navigation. Ionospheric scintillation. Predictive models. Machine learning.

Lista de ilustrações

Figura 1 – Efeitos das camadas ionosfera e troposfera na propagação de ondas de rádio	15
Figura 2 – Organização da ionosfera	21
Figura 3 – Incidência da cintilação ionosférica globalmente	22
Figura 4 – Construção de um modelo de aprendizagem de máquina	26
Figura 5 – Relação entre os conjuntos de treino, validação e teste	27
Figura 6 – Validação cruzada <i>3-fold</i>	27
Figura 7 – Divisão dos dados em blocos	28
Figura 8 – Algoritmo Random Forest	31
Figura 9 – Matriz de confusão	33
Figura 10 – Construção da base de dados usada nos modelos de regressão	37
Figura 11 – Tratamento de dados da base GNSS	40
Figura 12 – Estratégia de otimização de hiperparâmetros	44
Figura 13 – Metodologia de seleção de atributos	45
Figura 14 – Distribuição de exemplos por classe	47
Figura 15 – Metodologia de avaliação dos modelos preditivos	48
Figura 16 – Resultados obtidos por meio do tratamento de dados da base GNSS	49
Figura 17 – Resultados obtidos por meio do tratamento de dados da base GNSS - ampliação	50
Figura 18 – Gráficos de Box e Whisker relacionados aos modelos RLM	56
Figura 19 – Gráficos de Box e Whisker relacionados aos modelos RF - Regressão	56
Figura 20 – Gráficos de Box e Whisker contrapondo os resultados dos algoritmos RLM e RF	57
Figura 21 – Resultados obtidos no conjunto de testes	60
Figura 22 – Qualidade de ajustes aos dados do conjunto de testes	61
Figura 23 – Box plots referentes aos resultados da Tabela 21	65
Figura 24 – Box plots referentes aos resultados da Tabela 22	66
Figura 25 – Matrizes de confusão dos modelos RLog	69
Figura 26 – Matrizes de confusão dos modelos RF	70
Figura 27 – Parâmetros do campo magnético terrestre.	85
Figura 28 – Exemplo de ionograma	86

Lista de tabelas

Tabela 1 – Síntese dos trabalhos citados	18
Tabela 2 – atributos relacionadas à atividade geomagnética	38
Tabela 3 – atributos relacionadas às atividades solar e interplanetária	38
Tabela 4 – Atributos relacionadas ao estado da ionosfera derivadas de ionossonda - Regressão	39
Tabela 5 – atributos relacionadas ao estado da ionosfera derivadas de receptor GPS - Regressão	39
Tabela 6 – atributos relacionadas ao estado da ionosfera derivadas de receptor GPS	41
Tabela 7 – Atributos derivados de receptor GNSS - Classificação	43
Tabela 8 – Hiperparâmetros otimizados para o modelo RF de regressão	45
Tabela 9 – Hiperparâmetros otimizados para o modelo RF de classificação	46
Tabela 10 – Rankings de importância de atributos para os modelos de regressão	51
Tabela 11 – Quantidade ótima de atributos para cada algoritmo	51
Tabela 12 – Atributos selecionados nos modelos RLM	52
Tabela 13 – Atributos selecionados nos modelos RF	53
Tabela 14 – Resultados obtidos por meio do algoritmo RLM	55
Tabela 15 – Resultados obtidos por meio do algoritmo RF - Regressão	55
Tabela 16 – p -valores referentes aos testes que comparam os desempenhos dos algo- ritmos RLM e RF	58
Tabela 17 – p -valores referentes aos testes que avaliam se há diferenças nos desem- penhos dos modelos RLM em relação aos diferentes métodos de seleção de atributos	58
Tabela 18 – p -valores referentes aos testes que avaliam se há diferenças nos desem- penhos dos modelos RF em relação aos diferentes métodos de seleção de atributos	58
Tabela 19 – Rankings de importância das atributos para a classificação	63
Tabela 20 – Quantidade ótima de atributos para os modelos de classificação	64
Tabela 21 – Resultados obtidos por meio do algoritmo RLog	64
Tabela 22 – Resultados obtidos por meio do algoritmo RF	64
Tabela 23 – p -valores referentes ao teste TH_1	67
Tabela 24 – p -valores referentes ao teste TH_2	67
Tabela 25 – p -valores referentes ao teste TH_3	68
Tabela 26 – p -valores referentes ao teste TH_4	68
Tabela 27 – p -valores referentes ao teste TH_5	69
Tabela 28 – Resultados obtidos por meio do algoritmo RLog no conjunto de testes	70
Tabela 29 – Resultados obtidos por meio do algoritmo RF no conjunto de testes	71

Tabela 30 – Coeficientes do modelo RLM RR_1	82
Tabela 31 – Coeficientes do modelo RLM RR_2	83
Tabela 32 – Coeficientes do modelo RLM RR_3	83
Tabela 33 – Coeficientes do modelo RLM Base	84
Tabela 34 – Descrição da base de dados MAG usada nos modelos de regressão . . .	86
Tabela 35 – Descrição da base de dados SW usada nos modelos de regressão . . .	86
Tabela 36 – Descrição da base de dados ION usada nos modelos de regressão . . .	87
Tabela 37 – Descrição da base de dados GPS usada nos modelos de regressão . . .	87
Tabela 38 – Descrição da base de dados MAG usada nos modelos de classificação .	88
Tabela 39 – Descrição da base de dados SW usada nos modelos de classificação .	88
Tabela 40 – Descrição da base de dados ION usada nos modelos de classificação .	88
Tabela 41 – Descrição da base de dados GNSS usada nos modelos de classificação .	89

Lista de abreviaturas e siglas

GNSS	Global Navigation Satellite System
AIE	Anomalia de Ionização Equatorial
AMAS	Anomalia Magnética do Atlântico Sul
ANOVA	Análise de Variância
RF	Random Forest
RLM	Regressão Linear Múltipla
RLog	Regressão Logística
MI	Informação Mútua
NSC	Classe <i>não cintilação</i>
SC	Classe <i>cintilação</i>
VN	Quantidade de verdadeiros negativos
VP	Quantidade de verdadeiros positivos
FP	Quantidade de falsos positivos
FN	Quantidade de falsos positivos
MAG	Base de dados referente à atividade geomagnética
ION	Base de dados referente ao estado da ionosfera, obtida por meio de ionossonda
SW	Base de dados referente à atividade solar e interplanetária
GPS	Base de dados referente ao estado da ionosfera, obtida por meio de receptor GNSS
ACC	Acurácia
PRE	Precisão
SEN	Sensitividade
F1	Medida F1
ESP	Especificidade

Lista de símbolos

$\mathbf{E}$	Vetor campo elétrico [statV/cm]
c	Velocidade da luz no vácuo [cm/s]
$\mathbf{B}$	Vetor densidade de fluxo magnético [G]
$\mathbf{H}$	Vetor campo magnético [Oe]
$\mathbf{D}$	Vetor densidade de fluxo elétrico [statV/cm]
$\mathbf{J}$	Vetor densidade de corrente elétrica [statA/cm ²]
ρ_e	Densidade líquida de carga elétrica [statC/cm ³]
μ_m	Permeabilidade magnética do meio [G/Oe]
ε	Constante dielétrica do meio
$\mathbf{r}$	Vetor posição
S_4	Índice de cintilação de amplitude
I	Intensidade do sinal GNSS [dBW]
X	Entradas dos modelos de aprendizagem de máquina
Y	Saídas dos modelos de aprendizagem de máquina
$\mathbb{R}$	Conjunto dos número reais
$\mathbb{Z}$	Conjunto dos número inteiros
ρ	Coefficiente de correlação de Pearson
σ_{XY}	Covariância entre as variáveis X e Y
σ_X	Desvio padrão da variável X
σ_Y	Desvio padrão da variável Y
C	Quantidade de classes presentes no conjunto de dados
N	Quantidade total de exemplos
n_k	Quantidade de exemplos pertencentes à k -ésima classe

x_i	i -ésimo atributo
μ_k^i	Valor médio do atributo i correspondente aos exemplos da k -ésima classe
μ_i	Valor médio do atributo i
$x_{j,i}^k$	Valor do atributo i do j -ésimo exemplo pertencente à classe k
β_j	j -ésimo coeficiente do modelo de regressão linear múltipla
α_i	i -ésimo coeficiente do modelo de regressão logística
z	Estatística de teste usada no teste de postos com sinal de Wilcoxon
t_D	Número do dia considerado
t_H	Número da hora considerada
S_{4T}	Valor total do índice S_4
S_{4N}	Correção do índice S_4 , devida ao ruído ambiente
w_k	Peso atribuído ao classificar incorretamente um exemplo da classe k
RR_i	Ranking de importância de atributos para os modelos de regressão
RC_i	Ranking de importância de atributos para os modelos de classificação

Sumário

1	INTRODUÇÃO	15
2	OBJETIVOS	20
2.1	Objetivo Geral	20
2.2	Objetivos Específicos	20
3	REFERENCIAL TEÓRICO	21
3.1	Cintilação ionosférica	21
3.2	Aprendizagem de máquina	25
3.2.1	Validação cruzada	26
3.2.2	Otimização de hiperparâmetros	28
3.2.3	Seleção de atributos	29
3.2.4	Algoritmos	31
3.2.4.1	Random Forest	31
3.2.4.2	Regressão Linear Múltipla	31
3.2.4.3	Regressão Logística	32
3.2.5	Avaliação e comparação de modelos de aprendizagem de máquina	32
4	METODOLOGIA	36
4.1	Preparação da base de dados	36
4.1.1	Base de dados usada para regressão	36
4.1.2	Base de dados usada para classificação	41
4.2	Desenvolvimento dos modelos preditivos	43
4.2.1	Modelos preditivos de regressão	45
4.2.2	Modelos preditivos de classificação	46
4.3	Avaliação dos modelos preditivos	48
5	RESULTADOS E DISCUSSÃO	49
5.1	Modelos de regressão	49
5.1.1	Processamento de dados da base GNSS	49
5.1.2	Seleção de atributos	51
5.1.3	Avaliação dos modelos preditivos	55
5.2	Modelos de classificação	63
5.2.1	Seleção de atributos	63
5.2.2	Avaliação dos modelos preditivos	64
6	CONCLUSÃO	72

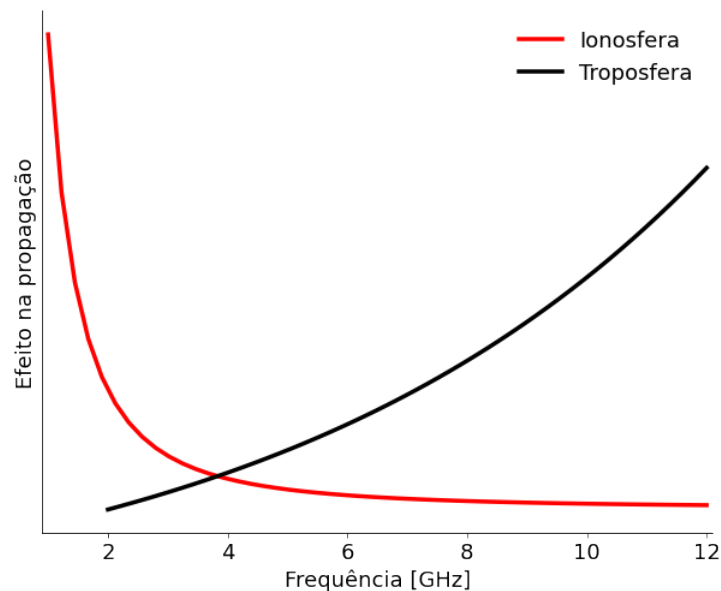
	REFERÊNCIAS	74
	APÊNDICE A – COEFICIENTES DOS MODELOS RLM	82
	APÊNDICE B – DETALHAMENTO DAS BASES DE DADOS	85
B.1	Definições de atributos	85
B.1.1	Magnetômetro	85
B.1.2	Ionossonda	85
B.2	Base de dados usada para regressão	86
B.3	Base de dados usada para classificação	88

1 Introdução

Na atualidade, nota-se uma tendência crescente no uso de tecnologias, sistemas, produtos e outras aplicações que, em algum momento, têm relação direta ao uso de satélites. É possível citar desde programas de televisão transmitidos por satélites de comunicação, a previsão do tempo determinada com o auxílio de satélites meteorológicos, até serviços de posicionamento e orientação devidos ao emprego de satélites de navegação [1]. Essa importância também é evidenciada pelo crescimento econômico observado nesse segmento, que registrou um aumento de 252% na quantidade de satélites na última década, além de uma receita global da ordem de 271 bilhões de dólares no ano de 2020 [2].

Satélites utilizam ondas de rádio para receber e enviar informações. O que habilita uma série de serviços como os sistemas de comunicação, distribuição ou *broadcasting* e navegação, os quais tornam-se suscetíveis à ocorrência de fenômenos relacionados com a propagação de ondas eletromagnéticas que podem inviabilizar ou comprometer seriamente o desempenho dessas aplicações. Em geral, sinais de satélite são degradados ao se propagarem pela atmosfera terrestre e em diversas situações esses efeitos são oriundos das camadas troposférica e ionosférica [3]. Na Figura 1 é possível observar como essas camadas atmosféricas afetam a propagação de ondas de rádio com diferentes frequências.

Figura 1 – Efeitos das camadas ionosfera e troposfera na propagação de ondas de rádio



Fonte: Adaptado de Arbesser-Rastburg [4]

A partir da Figura 1, nota-se que os efeitos troposféricos são significativos a partir de bandas de frequência na faixa dos 3 GHz. Por outro lado, os efeitos relacionados à ionosfera são mais expressivos em sinais que operam em faixas de até 6 GHz de frequência.

[5]. Dessa forma, uma das aplicações de satélites mais afetadas por fenômenos ionosféricos é a navegação, devido ao emprego da banda de frequências L, 1–2 GHz [6, 7].

Um dos principais e mais consolidados exemplos de uso dos satélites é a navegação. Um serviço inicialmente voltado para aplicações de posicionamento e localização, mas que evoluiu para atualmente estar presente em diversas atividades econômicas. Como a agricultura de precisão, o planejamento urbano, atividades de resgate e salvamento, monitoramento de fronteiras entre países, etc. Os diversos sistemas de navegação por satélite existentes são designados geralmente como Sistemas Globais de Navegação por Satélite, ou pelo acrônimo derivado do inglês, *GNSS* [8].

Em termos do impacto econômico das atividades associadas ao uso de satélites, o setor de navegação é o segundo maior gerador de receitas, com um valor estimado no ano de 2017 de cerca de 30 bilhões de dólares, superado somente pelo segmento de telecomunicações [8]. Em virtude da crescente dependência da sociedade atual em relação aos sistemas GNSS, é fundamental desenvolver formas de tratar os principais problemas técnicos referentes a essa aplicação, em geral, relacionados à transmissão dos sinais de satélite para os receptores em solo. Nessa etapa a propagação das ondas de rádio através da ionosfera é uma das maiores fontes de degradação do sinal, onde se observa que um dos efeitos mais relevantes é a *cintilação ionosférica* [3, 8, 9].

O termo cintilação ionosférica refere-se à ocorrência de variações em parâmetros tais como, amplitude, fase, direção de propagação e polarização, de ondas que se propagam na ionosfera. Tais variações são causadas por flutuações no índice de refração da camada ionosférica, relacionadas a irregularidades na densidade eletrônica do meio [10]. Esse fenômeno pode ter consequências em diversas atividades, desde atrasos e paralisações na exploração de recursos energéticos como óleo e gás, adiamentos de voos e aumento de custos operacionais relacionados à aviação, prejuízos logísticos devidos a erros de posicionamento e gerenciamento de rotas em operações de transporte de carga [11], até a interrupção nas comunicações via satélite em operações militares [12]. Em particular, um estudo da Agência Espacial Europeia [11] indicou que a ocorrência de cintilação ionosférica durante uma operação de lançamento pode causar uma suspensão parcial ou completa das comunicações entre o segmento solo e o veículo lançador. O que resulta na possibilidade de perda total do mesmo e de sua carga útil, o que acarretaria grandes prejuízos econômicos

Em razão das diversas aplicações baseadas em sistemas de telecomunicação e navegação via satélite, tecnologias vulneráveis aos efeitos da cintilação ionosférica, torna-se necessário desenvolver técnicas, tanto de caráter qualitativo quanto quantitativo, de modelagem e predição desse fenômeno [13, 14]. No entanto, o desenvolvimento de modelos preditivos para a cintilação ionosférica não é uma tarefa simples. Tal fato deve-se à complexidade do fenômeno, que apresenta uma grande variação espacial e temporal em virtude de sua dependência com relação a diversos fatores. Dos quais é possível

citar as atividades geomagnética, solar e interplanetária, o tempo local, a localização geográfica, além das características espaciais e temporais da ionosfera [15, 16, 17]. Nesse cenário, a aprendizagem de máquina, uma coleção de algoritmos capazes de detectar, de forma automática, padrões presentes em conjuntos de dados [18], vem sendo cada vez mais explorada como ferramenta para a construção de modelos preditivos relacionados à cintilação ionosférica.

Rezende *et al.* [19] desenvolveram modelos preditivos por meio de técnicas de aprendizagem de máquina para realizar a predição do índice S_4 , um importante parâmetro para caracterizar a cintilação de amplitude, com antecedência de 1, 2, 3 e 4 horas. Além disso, os autores apresentaram uma abordagem qualitativa cujo objetivo foi prever o nível de cintilação ionosférica no dia seguinte dentre as possibilidades de *cintilação fraca*, *moderada* ou *forte*. O conjunto de dados usado foi construído de modo a conter informações de outros fenômenos que influenciam a dinâmica da cintilação ionosférica, resultando em 6 variáveis de entrada do modelo, também conhecidas como atributos. Ademais, os autores usaram informações de localidades diferentes na mesma base de dados, a saber, as cidades de São Luís – MA e São José dos Campos – SP, indicando a possibilidade de usar a dependência geográfica do fenômeno estudado na construção de modelos preditivos.

Lima *et al.* [20] desenvolveram um modelo preditivo usando aprendizagem de máquina. O objetivo do trabalho foi, realizar a predição do nível de cintilação ionosférica com antecedência de algumas horas dentre duas possíveis classes, *cintilação forte* ou *cintilação não forte* na cidade de São Luís – MA. Os autores empregaram 13 atributos derivados de diversas fontes, para considerar diversos fatores que afetam a cintilação ionosférica

Jiao, Hall e Morton [17] desenvolveram modelos focados na cintilação ionosférica de amplitude. O objetivo de tais modelos foi realizar a distinção entre a ocorrência ou não de cintilação ionosférica em países localizados na região equatorial, tais como China, Peru e Singapura. Os atributos usados foram derivados a partir de receptores GNSS instalados nos locais de interesse. Sridhar *et al.* [15] realizaram a predição do índice S_4 com antecedência de 1, 2 e 3 horas, usando técnicas de aprendizagem de máquina. Os dados empregados na análise foram coletados na cidade de Darwin, Austrália, e os autores buscaram construir a base de dados para conter informações relativas a fenômenos correlatos à cintilação ionosférica.

McGranaghan *et al.* [21] desenvolveram um modelo preditivo usando ferramentas da aprendizagem de máquina, cujo objetivo foi determinar se haveria ou não a ocorrência de cintilação ionosférica de fase com uma antecedência de 1 ou 3 horas no Canadá. Os autores utilizaram uma base de dados com 51 atributos oriundos de diversas fontes, com o objetivo de reunir informações de uma série de fenômenos que possuem alguma relação com a cintilação ionosférica. Outro aspecto da pesquisa foi a realização de um estudo de

seleção dos melhores atributos para a construção dos modelos.

[Linty et al. \[22\]](#) usaram técnicas de aprendizagem de máquina para determinar a ocorrência da cintilação ionosférica de amplitude. Os modelos desenvolvidos empregaram atributos derivadas a partir do sinal captado por receptores GNNS e compreenderam dois países, o Vietnã e o Brasil. [Wu e Liu \[23\]](#) desenvolveram modelos preditivos baseados em aprendizagem de máquina para determinar a ocorrência ou não de cintilação com uma hora de antecedência em uma região do Alasca. Os autores empregaram atributos que caracterizavam outros fenômenos correlatos e, além disso, avaliaram quais atributos seriam mais adequadas para o desenvolvimento dos modelos.

[Atabati et al. \[24\]](#) empregaram técnicas de aprendizagem de máquina para construir modelos preditivos voltados à cintilação ionosférica na ilha de Guam, dentre outras coisas os autores realizaram a predição do índice S_4 com 1 dia de antecedência. Mais uma vez, as atributos usados continham informações de outros fenômenos importantes para a ocorrência de cintilação. [Ludwig-Barbosa et al. \[25\]](#) desenvolveram modelos relacionados à cintilação ionosférica tendo como principal ferramenta a aprendizagem de máquina. Os autores também realizaram um estudo específico à respeito de quais atributos seriam mais adequados para a situação tratada. Para organizar melhor as informações apresentadas, na Tabela 1 observam-se algumas informações referentes aos trabalhos citados. Na coluna *Horizonte de predição*, o traço (–) indica que o estudo realizado correspondente tratou o problema da cintilação em tempo real.

Tabela 1 – Síntese dos trabalhos citados

Ref.	Tarefa	Saída	Horizonte de predição
[15]	Regressão	Índice S_4	1, 2 e 3 horas
[17]	Classificação	Ocorrência de cintilação	–
[19]	Regressão	Índice S_4	1, 2, 3 e 4 horas
[20]	Classificação	Intensidade da cintilação	1 a 4 horas
[21]	Classificação	Ocorrência de cintilação	1 ou 3 horas
[22]	Classificação	Ocorrência de cintilação	–
[23]	Classificação	Ocorrência de cintilação	1 horas
[24]	Regressão	Índice S_4	24 horas
[25]	Classificação	Ocorrência de cintilação	–

Fonte: Autoria própria (2021)

Avaliando os estudos mencionados previamente, em conjunto com a Tabela 1, percebe-se que a maioria dos trabalhos tem um enfoque de classificação. Além disso, nos trabalhos voltados à predição, observa-se que os horizontes considerados são da ordem

de horas. No entanto, vale ressaltar a necessidade crescente de predições de curto prazo [13]. Dessa forma, o objetivo deste estudo é desenvolver modelos preditivos de curto prazo, tanto de caráter quantitativo quanto qualitativo, acerca da cintilação ionosférica de amplitude. Para isso, a abordagem selecionada consiste no uso de técnicas de aprendizagem de máquina. Ademais, são consideradas informações relativas à atividade geomagnética, à dependência temporal e geográfica do fenômeno, às atividades solar e interplanetária e ao estado da ionosfera, avaliando a influência de diferentes atributos no desempenho dos modelos obtidos.

As principais contribuições deste trabalho compreendem: o desenvolvimento de modelos preditivos de curto prazo, 30 minutos de antecedência, voltados para a cintilação ionosférica de amplitude; construção e análise de uma base de dados que abrange uma série de fenômenos correlatos à cintilação; avaliação do uso de alguns parâmetros determinados por meio de magnetômetros na construção de modelos de aprendizagem de máquina; avaliação do uso de alguns parâmetros derivados de ionossondas na construção de modelos de aprendizagem de máquina; uso e avaliação de um conjunto de parâmetros que modelam a influência temporal do fenômeno, de modo a verificar se os mesmos são úteis no desenvolvimento dos modelos preditivos; realização de um estudo de seleção de atributos.

Os resultados obtidos poderão contribuir para melhorar o desempenho de futuros modelos, algo de grande utilidade no cenário maranhense devido às perspectivas de uso comercial do Centro Espacial de Alcântara, que poderá ser afetado pela cintilação ionosférica. Este trabalho está organizado de forma que, no capítulo *Referencial teórico* serão apresentados os principais conceitos que auxiliarão a compreensão do trabalho. Em seguida as etapas da pesquisa serão descritas no capítulo de *Metodologia*. Os resultados serão apresentados e analisados no capítulo *Resultados e discussão* e, por fim, o trabalho será encerrado no capítulo *Conclusão*.

2 Objetivos

2.1 Objetivo Geral

Desenvolver modelos preditivos de curto prazo, tanto de caráter quantitativo quanto qualitativo, acerca da cintilação ionosférica de amplitude.

2.2 Objetivos Específicos

- Empregar a aprendizagem de máquina como principal ferramenta no desenvolvimento dos modelos preditivos.
- Desenvolver modelos preditivos capazes de prever o valor do índice S_4 com 30 minutos de antecedência, por meio de técnicas de regressão.
- Desenvolver modelos preditivos capazes de indicar a ocorrência de cintilação com antecedência de 30 minutos, por meio de modelos de classificação.
- Avaliar o desempenho de diferentes algoritmos de aprendizagem de máquina na construção dos modelos preditivos.
- Avaliar a influência de diferentes algoritmos de seleção de atributos nos resultados obtidos.
- Analisar a influência das informações relativas à atividade geomagnética, à dependência temporal e geográfica do fenômeno, às atividades solar e interplanetária e ao estado da ionosfera no desempenho dos modelos preditivos desenvolvidos.

3 Referencial teórico

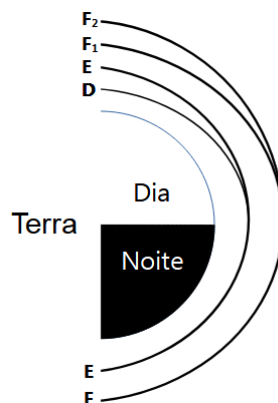
Neste capítulo serão apresentados os principais conceitos que auxiliarão a compreensão do restante do trabalho. As informações serão divididas nas seções *Cintilação ionosférica* e *Aprendizagem de máquina*.

3.1 Cintilação ionosférica

Antes de abordar a cintilação ionosférica, é importante ter um entendimento básico sobre a ionosfera, que pode ser definida como a porção da atmosfera onde existem íons e elétrons livres, em quantidade suficiente, para afetar a propagação de ondas de rádio [26]. Essa camada é constituída por um plasma produzido em um processo de fotoionização dos gases que compõem a alta atmosfera, desencadeado pela ação da radiação UV e de fótons com comprimento de onda curto provenientes do Sol. Além disso, em regiões de altas latitudes a precipitação de partículas oriundas da magnetosfera também contribui com o processo de ionização [27].

Devido à variação da composição da atmosfera neutra e das taxas de ionização em função da altitude, a ionosfera apresenta uma estrutura composta por camadas [27]. Entre 150 e 1000 km de altitude localiza-se a camada F, que, divide-se em F₁ e F₂ durante o dia, além disso, na região F₂ observa-se a máxima densidade eletrônica da ionosfera [26, 28, 29]. A região E situa-se entre 90 e 150 km, abaixo dessa faixa a cerca de 50 a 90 km observa-se a região D [26]. Durante a noite, além da recombinação das camadas F₁ e F₂ também constata-se o desaparecimento da região D [28]. Na Figura 2 encontra-se uma representação da organização geral da camada ionosférica e da dinâmica de suas subcamadas.

Figura 2 – Organização da ionosfera

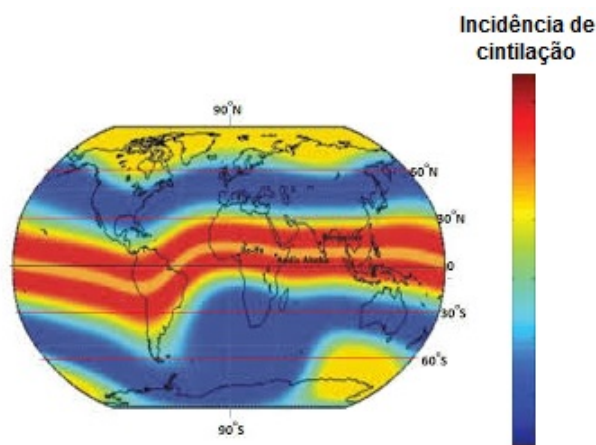


Fonte: Adaptado de [Haralambous, Papadopoulos e Economou \[30\]](#)

De posse de um entendimento básico sobre a ionosfera, é possível abordar o fenômeno estudado neste trabalho. Formalmente, a cintilação é definida como uma modulação aleatória transmitida aos campos de ondas, através de estruturas presentes em seus meios de propagação. Um exemplo cotidiano desse fenômeno é a cintilação de estrelas no céu noturno, causada pela turbulência presente na atmosfera terrestre [31]. No que se refere aos modelos usados no estudo da cintilação, é possível classificá-los como pertencentes às seguintes categorias, modelos analíticos, modelos climatológicos globais ou modelos baseados em dados coletados *in situ* [10].

A cintilação ionosférica apresenta uma forte dependência espacial. Em uma escala global, observa-se uma maior incidência do fenômeno nas regiões de baixas e de altas latitudes [32, 33]. Na Figura 3 é possível constatar a variação geográfica da cintilação ionosférica. O Brasil está entre as nações mais afetadas por esse fenômeno no planeta, condição relacionada diretamente às influências da Anomalia de Ionização Equatorial (AIE) e da Anomalia Magnética do Atlântico Sul (AMAS) sobre o território nacional [34, 35]. Em 2019, um relatório produzido pela Agência Espacial Brasileira [36] apresentou as principais demandas do setor espacial local, destacando a necessidade da realização de estudos direcionados à cintilação ionosférica para garantir a integridade de uma série de aplicações afetadas por esse fenômeno.

Figura 3 – Incidência da cintilação ionosférica globalmente



Fonte: Adaptado de Kintner, Humphreys e Hinks [37]

A Anomalia de Ionização Equatorial é caracterizada pela existência de um valor mínimo da densidade de ionização na região do equador magnético, contraposto por valores máximos localizados a cerca de 20° de latitude magnética nas direções norte e sul [35]. Na Figura 3 as faixas vermelhas representam as regiões de máxima ionização, também conhecidas como cristas da AIE, onde observa-se a ocorrência de cintilação com maior intensidade [38]. A Anomalia Magnética do Atlântico Sul é reconhecida como a maior anomalia do campo magnético terrestre, onde o mesmo alcança níveis mínimos de intensidade, o que facilita a penetração de partículas vindas do espaço na atmosfera [39].

Em geral, a cintilação ionosférica é classificada como *cintilação de fase* ou *cintilação de amplitude*. Os efeitos da primeira categoria são mais relevantes em regiões polares, já nas regiões de baixas latitudes o impacto da cintilação de amplitude é mais significativo [33]. Do ponto de vista físico, a cintilação ionosférica pode ser compreendida da seguinte forma: as irregularidades na densidade eletrônica da ionosfera fazem com que o índice de refração do meio apresente um comportamento anisotrópico. Dessa forma, quando um sinal se propaga nessa região, observa-se não somente a refração do mesmo, mas também uma alteração na sua fase. Ao nível dos receptores, a combinação de diversos componentes do sinal, que chegam por diferentes direções e com amplitudes e fases distintas após se propagarem na ionosfera, gera um sinal resultante com fase e amplitude aleatórias, caracterizando a cintilação ionosférica [6].

Para complementar o que foi apresentado anteriormente, a perspectiva matemática do fenômeno será brevemente abordada. Para compreender a formulação matemática da cintilação ionosférica, a etapa inicial consiste em caracterizar a propagação de ondas eletromagnéticas em um meio aleatório. Para isso, considere as equações de Maxwell:

$$\nabla \times \mathbf{E} = -\frac{1}{c} \frac{\partial \mathbf{B}}{\partial t}, \quad (3.1)$$

$$\nabla \times \mathbf{H} = \frac{1}{c} \frac{\partial \mathbf{D}}{\partial t} + \frac{4\pi}{c} \mathbf{J}, \quad (3.2)$$

$$\nabla \cdot \mathbf{D} = 4\pi\rho_e, \quad (3.3)$$

$$\nabla \cdot \mathbf{B} = 0. \quad (3.4)$$

Nas expressões (3.1) a (3.4), os parâmetros em negrito denotam quantidades vetoriais. Além disso, o termo $\mathbf{E}$ indica o vetor campo elétrico; c a velocidade da luz no vácuo; $\mathbf{B}$ o vetor densidade de fluxo magnético; $\mathbf{H}$ o vetor campo magnético; $\mathbf{D}$ o vetor densidade de fluxo elétrico; $\mathbf{J}$ o vetor densidade de corrente elétrica; ρ_e a densidade líquida de carga elétrica. Ademais, as unidades de medida desses parâmetros, expressas no sistema de unidades gaussianas, estão indicadas na página 11. No mais, ∇ e $\partial/\partial t$ denotam o operador nabla e a derivada parcial em relação ao tempo t . As operações $\nabla \times$ e $\nabla \cdot$ designam, respectivamente, o rotacional e o divergente de um campo vetorial.

Na região de transmissão da onda eletromagnética, é possível desconsiderar o termo ρ_e , de forma que a equação (3.3) assume a seguinte forma:

$$\nabla \cdot \mathbf{D} = 0. \quad (3.5)$$

Para caracterizar o meio de propagação, usam-se duas relações constitutivas, a primeira delas é dada por:

$$\mathbf{B} = \mu_m \mathbf{H}. \quad (3.6)$$

Em (3.7), μ_m indica permeabilidade magnética do meio. No sistema de unidades adotado, μ_m é unitária na atmosfera terrestre. Desse modo, a equação (3.6) se reduz a $\mathbf{B} = \mathbf{H}$. Além disso, tem-se:

$$\mathbf{D} = \varepsilon(\mathbf{r}, t) \mathbf{E}. \quad (3.7)$$

A constante dielétrica do meio, $\varepsilon(\mathbf{r}, t)$, carrega toda a informação necessária para descrever a propagação de ondas eletromagnéticas em um meio aleatório. Geralmente, esse parâmetro é decomposto da seguinte forma:

$$\varepsilon(\mathbf{r}, t) = \varepsilon_0(\mathbf{r}) + \Delta\varepsilon(\mathbf{r}, t). \quad (3.8)$$

Na expressão (3.8), ε_0 denota um valor médio que pode ser considerado unitário na atmosfera. O termo $\Delta\varepsilon$ representa as flutuações estocásticas da constante dielétrica, as quais dão origem ao fenômeno da cintilação. Para obter a equação diferencial que governa a propagação de ondas eletromagnéticas no meio considerado, é necessário combinar as equações de Maxwell. Aplicando o operador rotacional na expressão (3.1), obtém-se:

$$\begin{aligned} \nabla \times \nabla \times \mathbf{E} &= -\frac{1}{c} \nabla \times \frac{\partial \mathbf{H}}{\partial t} = -\frac{1}{c} \frac{\partial}{\partial t} (\nabla \times \mathbf{H}) \\ &= -\frac{1}{c} \frac{\partial}{\partial t} \left(\frac{1}{c} \frac{\partial \mathbf{D}}{\partial t} + \frac{4\pi}{c} \mathbf{J} \right) \\ &= -\frac{1}{c^2} \frac{\partial^2 \mathbf{D}}{\partial t^2} - \frac{4\pi}{c^2} \frac{\partial \mathbf{J}}{\partial t}. \end{aligned} \quad (3.9)$$

É possível aplicar uma identidade vetorial no membro esquerdo da equação (3.9) para obter:

$$\nabla \times \nabla \times \mathbf{E} = -\nabla^2 \mathbf{E} + \nabla(\nabla \cdot \mathbf{E}). \quad (3.10)$$

Note que, usando (3.5) obtém-se:

$$\nabla \cdot \mathbf{D} = \nabla \cdot (\varepsilon \mathbf{E}) = \mathbf{E} \cdot \nabla \varepsilon + \varepsilon \nabla \cdot \mathbf{E} = 0 \Rightarrow \nabla \cdot \mathbf{E} = -\mathbf{E} \cdot \nabla(\log \varepsilon). \quad (3.11)$$

Na expressão (3.11), o operador $\log$ denota o logaritmo natural. Combinando as equações (3.8), (3.9), (3.10) e (3.11), obtém-se:

$$\nabla^2 \mathbf{E} - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} [(1 + \Delta\varepsilon) \mathbf{E}] = \frac{4\pi}{c^2} \frac{\partial \mathbf{J}}{\partial t} - \nabla \{ \mathbf{E} \cdot \nabla [\log(1 + \Delta\varepsilon)] \}. \quad (3.12)$$

A equação (3.12) é a expressão mais geral que descreve o comportamento do campo elétrico $\mathbf{E}$, o qual é gerado por um transmissor e sofre alterações devido às flutuações $\Delta\varepsilon$

presentes no meio de propagação [40]. Diferentes teorias são usadas para interpretar os dados experimentais relacionados à cintilação ionosférica, nos últimos 60 anos destaca-se a teoria de espalhamento fraco, a aproximação de Rytov, teorias de tela de fase única, delgada e múltipla, além da teoria de espalhamento múltiplo [10].

O foco deste trabalho será a cintilação de amplitude. Um dos parâmetros mais importantes para caracterizar esse fenômeno é o chamado índice S_4 , definido como o desvio padrão da intensidade do sinal recebido normalizado pelo seu valor médio. O índice S_4 é dado por:

$$S_4 = \frac{\sqrt{\langle I^2 \rangle - \langle I \rangle^2}}{\langle I \rangle}. \quad (3.13)$$

Na equação (3.13), I denota a intensidade do sinal, geralmente medida em dBW, e $\langle \cdot \rangle$ representa a operação de média temporal. Usualmente esse parâmetro é calculado sob uma janela de 60 segundos a partir de receptores GNSS com taxas de aquisição de 50 Hz [22, 41, 42, 43]. Na próxima seção serão abordados algumas técnicas e conceitos relacionados à aprendizagem de máquina.

3.2 Aprendizagem de máquina

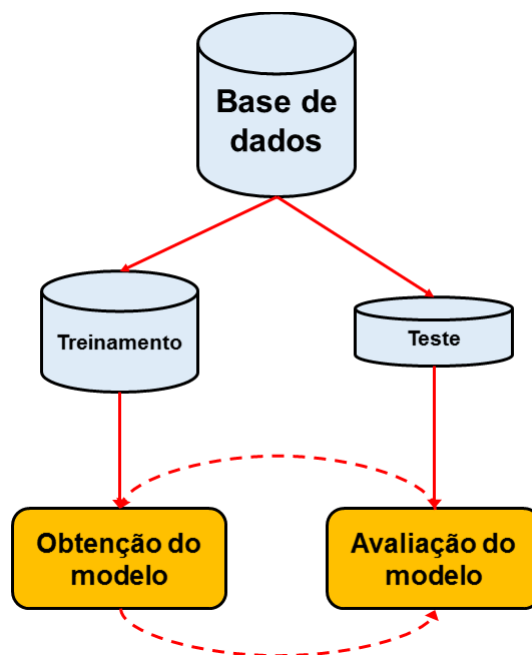
Em geral, é possível subdividir o campo da aprendizagem de máquina em três modalidades distintas, quais sejam, o aprendizado supervisionado, o aprendizado não supervisionado e o aprendizado por reforço. Neste estudo o foco será na aprendizagem de máquina supervisionada, cujo objetivo é determinar um mapeamento que relacione variáveis de entrada, também referenciadas como atributos, usualmente denotadas pela letra x , a variáveis de saída, também indicadas como variáveis dependentes, y . Dentro da categoria de aprendizado supervisionado, tem-se algoritmos dedicados às tarefas de regressão e classificação. Na regressão as variáveis y assumem valores contínuos, ou seja, $y \in \mathbb{R}$, já na classificação $y \in \mathbb{Z}$, ou seja, a variável y assume valores discretos [18].

Para definir melhor a terminologia usada neste trabalho, um atributo é tido como uma propriedade individual mensurável, relacionada ao fenômeno estudado [44]. Em geral, o i -ésimo atributo é denotado por x_i . Já um exemplo, é considerado como um conjunto de atributos que caracterizam o fenômeno analisado em um dado contexto. Um exemplo pode ser representado por: $(x_1, x_2, \dots, x_i)$.

O processo geral de construção de um modelo de aprendizagem de máquina está indicado na Figura 4. Inicialmente a base de dados utilizada é dividida em dois subconjuntos, um dedicado ao treinamento e o outro ao teste do modelo. Em aprendizagem de máquina, treinar um modelo significa fornecer exemplos pertencentes ao conjunto de dados para que um algoritmo consiga detectar os padrões existentes. O conjunto de testes é empregado

para avaliar o desempenho do modelo obtido na etapa de treinamento. As setas tracejadas na Figura 4 indicam a interação entre as etapas de obtenção e avaliação dos modelos. Nessa fase, o modelo é obtido e posteriormente avaliado usando os exemplos do conjunto de testes. Caso o desempenho não seja satisfatório, retorna-se à etapa anterior para obter um modelo diferente com as devidas melhorias. Esse ciclo se repete até que o modelo apresente um desempenho satisfatório. A seguir serão discutidas brevemente algumas particularidades do desenvolvimento de modelos preditivos por meio da aprendizagem de máquina.

Figura 4 – Construção de um modelo de aprendizagem de máquina



Fonte: Autoria própria (2021)

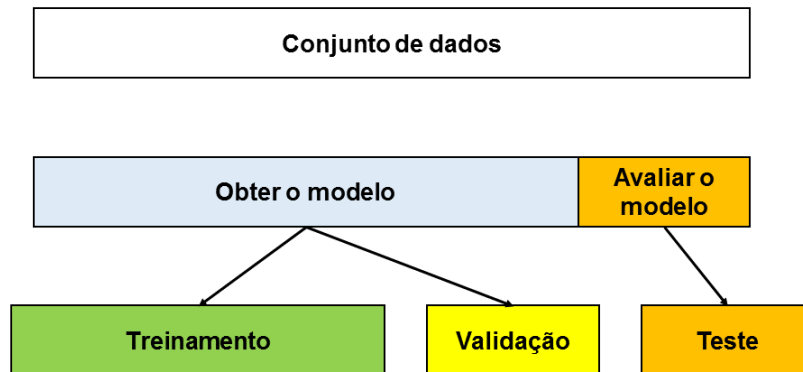
3.2.1 Validação cruzada

A etapa de segmentação dos dados em um subconjunto usado para obter o modelo e outro para avaliá-lo ilustrada na Figura 4, é conhecida como validação cruzada, um método estatístico amplamente usado em aprendizagem de máquina [45]. As hipóteses relacionadas ao método de validação cruzada são as de que os dados sejam identicamente distribuídos e os conjuntos de treino e validação sejam independentes. Dessa forma, essa metodologia pode ser aplicada em diversas categorias de algoritmos, sejam eles de classificação ou regressão [46].

Para compreender algumas técnicas de validação cruzada é importante entender a relação entre os conjuntos de treino, validação e teste. Na Figura 5, é possível observar a interação entre tais conjuntos de dados. Aonde nota-se que a porção dos dados destinada à obtenção do modelo é dividida em dois subconjuntos, um de treino e outro de validação. Este último serve para oferecer uma estimativa do desempenho do modelo desenvolvido por

meio do conjunto de treinamento, o que auxilia na definição de uma série de parâmetros do modelo. Por outro lado, o conjunto de testes tem como função fornecer uma estimativa do desempenho do modelo definitivo em relação a exemplos desconhecidos [47].

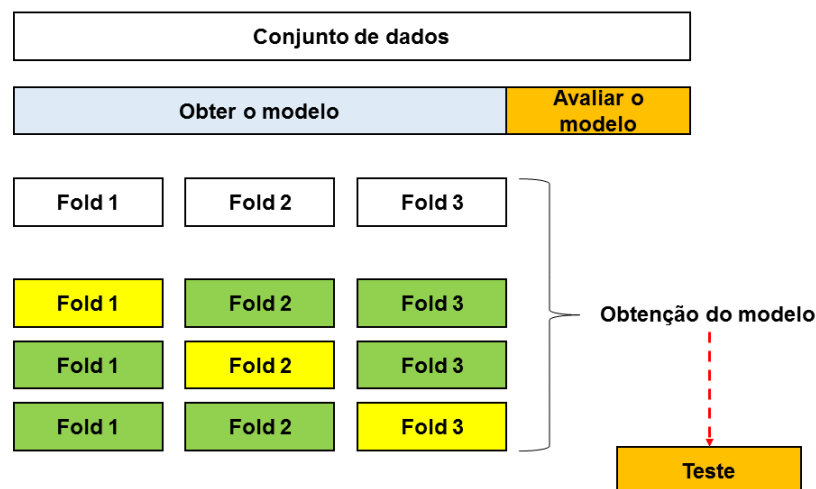
Figura 5 – Relação entre os conjuntos de treino, validação e teste



Fonte: Autoria própria (2021)

Uma das técnicas básicas de validação cruzada é conhecida como *k-fold* [45]. Essa abordagem é baseada em dividir, de forma aleatória, o conjunto de dados disponível em k grupos ou *folds* com aproximadamente o mesmo número de elementos. Em seguida, um dos grupos é tratado como conjunto de validação enquanto o modelo é obtido usando-se os demais $k - 1$ grupos. Esse processo é repetido k vezes, em cada uma delas um grupo diferente é definido como conjunto de validação. A estimativa de erro do modelo é então obtida por meio da média aritmética dos k valores de erro resultantes das etapas anteriores [48]. A Figura 6 ilustra a aplicação da metodologia *3-fold*.

Figura 6 – Validação cruzada *3-fold*

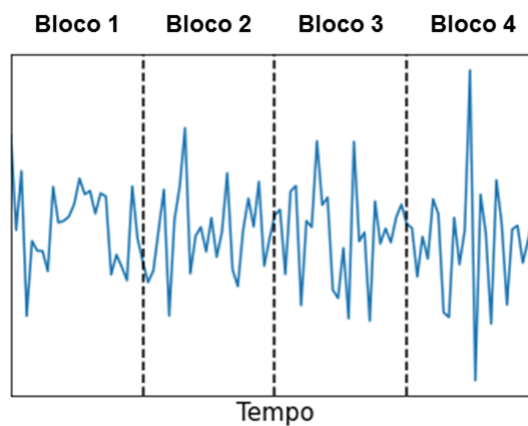


Fonte: Autoria própria (2021)

Uma particularidade do presente estudo é a construção de modelos preditivos de séries temporais. Nesse caso, a metodologia *k-fold* tradicional deve ser alterada ligeiramente pois, em geral, exemplos obtidos sequencialmente no tempo apresentam estruturas de

dependência interna, caracterizadas pela tendência de observações próximas apresentarem maior similaridade [49]. Nessas circunstâncias, o processo de validação cruzada deve ser modificado, visto que a dependência existente no conjunto de dados viola uma hipótese fundamental dessa técnica [46, 49]. Essa questão pode ser tratada usando uma abordagem caracterizada pelo emprego de blocos de dados, o que aumenta a independência entre os conjuntos de treino e validação [49, 50]. Essa abordagem alternativa será referenciada como validação cruzada *k-fold* em blocos. A qual consiste em inicialmente dividir o conjunto de dados em grupos de exemplos ordenados, da forma indicada na Figura 7, distribuídos em seguida nas diferentes *folds*, ou seja, ao invés de atribuir exemplos individuais às *folds*, são atribuídos grupos de exemplos.

Figura 7 – Divisão dos dados em blocos



Fonte: Autoria própria (2021)

Note que a validação cruzada *k-fold* em blocos é compatível com as variações da metodologia *k-fold*, como o *k-fold* repetido. Neste trabalho, a validação cruzada *k-fold* em blocos foi utilizada na etapa de obtenção dos modelos preditivos.

3.2.2 Otimização de hiperparâmetros

Outro aspecto importante relacionado à aprendizagem de máquina é a definição dos *hiperparâmetros* do modelo, uma série de parâmetros que devem ser especificados antes da etapa de treinamento. O processo de determinação de valores adequados a tais parâmetros de modo a obter modelos com desempenhos satisfatórios, dentro de um intervalo de tempo factível é conhecido como *otimização de hiperparâmetros*. Vale ressaltar que essa etapa é determinante para a acurácia de predição dos modelos desenvolvidos [51, 52].

As principais abordagens quando se trata da otimização de hiperparâmetros correspondem aos métodos de seleção manuais e automáticos. Dentre os métodos automáticos, destaca-se a busca cartesiana ou busca de *grid*. Tal estratégia consiste em uma busca exaustiva que avalia todas as combinações possíveis de hiperparâmetros do modelo para

determinar a configuração ótima [51, 53]. Apesar da busca de *grid* teoricamente ser capaz de determinar o conjunto ótimo de hiperparâmetros, ela sofre com a chamada maldição da dimensionalidade. Visto que à medida que o número de hiperparâmetros a serem definidos cresce, o custo computacional associado a essa estratégia também aumenta, tornando-a pouco prática em problemas multidimensionais [51].

Nesse contexto, Bergstra e Bengio [53] propuseram um algoritmo de busca aleatória. O qual consiste em, dado um espaço de buscas, amostrar e testar N combinações aleatórias de hiperparâmetros. Posteriormente, os modelos obtidos são avaliados e os hiperparâmetros daquele de melhor desempenho são selecionados. Essa abordagem possui as vantagens relacionadas à busca de *grid*, ou seja, a facilidade de implementação e a reprodutibilidade, ao mesmo tempo em que é eficiente em problemas multidimensionais.

3.2.3 Seleção de atributos

Além da otimização de hiperparâmetros, o desempenho de modelos baseados em aprendizagem de máquina pode ser aperfeiçoada por meio da seleção de atributos. Essa etapa possui como principal enfoque selecionar um subconjunto de atributos, de modo que o mesmo seja capaz de representar os dados de entrada de forma satisfatória. O que possibilita o desenvolvimento de modelos com melhores capacidades preditivas [44]. Os principais benefícios associados à seleção de atributos são a melhora no desempenho preditivo dos modelos de aprendizagem de máquina, a redução no custo computacional envolvido na construção do modelo e uma maior compreensão da base de dados usada [54].

Em geral, as abordagens de seleção de atributos são categorizadas em: filtragem, métodos *wrapper* e métodos embarcados. Na primeira categoria, utiliza-se algum critério pré-definido para estabelecer um ranking de importância entre os atributos disponíveis. Em seguida selecionam-se os k atributos melhor ranqueados. Nos métodos *wrapper*, emprega-se uma estratégia de busca para selecionar os atributos mais adequados para um modelo específico. Finalmente, nos métodos embarcados a ideia básica é incorporar o processo de seleção de atributos na etapa de treinamento do modelo [44]. Um exemplo de abordagem embarcada é uso do algoritmo Random Forest (RF), que oferece uma forma de avaliar a importância dos atributos usados na construção do modelo [55].

Neste trabalho serão empregados métodos de filtragem e embarcado para realizar a seleção de atributos. Os rankings de importância de atributos foram derivados por meio das seguintes estratégias: correlação linear, informação mútua, métrica de Fisher e RF. As três primeiras serão exploradas a seguir, o algoritmo RF será tratado posteriormente.

Uma forma simples e rápida de avaliar a relação entre duas variáveis é admitir que, se uma delas é capaz de fornecer informação significativa para realizar a previsão da segunda, deve existir algum grau de correlação entre as mesmas [56]. Nesse contexto,

um dos métodos mais simples de seleção de atributos consiste no uso do coeficiente de correlação de Pearson, ρ , dado por [44]:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}. \quad (3.14)$$

Na equação (3.14), σ_{XY} , σ_X e σ_Y representam a covariância entre as variáveis X e Y , o desvio padrão da variável X e o desvio padrão da variável Y , respectivamente.

O critério de Informação Mútua (MI) é tido como uma forma satisfatória de avaliar o grau de relação entre conjuntos de dados. Pois, é capaz de detectar dependências de qualquer tipo entre as variáveis consideradas, não se limitando a relações de natureza linear [57, 58]. A MI entre cada uma das variáveis dependentes, X_i , e a variável independente, Y , é definida de acordo como [54]:

$$MI(X_i, Y) = \iint p(x_i, y) \log \left(\frac{p(x_i, y)}{p(x_i) \cdot p(y)} \right) dx_i dy \quad (3.15)$$

Na equação (3.15), $p(x_i)$, $p(y)$ e $p(x_i, y)$ representam, respectivamente, as densidades de probabilidade das variáveis X_i , Y e a densidade de probabilidade conjunta entre as mesmas. O obstáculo relacionado à expressão (3.15) está no fato das densidades de probabilidade $p(x_i)$, $p(y)$ e $p(x_i, y)$ serem, em geral, desconhecidas. O que justifica a necessidade de estimá-las com base nos dados disponíveis [54].

A métrica de Fisher, também referida como *F score*, é amplamente empregada para selecionar atributos no contexto dos modelos de classificação. Essa estratégia consiste em computar a estatística F, em geral, usada no contexto da Análise de Variância (ANOVA), para cada um dos atributos. A ideia básica dessa abordagem, é a de que bons atributos são aqueles que apresentam valores distintos em exemplos de diferentes classes e valores semelhantes em exemplos da mesma classe [59]. A métrica de Fisher do atributo x_i é dada por [60]:

$$F(x_i) = \frac{\frac{1}{C-1} \sum_{k=1}^C n_k (\mu_k^i - \mu_i)^2}{\frac{1}{N-C} \sum_{k=1}^C \sum_{j=1}^{n_k} (x_{j,i}^k - \mu_k^i)^2}. \quad (3.16)$$

Na equação (3.16), C denota o número de classes, N o número total de exemplos, n_k representa a quantidade de exemplos pertencentes à k -ésima classe, μ_k^i é o valor médio do atributo i correspondente aos exemplos da k -ésima classe, μ_i denota o valor médio do atributo i e $x_{j,i}^k$ representa o valor do atributo i do j -ésimo exemplo pertencente à classe k .

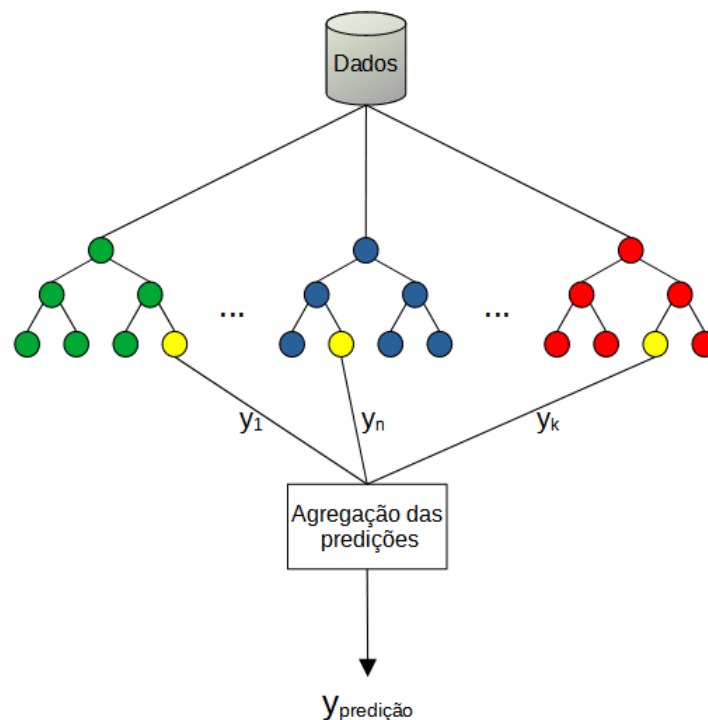
3.2.4 Algoritmos

Nesta seção serão abordados alguns aspectos dos algoritmos usados para obter os modelos preditivos, quais sejam, o algoritmo RF, a regressão linear múltipla e a regressão logística.

3.2.4.1 Random Forest

O algoritmo RF consiste em um modelo gerado a partir de um conjunto de k sub-modelos estruturados em forma de árvore, que fornecem previsões individuais. A resposta final do modelo é obtida agregando as previsões de seus respectivos sub-modelos. No processo de construção do conjunto de tais sub-modelos, ou *ensemble*, o crescimento de cada árvore é governado por um vetor aleatório. Tal abordagem é aplicável tanto em problemas de classificação quanto de regressão. No primeiro caso as previsões das k árvores são agregadas por meio de um sistema de votação e no segundo realiza-se a média aritmética de todas as previsões [61]. Na Figura 8 é possível observar a representação do funcionamento do algoritmo Random Forest.

Figura 8 – Algoritmo Random Forest



Fonte: Autoria própria (2021)

3.2.4.2 Regressão Linear Múltipla

Um modelo de Regressão Linear Múltipla tem a forma descrita na seguinte expressão:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon. \quad (3.17)$$

Na equação (3.17), X_j denota a j -ésima variável independente, β_j quantifica a relação entre X_j e a variável dependente Y e pode ser interpretado como o efeito médio em Y causado pelo aumento de uma unidade na variável X_j , mantendo-se as demais fixadas. Por fim, ϵ representa um termo de erro aleatório com média zero [48]. Os coeficientes β_j podem ser estimados por meio da equação:

$$\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (3.18)$$

Na equação (3.18), β denota um vetor com os coeficientes β_j , $\mathbf{X}$ representa uma matriz cujas linhas são os exemplos de entrada e $\mathbf{y}$ é um vetor que armazena os valores assumidos pela variável dependente. A obtenção da expressão (3.18), também conhecida como equação normal, por meio do método dos mínimos quadrados pode ser consultada em [Hastie, Tibshirani e Friedman \[62\]](#).

3.2.4.3 Regressão Logística

O algoritmo regressão logística fornece a probabilidade p de um exemplo X pertencer a uma determinada classe. Essa probabilidade é determinada por meio da expressão (3.19) [48].

$$p(X) = \frac{e^{\alpha_0 + \alpha_1 X_1 + \dots + \alpha_m X_m}}{1 + e^{\alpha_0 + \alpha_1 X_1 + \dots + \alpha_m X_m}} \quad (3.19)$$

Os parâmetros α_i são estimados através do método de máxima verossimilhança, para mais detalhes, veja [James *et al.* \[48\]](#).

3.2.5 Avaliação e comparação de modelos de aprendizagem de máquina

A avaliação do desempenho dos modelos preditivos construídos será realizada por meio de métricas de desempenho. Naturalmente, as métricas empregadas para os modelos de regressão serão diferentes daquelas usadas para os modelos de classificação. As métricas de desempenho selecionadas para avaliar os modelos de regressão foram o coeficiente de determinação, o erro quadrático médio, o erro absoluto médio e o erro percentual absoluto médio, indicados respectivamente nas expressões (3.20), (3.21), (3.22) e (3.23):

$$R^2 = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y}_i)^2}, \quad (3.20)$$

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2, \quad (3.21)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|, \quad (3.22)$$

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{\hat{y}_i - y_i}{y_i} \right|. \quad (3.23)$$

Nas equações (3.20), (3.21), (3.22) e (3.23), $\hat{y}$, y e $\bar{y}$ denotam o valor predito, o valor real e o valor médio da variável y , ademais, N representa o número de exemplos utilizados. Antes de introduzir as métricas usadas para avaliar os modelos de classificação, é necessário compreender a matriz de confusão, uma forma gráfica de indicar os resultados obtidos por meio de um classificador. Na Figura 9 é possível analisar a configuração geral de uma matriz de confusão, que exhibe a quantidade de exemplos pertencentes à classe i classificados na classe j . Na Figura 9, NSC e SC indicam as classes *não cintilação* e *cintilação*, respectivamente. Os termos indicados no interior da matriz são:

- VN: Quantidade de Verdadeiros Negativos, ou seja, quantidade de exemplos pertencentes à classe *não cintilação* e classificadas corretamente.
- VP: Quantidade de Verdadeiros Positivos, ou seja, quantidade de exemplos pertencentes à classe *cintilação* e classificadas corretamente.
- FP: Quantidade de Falsos Positivos, ou seja, quantidade de exemplos pertencentes à classe *não cintilação* e classificadas incorretamente.
- FN: Quantidade de Falsos Negativos, ou seja, quantidade de exemplos pertencentes à classe *cintilação* e classificadas incorretamente.

Figura 9 – Matriz de confusão

		Classe predita	
		NSC	SC
Classe real	NSC	VN	FP
	SC	FN	VP

Fonte: Autoria própria (2021)

A partir da matriz de confusão, é possível computar as métricas, acurácia, precisão, sensibilidade, medida F1 e especificidade dadas, respectivamente, por:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}, \quad (3.24)$$

$$\text{Precisão} = \frac{VP}{VP + FP}, \quad (3.25)$$

$$\text{Sensibilidade} = \frac{VP}{VP + FN}, \quad (3.26)$$

$$\text{Medida F1} = \frac{2 \cdot \text{Precisão} \cdot \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}}, \quad (3.27)$$

$$\text{Especificidade} = \frac{VN}{VN + FP}. \quad (3.28)$$

Além de determinar os desempenhos dos modelos, é interessante compará-los entre si. Em geral, tais comparações entre algoritmos de aprendizagem de máquina são feitas por meio de testes estatísticos de significância [63]. Os testes mais empregados quando se deseja avaliar se há diferenças significantes nos desempenhos de dois algoritmos de aprendizagem de máquina são o teste- t e o teste de postos com sinal de Wilcoxon. O primeiro tem natureza paramétrica e baseia-se em pressupostos que raramente são satisfeitos em análises de modelos de aprendizagem de máquina, o que leva ao uso do segundo teste, de caráter não paramétrico [64].

O teste de postos com sinal de Wilcoxon é realizado da seguinte forma, considere que deseja-se comparar o desempenho de dois modelos, referenciados como 1 e 2. Primeiro calculam-se as diferenças entre as métricas de desempenho dos dois modelos. Denota-se a diferença entre as i -ésimas métricas de desempenho como d_i . Em seguida, essas diferenças são ranqueadas de acordo com seus valores absolutos, nos casos em que se observar um empate entre dois valores d_i , realiza-se uma média aritmética para a atribuição dos rankings. A próxima etapa do teste consiste em realizar dois somatórios, um dos quais considera todos os rankings onde o modelo 1 superou o modelo 2 e o outro considera o caso oposto, indicam-se esses somatórios por R^+ e R^- , dados por [65]:

$$R^+ = \sum_{d_i > 0} \text{ranking}(d_i),$$

$$R^- = \sum_{d_i < 0} \text{ranking}(d_i).$$

Casos nos quais $d_i = 0$ são ignorados. Em seguida calcula-se o parâmetro $T = \min(R^+, R^-)$ que possibilita o cálculo da estatística z dada por:

$$z = \frac{T - N(N + 1)/4}{\sqrt{N(N + 1)(2N + 1)/24}}. \quad (3.29)$$

Na expressão (3.29), N denota o número de amostras avaliadas menos a quantidade de diferenças nulas [65]. Nos casos em que $N > 25$ a estatística z pode ser aproximada por uma distribuição normal [64].

4 Metodologia

Neste capítulo serão descritas as etapas que possibilitaram a obtenção dos resultados do trabalho. Para fins de organização, a metodologia foi organizada em três estágios, *Preparação da base de dados*, *Desenvolvimento dos modelos preditivos* e *Avaliação dos modelos preditivos*. No primeiro serão fornecidos detalhes acerca das informações consideradas e de como as mesmas foram obtidas. Em seguida, o processo de desenvolvimento dos modelos, que compreende a otimização de hiperparâmetros e a seleção de atributos, será descrita. Finalmente, o processo de avaliação dos modelos será detalhado.

4.1 Preparação da base de dados

Nesta seção serão descritas as etapas de construção das bases de dados utilizadas para obter os modelos de regressão e classificação. Um ponto comum às duas bases de dados é a consideração de fenômenos e aspectos que influenciam em algum grau a cintilação ionosférica. Dentre os quais é possível destacar as atividades geomagnética, solar e interplanetária, o tempo local, a localização geográfica, além das características espaciais e temporais da ionosfera [15, 16, 17]. Portanto, é importante que modelos orientados à predição da cintilação ionosférica considerem informações que representem, tanto quanto possível, tais aspectos subjacentes [16].

4.1.1 Base de dados usada para regressão

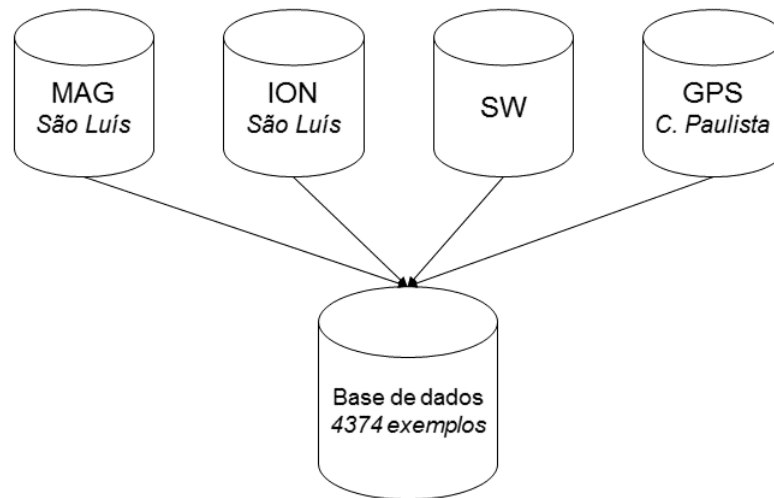
Os modelos de regressão foram desenvolvidos considerando dados referentes à cidade de Cachoeira Paulista – SP ($22^{\circ}39'54''\text{S}$ - $45^{\circ}00'34''\text{O}$). Esse local foi escolhido por estar localizado sob a crista sul da Anomalia de Ionização Equatorial, uma região altamente afetada pela cintilação ionosférica [35, 66]. O período considerado compreendeu os meses de outubro e novembro de 2014, visto que a variação sazonal do fenômeno estudado leva a uma intensificação de seus efeitos no território brasileiro nos meses de setembro a março [67]. O ano de 2014 foi selecionado por corresponder ao ano de máxima atividade do ciclo solar 24, o que intensifica o fenômeno da cintilação ionosférica [68]. A seguir estão descritas as informações contidas no conjunto de dados, que conta com 4374 exemplos.

Os aspectos geográficos foram explorados da seguinte forma, a base de dados foi construída com informações coletadas nas cidades de São Luís – MA (2.3°S , 44.2°W) e Cachoeira Paulista – SP (22.7°S , 45.0°W). Essa escolha visou, de forma análoga a Rezende *et al.* [19], considerar os efeitos das bolhas de plasma que se originam em regiões equatoriais, onde se localiza a cidade de São Luís, e impactam a região da chamada Anomalia de

Ionização Equatorial, sob a qual se encontra a cidade de Cachoeira Paulista [35, 66].

A Figura 10 ilustra parte do processo de construção da base de dados. Foram consideradas informações referentes à atividade geomagnética, indicadas pela base *MAG*, às condições solar e interplanetária, indicadas pela base *SW* e ao estado da ionosfera, indicadas pelas bases *ION* e *GPS*. Na Figura 10 estão indicadas as cidades a partir das quais as informações foram coletadas. A seguir as bases de dados indicadas serão exploradas individualmente.

Figura 10 – Construção da base de dados usada nos modelos de regressão



Fonte: Autoria própria (2021)

A atividade geomagnética, representada pela base *MAG*, foi considerada mediante o uso de dados coletados por magnetômetros, dispositivos usados para medir a intensidade e direção do campo magnético terrestre [69]. Tais informações foram disponibilizados por meio do portal *Space Weather Data Share*¹, mantido pelo programa Estudo e Monitoramento Brasileiro do Clima Espacial, Embrace/INPE². Os sensores empregados são provenientes da rede MagNet, composta por magnetômetros do tipo núcleo saturado fabricados pelo Rádio Observatório de Jicamarca. Equipamentos sensíveis a campos magnéticos da ordem de 1 mT, com resolução máxima de até 10 pT [70]. Os parâmetros considerados estão dispostos na Tabela 2.

¹ <<http://www2.inpe.br/climaespacial/SpaceWeatherDataShare/>>

² <<http://www2.inpe.br/climaespacial/portal/pt/>>

Tabela 2 – atributos relacionadas à atividade geomagnética

atributo	Descrição
F_M	Módulo do campo principal
H_M	Componente horizontal do vetor campo magnético
D_M	Declinação do campo magnético

Fonte: Autoria própria (2021)

As condições solar e interplanetária, indicadas pela base *SW*, foram representadas através de dados oriundos da ferramenta *OMNIWeb* ³. Que consiste em uma base de dados mantida pela NASA que reúne informações, coletadas por diversos satélites em órbita, a respeito do campo magnético e do plasma que constituem os ventos solares [71]. A componente z do campo magnético interplanetário, B_z , está expressa no sistema de coordenadas GSM, do inglês *Geocentric Solar Magnetospheric*, o mais adequado para estudar os efeitos das componentes do campo magnético interplanetário na ionosfera e na magnetosfera [72]. As informações dessa base de dados empregadas neste estudo estão indicadas na Tabela 3.

Tabela 3 – atributos relacionadas às atividades solar e interplanetária

atributo	Descrição
IMF	Magnitude média do campo magnético interplanetário
B_z	Componente z do campo magnético interplanetário
P_P	Pressão do fluxo de plasma
IEF	Campo elétrico interplanetário

Fonte: Autoria própria (2021)

O estado da ionosfera foi caracterizado mediante informações provenientes de ionossondas, representadas pela base *ION*, e receptores GPS, representadas pela base *GPS*. Ionossondas representam uma importante ferramenta tanto para operações de pesquisa quanto de monitoramento. Pois, possibilitam a aquisição de uma variedade de parâmetros ionosféricos que viabilizam a realização de inferências tanto acerca do comportamento da ionosfera como do clima espacial [73, 74]. Os dados associados à ionossonda foram obtidos por meio do portal *Space Weather Data Share* de forma análoga aqueles relativos aos magnetômetros. Nessa etapa foi necessário usar o software *SAO Explorer 3.5.3* ⁴ para processar os arquivos que continham as informações desejadas. A ionossonda responsável pela aquisição dos dados é da categoria DGS 256 [75], descrita em [76, 77, 78]. A qualidade das informações extraídas desses equipamentos pode ser medida em uma escala de 0 a

³ <<https://omniweb.gsfc.nasa.gov/index.html>>⁴ <<https://ulcar.uml.edu/SAO-X/SAO-X.html>>

100 por um parâmetro conhecido como *C-score*. Nesse estudo foram consideradas apenas os exemplos que obtiveram um *C-score* maior do que 40, critério indicado na literatura para assegurar a qualidade dos dados utilizados [79]. Os atributos utilizados nesse estudo associados à ionossonda estão listados na Tabela 4.

Tabela 4 – Atributos relacionadas ao estado da ionosfera derivadas de ionossonda - Regressão

atributo	Descrição
f_oF2	Frequência crítica da camada F2
$fminF$	Frequência mínima dos ecos da camada F
$fmin$	Frequência mínima dos ecos do ionograma
$h'F2$	Altura virtual mínima do traço F2
$h'F$	Altura virtual mínima do traço F
fxI	Frequência máxima do traço F
FF	Distribuição de frequência entre $fxF2$ e fxI
$hmF2$	Altura do pico da camada F
$yF2$	Metade da espessura da camada F2, modelo parabólico
TEC	Conteúdo Eletrônico Total
$f(h'F)$	Frequência na qual ocorre $h'F$
$f(h'F2)$	Frequência na qual ocorre $h'F2$

Fonte: Autoria própria (2021)

Dados provenientes de receptores GNSS são amplamente usados pela comunidade científica para estudos voltados à dinâmica da ionosfera e, particularmente são bastante usados nas pesquisas com foco na cintilação ionosférica [80]. Os parâmetros usados nesse estudo associados a esses dispositivos estão listados na Tabela 5.

Tabela 5 – atributos relacionadas ao estado da ionosfera derivadas de receptor GPS - Regressão

atributo	Descrição
$ROTI$	Índice ROTI
θ_E	Ângulo de elevação do satélite
θ_A	Ângulo azimute do satélite
$MA-S_4$	Valor do índice S_4 no instante atual

Fonte: Autoria própria (2021)

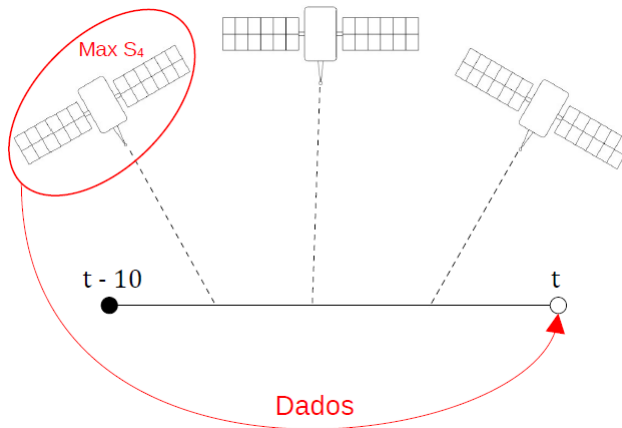
Uma particularidade desses dados é que a cada instante de tempo estão disponíveis

informações de diversos satélites. O que demanda uma etapa de processamento para selecionar o satélite cujos dados serão aproveitados. Nessa situação, optou-se por aplicar uma metodologia baseada naquelas utilizadas em [19, 81]. Esse processamento se deu em duas etapas, descritas abaixo.

1. Considerar somente os satélites cujo ângulo de elevação seja superior a 30° , ou seja, $\theta_E > 30^\circ$, essa etapa é necessária para atenuar os efeitos do multicaminho, um fenômeno que afeta negativamente os sinais captados por receptores GNSS [82].
2. Para um determinado instante de tempo t verifica-se no intervalo de 10 minutos anteriores a esse instante qual satélite, dentre aqueles selecionados na etapa anterior, apresenta o máximo valor do índice S_4 . Então selecionam-se os dados desse satélite como representativos do instante t .

Na Figura 11 encontra-se a ilustração das etapas descritas anteriormente. Realizando-se esse processamento foi possível obter uma série temporal do índice S_4 . Com o objetivo de suavizar a curva que descreve a variação desse parâmetro, aplicou-se uma operação de média móvel compreendendo 15 amostras, dando origem ao atributo $MA-S_4$ [19, 81]. Em seguida, para cada exemplo foi determinado qual seria o valor do índice S_4 após 30 minutos, dando origem ao parâmetro $S_4 - 30$. O objetivo dos modelos de regressão desenvolvidos neste trabalho é realizar a predição dessa variável.

Figura 11 – Tratamento de dados da base GNSS



Fonte: Autoria própria (2021)

A dependência temporal do fenômeno foi considerada levando-se em conta o dia do ano, a data e a hora locais. Para considerar as variações anuais e diárias da cintilação ionosférica foram empregados os parâmetros indicados na Tabela 6.

Tabela 6 – atributos relacionadas ao estado da ionosfera derivadas de receptor GPS

atributo	Descrição
CSD	Componente senoidal do dia do ano
CCD	Componente cossenoidal do dia do ano
CSH	Componente senoidal da hora do dia
CCH	Componente cossenoidal da hora do dia

Fonte: Autoria própria (2021)

Os parâmetros indicados na Tabela 6 foram obtidos por meio das expressões [83]:

$$CSD = \sin\left(\frac{2\pi t_D}{365,25}\right), \quad (4.1)$$

$$CCD = \cos\left(\frac{2\pi t_D}{365,25}\right), \quad (4.2)$$

$$CSH = \sin\left(\frac{2\pi t_H}{24}\right), \quad (4.3)$$

$$CCH = \cos\left(\frac{2\pi t_H}{24}\right). \quad (4.4)$$

Nas equações (4.1) e (4.2), o termo t_D denota o número do dia considerado, portanto, $1 \leq t_D \leq 365$. Nas equações (4.3) e (4.4), o termo t_H se refere ao número da hora considerada, logo, $0 \leq t_H \leq 23$. Tais parâmetros já foram utilizados em trabalhos anteriores, como Sivavaraprasad *et al.* [83] e Poole e McKinnell [84], na construção de modelos preditivos relacionados à ionosfera. No Apêndice B é possível observar uma descrição mais detalhada das bases de dados usadas para obter os modelos de regressão.

4.1.2 Base de dados usada para classificação

Os modelos de classificação foram desenvolvidos considerando dados referentes somente à cidade de São Luís – MA. Foram coletados exemplos distribuídos ao longo dos anos de 2014, 2015, 2016, 2017 e 2019, de modo que a base de dados resultante conta com 40125 exemplos.

A base de dados usada no desenvolvimento dos modelos de classificação foi construída da seguinte forma: as informações referentes às atividades geomagnética, solar e interplanetária, além da dependência temporal do fenômeno foram obtidas de forma análoga aos dados usados no problema de regressão. Já os dados referentes ao estado da ionosfera são ligeiramente distintos, de forma que as informações provenientes da ionossonda e do receptor GNSS apresentam as seguintes diferenças:

- Ionossonda

Foram considerados os mesmos atributos empregados na base de dados dos modelos de regressão, com exceção do *conteúdo eletrônico total (tec)* visto que essa informação foi obtida com mais detalhes por meio do receptor GNSS.

- Receptor GNSS

As informações relacionadas ao receptor GNSS foram coletadas por meio da plataforma *ISMR Query Tool*, que disponibiliza dados referentes a satélites de diversas constelações GNSS que operam em uma série de frequências [34]. Para assegurar a qualidade dos dados utilizados e delimitar o estudo realizado, as seguintes etapas de processamento foram empregadas:

- Considerar somente os satélites cujo ângulo de elevação seja superior a 30° , ou seja, $\theta_E > 30^\circ$.
- Considerar apenas os exemplos cujo parâmetro *lock_time* seja maior que 60 segundos [85].
- Considerar somente os dados de satélites da constelação GPS, o mais difundido sistema de navegação por satélite, mais afetado pela cintilação ionosférica em regiões próximas ao equador magnético [86], região onde a cidade de interesse está localizada.
- Considerar somente dados relativos à frequência L1 (1575,42 MHz).
- Corrigir o índice S_4 considerando o ruído ambiente captado pelo sensor. Tal operação foi realizada por meio da expressão [87, 88]:

$$S_4 = \begin{cases} \sqrt{S_{4T}^2 - S_{4N}^2} & \text{se } S_{4T}^2 - S_{4N}^2 > 0 \\ 0 & \text{se } S_{4T}^2 - S_{4N}^2 \leq 0 \end{cases} . \quad (4.5)$$

Na equação 4.5, as parcelas S_{4T} e S_{4N} correspondem ao valor do índice S_4 total e a correção devida ao ruído ambiente, respectivamente.

- Empregar a mesma metodologia usada na construção da base de dados do problema de regressão para obter uma série temporal do índice S_4 .

Na Tabela 7 estão listados os atributos derivados do receptor GNSS usadas na construção dos modelos de classificação.

Tabela 7 – Atributos derivados de receptor GNSS - Classificação

atributo	Descrição
tec	Contéudo Eletrônico Total no instante atual
tec_{15}	TEC a 15 segundos atrás
tec_{30}	TEC a 30 segundos atrás
tec_{45}	TEC a 45 segundos atrás
$dtec_{15tow}$	$tec_{15} - tec$
$dtec_{3015}$	$tec_{15} - tec_{30}$
$dtec_{4530}$	$tec_{30} - tec_{45}$
$dtec_{6045}$	$tec_{45} - tec$ (instante anterior)
avg_cn0_l1	Relação sinal ruído média calculada no minuto anterior
θ_E	Ângulo de elevação do satélite
θ_A	Ângulo azimuth do satélite

Fonte: Autoria própria (2021)

Os exemplos obtidos foram agrupadas em duas classes, *cintilação* e *sem cintilação*, o seguinte critério foi usado para realizar essa divisão [89]: exemplos cujo índice S_4 fosse menor ou igual a 0,2 foram designadas como *sem cintilação*, os demais foram classificados como *cintilação*. Usando o critério descrito acima foram inseridas duas informações na base de dados, $label_now$ indica se está ocorrendo cintilação no instante atual e $label_30$ indica se estará ocorrendo cintilação após 30 minutos. O objetivo dos modelos preditivos de classificação será classificar corretamente os exemplos com base no parâmetro $label_30$, a atributo $label_now$ estará entre as entradas do modelo. No Apêndice B é possível observar uma descrição mais detalhada das bases de dados usadas para obter os modelos de classificação.

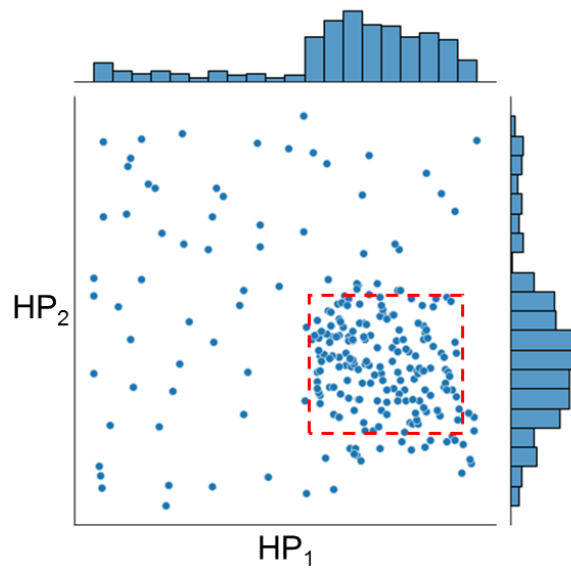
4.2 Desenvolvimento dos modelos preditivos

Nesta seção serão descritas as etapas necessárias para desenvolver os modelos preditivos de classificação e regressão. Embora pertençam à categoria diferentes, a construção de tais modelos possui alguns aspectos em comum, tais como a etapa de otimização de hiperparâmetros e a seleção de atributos. O procedimento geral associado a tais questões será indicado a seguir e os detalhes específicos de cada modelo serão tratados individualmente.

A estratégia adotada neste trabalho para otimizar os hiperparâmetros dos modelos preditivos, consistiu na combinação dos métodos de busca aleatória e de *grid*. O processo é iniciado realizando-se buscas aleatórias por um determinado número de vezes, digamos n vezes e armazenando-se os melhores modelos obtidos em cada uma dessas n iterações.

Em seguida, é possível avaliar a frequência dos hiperparâmetros selecionados nos modelos ótimos de cada iteração, possibilitando uma redução no espaço de buscas considerado, o que viabiliza o uso da busca de *grid*. O processo descrito está representado na Figura 12, onde cada ponto representa um modelo ótimo determinado por meio da busca aleatória. Note a região destacada pelo retângulo tracejado vermelho, que reúne a maioria desses modelos. Uma vez identificada essa região particular, realiza-se uma busca de *grid* na mesma para determinar os valores ótimos para os hiperparâmetros HP_1 e HP_2 .

Figura 12 – Estratégia de otimização de hiperparâmetros



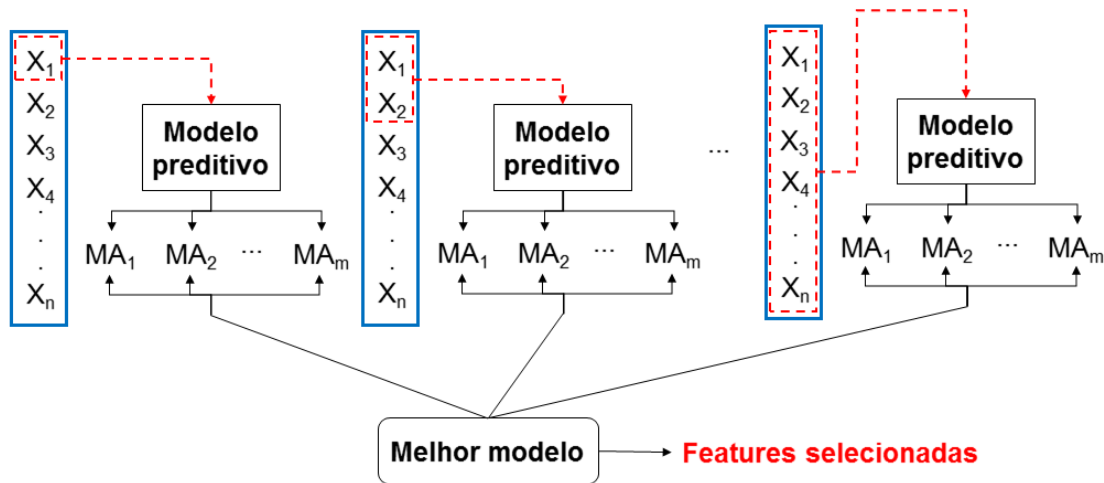
Fonte: Autoria própria (2021)

Após a otimização dos hiperparâmetros dos modelos, deu-se seguimento à fase de seleção de atributos. Essa etapa foi realizada da seguinte forma:

1. Determinar um ranking de importância dos atributos disponíveis.
2. A partir do ranking obtido, definir quantos atributos usar na construção dos modelos.

Note que, é no segundo passo da metodologia empregada que ocorre de fato a seleção de atributos. Nesse ponto a seguinte pergunta é respondida: “Quantos atributos, dentre os melhores ranqueados, otimizam o desempenho do modelo preditivo?”. Na Figura 13, é possível observar como essa etapa acontece. Os retângulos azuis representam os rankings de importância de atributos, com X_1 o mais relevante, seguido por X_2 e assim sucessivamente. A estratégia de seleção consistiu em construir diversos modelos considerando quantidades progressivas de atributos em cada um. O primeiro modelo usa somente o atributo mais importante, o segundo usa os dois mais importantes e assim por diante. Uma vez obtidos os modelos, aos desempenhos de cada um deles são determinadas por meio das métricas de desempenho MA_1 , MA_2 , etc. Avaliando uma métrica de desempenho ou um conjunto delas é possível apontar o melhor modelo, o que determina os atributos selecionados.

Figura 13 – Metodologia de seleção de atributos



Fonte: Autoria própria (2021)

Essas foram as metodologias gerais de otimização de hiperparâmetros e de seleção de atributos, dando segmento ao trabalho tais metodologias serão particularizadas para os modelos de regressão e classificação.

4.2.1 Modelos preditivos de regressão

Os algoritmos usados para o desenvolvimento dos modelos de regressão foram os seguintes:

- Regressão linear múltipla - RLM
- Random Forest - RF

O modelo baseado no algoritmo RLM não teve hiperparâmetros otimizados, devido à simplicidade do algoritmo. Os hiperparâmetros otimizados para o algoritmo RF estão dispostos na Tabela 8.

Tabela 8 – Hiperparâmetros otimizados para o modelo RF de regressão

Hiperparâmetro	Valor ótimo
<i>max_depth</i>	14
<i>min_samples_leaf</i>	5
<i>min_samples_split</i>	2
<i>n_estimators</i>	18

Fonte: Autoria própria (2021)

Na Tabela 8: o hiperparâmetro *max_depth* denota a profundidade máxima dos sub-modelos de árvore usados; *min_samples_leaf* representa a quantidade mínima de

exemplos requerida para estar em um nó folha, $min_samples_split$ indica a quantidade mínima de exemplos requerida para subdividir um nó interno; $n_estimators$ representa a quantidade de sub-modelos usados no algoritmo RF.

Na fase de seleção de atributos para os modelos de regressão, as técnicas de ranqueamento de atributos usadas foram a correlação linear, a informação mútua e o algoritmo RF. A base de dados usada compreende cerca de 60 dias, os quais foram divididos em 4 blocos de 15 dias. Os três primeiros blocos de exemplos foram usados na etapa de obtenção dos modelos preditivos, já o quarto foi reservado como conjunto de teste. Ademais, na etapa de obtenção dos modelos preditivos empregou-se a validação cruzada *3-fold* em blocos repetida 10 vezes. O que resultou em 30 modelos, a partir dos quais foram computadas as métricas de avaliação indicadas no capítulo *Referencial teórico*. A métrica considerada para eleger o melhor modelo foi o R^2 , os atributos selecionados serão apresentados no próximo capítulo.

4.2.2 Modelos preditivos de classificação

Os algoritmos usados para o desenvolvimento dos modelos de classificação foram os seguintes:

- Regressão logística - RLog
- Random Forest - RF

O modelo baseado no algoritmo RLog não teve hiperparâmetros otimizados. Os hiperparâmetros otimizados para o algoritmo RF estão dispostos na Tabela 9.

Tabela 9 – Hiperparâmetros otimizados para o modelo RF de classificação

Hiperparâmetro	Valor ótimo
max_depth	18
$min_samples_leaf$	5
$min_samples_split$	2
$n_estimators$	52

Fonte: Autoria própria (2021)

Na fase de seleção de atributos para os modelos de classificação, as técnicas de ranqueamento de atributos usadas foram a estatística F, a informação mútua e o algoritmo RF. Além disso, a base de dados foi dividida em dois subconjuntos, um para obter os modelos contendo cerca de 75% dos exemplos (30090) e outro para avaliar os modelos, com os exemplos restantes. Ademais, empregou-se a validação cruzada *5-fold*

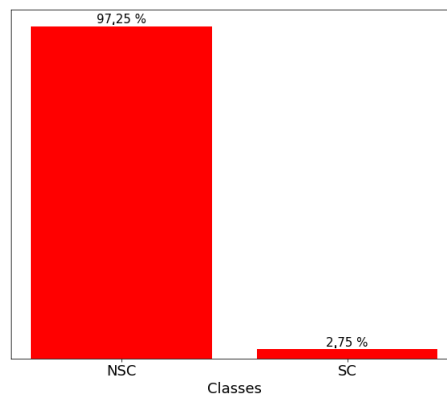
em blocos estratificada repetida 5 vezes, obtendo-se 25 modelos, a partir dos quais foram computadas as métricas acurácia, precisão, sensibilidade, medida F1 e especificidade. A métrica considerada para eleger o melhor modelo foi a medida F1. Visto que a mesma é adequada para o cenário de classes desbalanceadas [90]. Os atributos selecionados serão apresentados no próximo capítulo.

Como mencionado acima, uma característica da base de dados usada é o desbalanceamento, ou seja, a existência de uma quantidade desigual de exemplos das classes consideradas [91]. Na Figura 14, é possível perceber que existem muito mais exemplos da classe que indica a ausência de cintilação, NSC, na base de dados. Nesse contexto, os modelos de classificação tendem a desenvolver um viés na direção dos exemplos da classe majoritária. Dessa forma, a maioria dos exemplos pertencentes à classe minoritária é classificada incorretamente e, em geral, a classificação correta desses exemplos é mais relevante em uma série de aplicações práticas [91, 90]. Uma forma de lidar com o desbalanceamento de classes é penalizar mais os erros relacionados à classe minoritária durante a etapa de treinamento, isso pode ser realizado ponderando os erros cometidos. O peso atribuído ao classificar incorretamente um exemplo da classe k é dado por [92]:

$$w_k = \frac{N}{C \cdot n_k}. \quad (4.6)$$

Na equação (4.6), N denota a quantidade total de exemplos, C a quantidade de classe e n_k a quantidade de exemplos pertencentes à classe k . Essa estratégia foi escolhida, em favor de outras amplamente empregadas na literatura, com base em experimentos. Nos quais foram avaliadas tanto técnicas de sobreamostragem quanto de subamostragem, além da estratégia de ponderação citada anteriormente. Os resultados dessa etapa não indicaram diferenças significativas dentre essas abordagens. Logo, selecionou-se a técnica de ponderação por considerar-se que dessa forma os dados não sofreriam alterações significativas.

Figura 14 – Distribuição de exemplos por classe



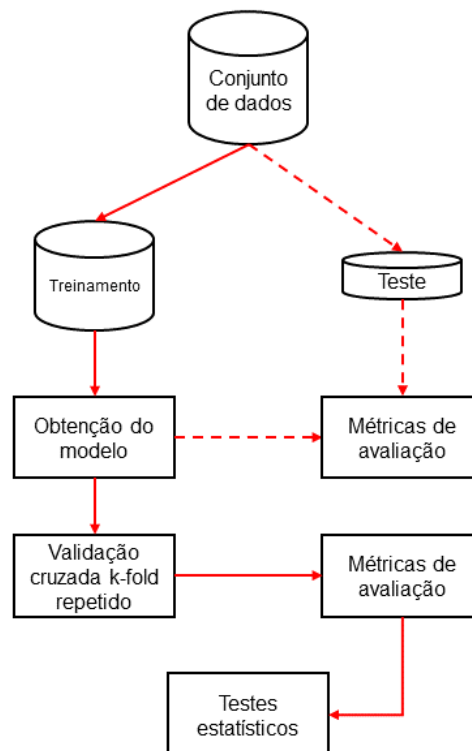
Fonte: Autoria própria (2021)

4.3 Avaliação dos modelos preditivos

O processo de avaliação dos modelos preditivos construídos está descrita na Figura 15, onde as setas contínuas indicam as operações realizadas no conjunto de treinamento e as tracejadas representam o uso do conjunto de testes. Note que os testes estatísticos são aplicados usando as métricas de avaliação obtidas por meio do conjunto de treinamento.

Para os modelos de regressão, foram avaliadas as seguintes métricas: o coeficiente de determinação, o erro quadrático médio, o erro absoluto médio e o erro percentual absoluto médio. Já para avaliar os modelos de classificação, foram empregadas as métricas: acurácia, precisão, sensibilidade, medida F1 e especificidade.

Figura 15 – Metodologia de avaliação dos modelos preditivos



Fonte: Autoria própria (2021)

Note que os testes estatísticos foram realizados por meio do teste de postos com sinal de Wilcoxon empregando a aproximação de uma distribuição normal indicada na equação (3.29). O presente trabalho foi desenvolvido usando a linguagem Python e em particular a biblioteca Scikit Learn [93].

5 Resultados e discussão

Neste capítulo serão apresentados os resultados da pesquisa, os quais serão organizados em duas seções. Uma com foco nos modelos de regressão e outra nos modelos de classificação.

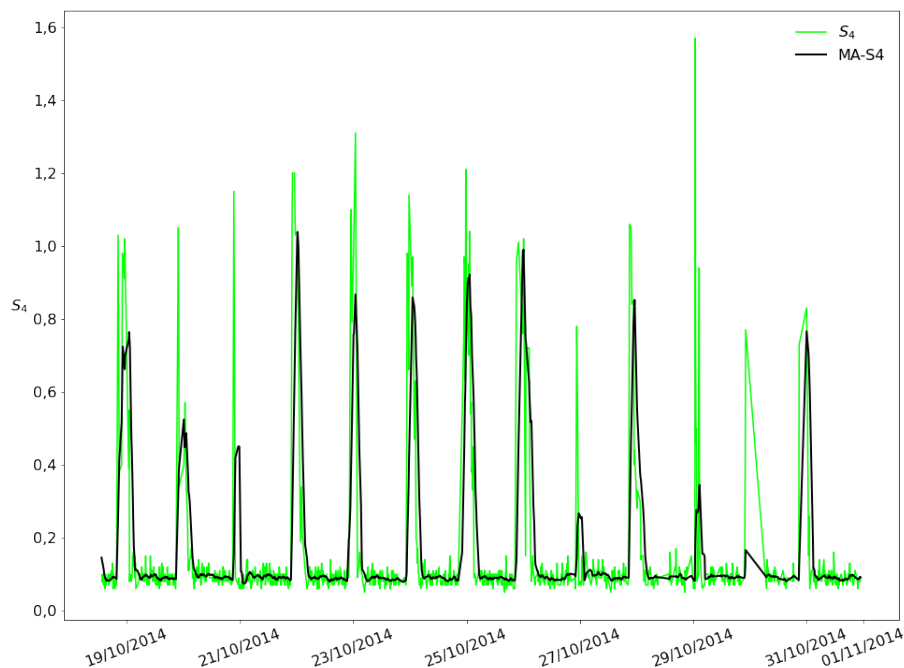
5.1 Modelos de regressão

Nesta seção serão apresentados os resultados obtidos por meio dos modelos de regressão. Inicialmente será discutida a etapa de processamento de dados da base GNSS, em seguida o foco será na seleção de atributos e por fim a avaliação dos modelos preditivos será abordada.

5.1.1 Processamento de dados da base GNSS

Na Figura 16 é possível verificar o resultado da etapa de processamento da base de dados GNSS. O gráfico verde representa a série temporal do índice S_4 obtida por meio das etapas de processamento descritas no capítulo anterior. Já a curva escura representa a operação de média móvel aplicada na curva verde.

Figura 16 – Resultados obtidos por meio do tratamento de dados da base GNSS

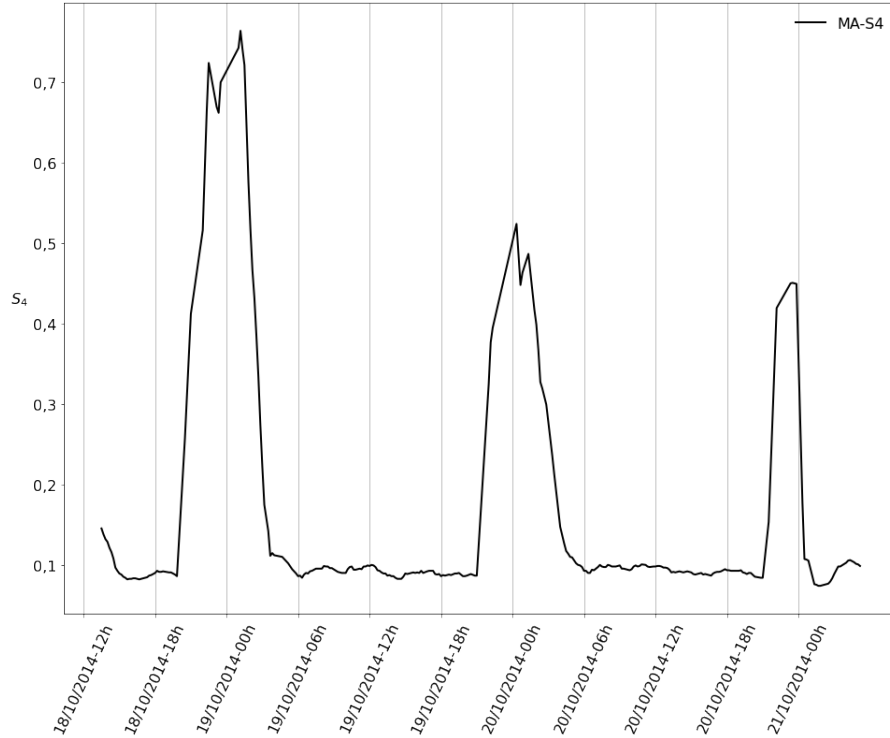


Fonte: Autoria própria (2021)

Na Figura 16 é possível constatar que a operação de média móvel foi eficiente no

que se refere a reduzir a quantidade de ruído presente nos dados. Na Figura 17 é possível verificar uma ampliação da região inicial do gráfico apresentado na Figura 16.

Figura 17 – Resultados obtidos por meio do tratamento de dados da base GNSS - ampliação



Fonte: Autoria própria (2021)

Analisando a Figura 17, nota-se um comportamento periódico na variação do índice S_4 . Tal aspecto é caracterizado por um aumento no valor do índice por volta das 18h, chegando a um valor máximo às 00h do dia seguinte e decaindo logo após. Um comportamento característico observado em sinais que operam em banda de frequências L na região da ionosfera equatorial de acordo com Basu *et al.* [13]. O que indica que as etapas de pré-processamento empregadas na base GNSS mantiveram as características inerentes ao fenômeno da cintilação ionosférica, e se mostram adequadas para a análise desenvolvida neste estudo. A discussão apresentada acima também aplica-se à etapa de processamento da base de dados GNSS empregada na construção da base de dados dos modelos de classificação.

5.1.2 Seleção de atributos

Essa etapa compreendeu inicialmente o ranqueamento e a posterior seleção de atributos. Na Tabela 10 é possível observar os resultados dos algoritmos de ranqueamento de atributos. Os atributos estão ordenados em ordem de importância da esquerda para a direita. Na Tabela 10, RR_1 indica o ranking de importância obtido por meio da correlação linear, RR_2 representa o ranking de importância derivado por meio da técnica de informação mútua e RR_3 denota o ranking de importância obtido por meio do algoritmo RF.

Tabela 10 – Rankings de importância de atributos para os modelos de regressão

Algoritmos	Ranking de atributos
RR_1	$MA-S_4 - ROTI - CCH - yF2 - CET - FF - fxI - foF2 - fminF - F_M - f(h'F2) - f(h'F) - hmF2 - \theta_E - H_M - fmin - \theta_A - h'F - D_M - h'F2 - CCD - CSD - IEF - B_z - IMF - CSH - P_P$
RR_2	$MA-S_4 - CCH - CSD - CCD - yF2 - CSH - CET - fxI - foF2 - H_M - ROTI - F_M - \theta_A - fminF - IMF - f(h'F2) - hmF2 - f(h'F) - FF - D_M - P_P - h'F2 - h'F - fmin - \theta_E - IEF - B_z$
RR_3	$MA-S_4 - ROTI - CSH - h'F2 - h'F - hmF2 - P_P - \theta_A - yF2 - IMF - F_M - fxI - H_M - IEF - \theta_E - FF - CET - CSD - CCD - B_z - D_M - foF2 - fmin - f(h'F) - f(h'F2) - fminF - CCH$

Fonte: Autoria própria (2021)

De posse dos rankings dispostos na Tabela 10, partiu-se para a seleção da quantidade ótima de atributos. Na Tabela 11 é possível consultar as quantidades ótimas de atributos resultantes das combinações entre os dois algoritmos empregados e os métodos de ranqueamento.

Tabela 11 – Quantidade ótima de atributos para cada algoritmo

Algoritmo	RR_1	RR_2	RR_3
RLM	13	11	5
RF	26	17	6

Fonte: Autoria própria (2021)

Os resultados dispostos na Tabela 11 são interpretados da seguinte forma: o número 13 localizado na intersecção da linha RLM e a coluna RR_1 indica que o conjunto ótimo de atributos para o modelo RLM, partindo do ranqueamento RR_1 , corresponde aos 13 atributos melhor ranqueados, e assim por diante para as outras entradas da Tabela 11.

Os atributos selecionados para cada modelo podem ser consultados nas Tabelas 12 e 13.

Tabela 12 – Atributos selecionados nos modelos RLM

Modelo	RR ₁	RR ₂	RR ₃
F_M	✓		
H_M		✓	
D_M			
IMF			
B_z			
P_P			
IEF			
$foF2$	✓	✓	
$fminF$	✓		
$fmin$			
$h'F2$			✓
$h'F$			✓
fxI	✓	✓	
FF	✓		
$hmF2$	✓		
$yF2$	✓	✓	
CET	✓	✓	
$f(h'F)$	✓		
$f(h'F2)$	✓		
CSD		✓	
CCD		✓	
CSH		✓	✓
CCH	✓	✓	
$ROTI$	✓	✓	✓
$MA-S_4$	✓	✓	✓
θ_A			
θ_E			

Fonte: Autoria própria (2021)

Tabela 13 – Atributos selecionados nos modelos RF

Modelo	RR ₁	RR ₂	RR ₃
F_M	✓	✓	
H_M	✓	✓	
D_M	✓		
IMF	✓	✓	
B_z	✓		
P_P			
IEF	✓		
$foF2$	✓	✓	
$fminF$	✓	✓	
$fmin$	✓		
$h'F2$	✓		✓
$h'F$	✓		✓
fxI	✓	✓	
FF	✓		
$hmF2$	✓	✓	✓
$yF2$	✓	✓	
CET	✓	✓	
$f(h'F)$	✓		
$f(h'F2)$	✓	✓	
CSD	✓	✓	
CCD	✓	✓	
CSH	✓	✓	✓
CCH	✓	✓	
$ROTI$	✓	✓	✓
$MA-S_4$	✓	✓	✓
θ_A	✓	✓	
θ_E	✓		

Fonte: Autoria própria (2021)

Analisando os resultados das Tabelas 10, 12 e 13, percebe-se que a variável $MA-S_4$ aparece no topo dos rankings fornecidos pelos algoritmos de seleção de atributos, e está presente em todos os conjuntos de variáveis selecionadas em cada modelo específico. O que reforça que modelos voltados para a predição de características da ionosfera têm seus desempenhos melhorados significativamente por meio do emprego de atributos contendo informações atrasadas no tempo, como constatado por Kutiev *et al.* [94]. Note também que, o atributo $ROTI$ foi selecionado em todos os conjuntos de atributos em relação aos dois algoritmos de aprendizagem de máquina. Isso ocorreu, possivelmente, devido à capacidade desse índice de capturar efeitos associados a irregularidades ionosféricas que ocorrem na escala dos minutos e por observar-se que o mesmo correlaciona-se bem com o índice S_4 nas fases de geração, evolução e decaimento do fenômeno da cintilação ionosférica [74, 95].

É possível notar também a presença de diversas variáveis provenientes de ionossondas nos conjuntos de atributos selecionados. O que evidencia a utilidade de tais parâmetros em estudos relacionados à cintilação ionosférica, visto que as frequências críticas da camada ionosférica, quantidades bem definidas e medidas regularmente, se constituem como alguns dos parâmetros mais difundidos para descrever o estado da camada ionosférica. Que possibilitam, por exemplo, associar a frequência $foF2$ diretamente à ocorrência de cintilação [96, 97]. Além das frequências críticas, as alturas relacionadas às diferentes regiões da ionosfera se mostram úteis para caracterizar o estado da camada, como os parâmetros $h'F$ e $hmF2$, os quais possibilitam a análise dos movimentos verticais do plasma ionosférico na camada F [98].

Outro aspecto a ser ressaltado diz respeito à seleção frequente de atributos relacionados ao tempo, reforçando a dependência temporal da cintilação ionosférica. Nota-se também que as variáveis referentes aos ventos solares não foram selecionadas com frequência. O que pode estar relacionado ao fato de que esse fenômeno tem um impacto maior na incidência de cintilação ionosférica nas regiões polares, como argumentado por Priyadarshi *et al.* [99]. O que indica que, eventualmente possa ser mais adequado buscar outros parâmetros representativos da atividade solar e do meio interplanetário para uso em modelos ionosféricos voltados à região equatorial, tais como o índice Kp e o parâmetro Dst. Destaca-se também a seleção de atributos coletados através de magnetômetros, reforçando o papel dos efeitos geomagnéticos na cintilação ionosférica, como verificado em [74].

Com respeito aos algoritmos de aprendizagem de máquina, é possível notar que as mesmas estratégias de seleção resultaram em conjuntos de atributos diferenciados. O que ressalta que diferentes conjuntos de atributos têm importâncias distintas quando se consideram modelos diferentes. Ademais, as estratégias de ranqueamento levaram a resultados distintos, algo já esperado, dado que as metodologias adotadas nessa etapa possuem diferentes formas de quantificar a relação entre os atributos e a variável de interesse. A técnica baseada na correlação linear é capaz de identificar apenas associações

de caráter linear. A abordagem baseada na Informação Mútua, teoricamente consegue identificar dependências mais complexas entre as variáveis [57, 58]. Já o método que usa o algoritmo RF, de acordo com Rogers e Gunn [56], consegue avaliar subconjuntos de atributos ao invés de considerá-los individualmente durante o processo de ranqueamento. Desse modo, é natural esperar que essas abordagens resultem em ranqueamentos diferentes.

5.1.3 Avaliação dos modelos preditivos

Os resultados obtidos por meio dos algoritmos RLM e RF estão sumarizados nas Tabelas 14 e 15, onde a legenda 'Base' indica os modelos desenvolvidas a partir de todos os 27 atributos disponíveis. Os valores apresentados estão na forma: valor médio mais ou menos o desvio padrão.

Tabela 14 – Resultados obtidos por meio do algoritmo RLM

Modelo	R^2	MSE	MAE	MAPE
Base	$0,9229 \pm 0,0266$	$0,0017 \pm 0,0007$	$0,0232 \pm 0,0054$	$0,8242 \pm 0,0643$
RR ₁	$0,9436 \pm 0,0041$	$0,0013 \pm 0,0002$	$0,0169 \pm 0,0005$	$0,8356 \pm 0,0606$
RR ₂	$0,9411 \pm 0,0007$	$0,0013 \pm 0,0001$	$0,0193 \pm 0,0014$	$0,8633 \pm 0,1004$
RR ₃	$0,9435 \pm 0,0052$	$0,0013 \pm 0,0002$	$0,0169 \pm 0,0014$	$0,8371 \pm 0,0624$

Fonte: Autoria própria (2021)

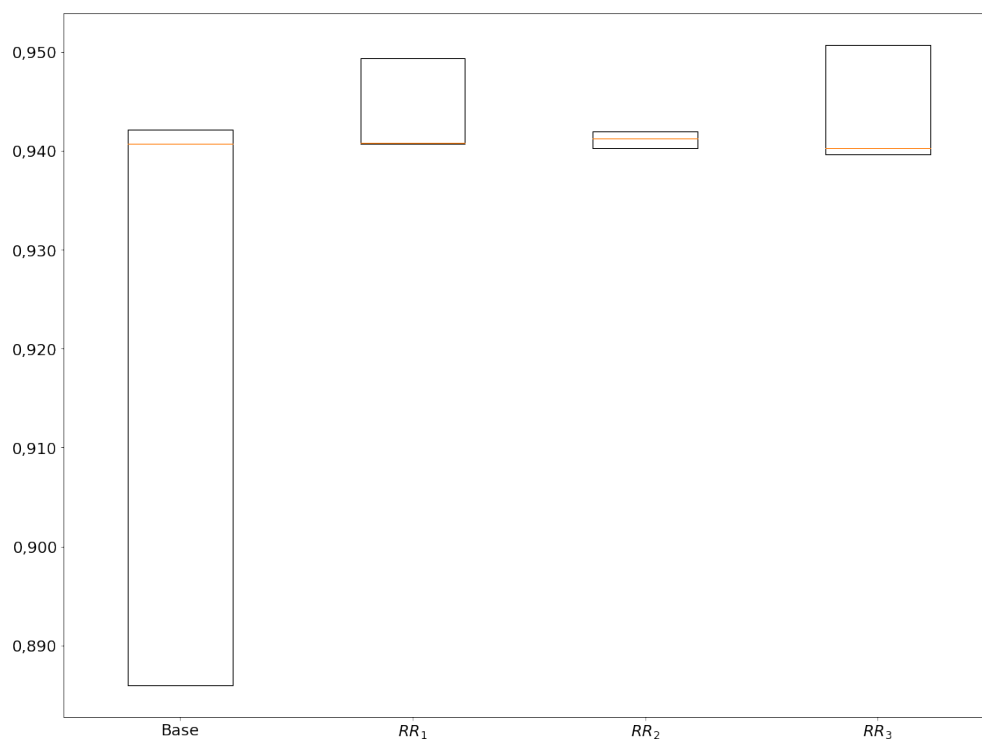
Tabela 15 – Resultados obtidos por meio do algoritmo RF - Regressão

Modelo	R^2	MSE	MAE	MAPE
Base	$0,9569 \pm 0,0119$	$0,0010 \pm 0,0003$	$0,0130 \pm 0,0022$	$0,8260 \pm 0,0476$
RR ₁	$0,9582 \pm 0,0101$	$0,0009 \pm 0,0003$	$0,0127 \pm 0,0018$	$0,8283 \pm 0,0505$
RR ₂	$0,9579 \pm 0,0096$	$0,0010 \pm 0,0003$	$0,0126 \pm 0,0017$	$0,8302 \pm 0,0526$
RR ₃	$0,9608 \pm 0,0087$	$0,0009 \pm 0,0002$	$0,0124 \pm 0,0017$	$0,8265 \pm 0,0468$

Fonte: Autoria própria (2021)

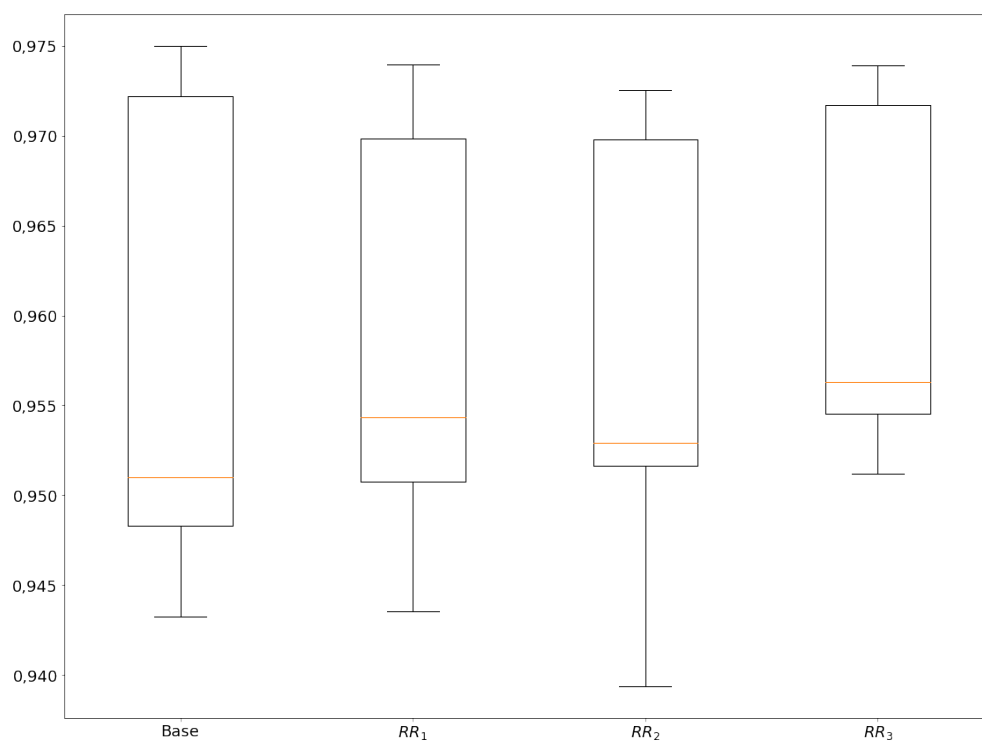
Os resultados apresentados nas Tabelas 14 e 15 podem ser representados por meio de gráficos de Box e Whisker em relação à métrica R^2 , dispostos nas Figuras 18, 19 e 20.

Figura 18 – Gráficos de Box e Whisker relacionados aos modelos RLM



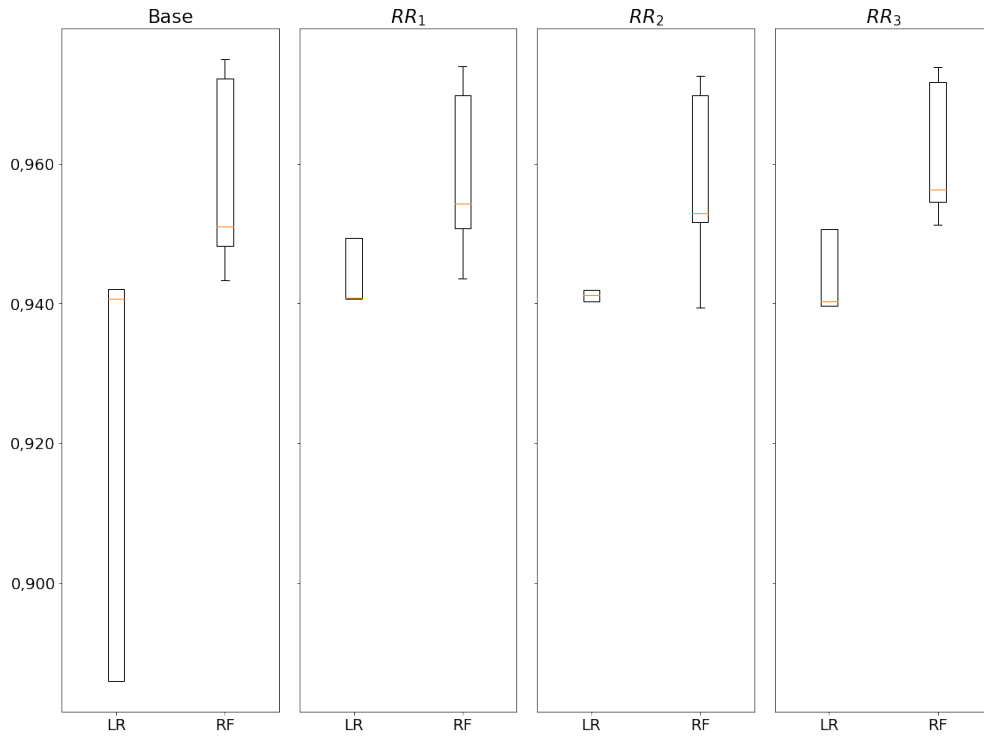
Autoria própria (2021)

Figura 19 – Gráficos de Box e Whisker relacionados aos modelos RF - Regressão



Autoria própria (2021)

Figura 20 – Gráficos de Box e Whisker contrapondo os resultados dos algoritmos RLM e RF



Fonte: Autoria própria (2021)

Na Figura 18 nota-se que os resultados do modelo *Base* apresentam uma maior dispersão em relação aos demais, além de que, exceto pelo modelo RR_1 , as distribuições das quais os dados são provenientes apresentam um caráter assimétrico. Na Figura 19, é possível notar que os resultados apresentam semelhantes graus de dispersão e mais uma vez nota-se que as distribuições dos resultados têm uma natureza assimétrica. No que diz respeito às comparações entre os algoritmos usados, apresentadas na Figura 20, os gráficos aparentam indicar que os modelos RF tiveram um desempenho superior aos modelos RLM, porém, é necessário realizar testes de significância estatística para avaliar essa hipótese de modo mais preciso.

Além de computar as métricas de avaliação dos modelos de aprendizagem de máquina, é necessário avaliar se as diferenças observadas têm significância estatística. Para isso foram realizados testes de hipóteses, com os quais foram obtidos os resultados dispostos nas Tabelas 16, 17 e 18.

Na Tabela 16 encontram-se os p -valores correspondentes ao teste que tem as seguintes hipóteses, nula H_0 , e alternativa H_1 :

H_0 : Considerando a métrica R^2 , o desempenho do algoritmo RLM é superior ao desempenho do algoritmo RF.

H_1 : Considerando a métrica R^2 , o desempenho do algoritmo RF é superior ao desempenho do algoritmo RLM.

Tabela 16 – p -valores referentes aos testes que comparam os desempenhos dos algoritmos RLM e RF

	Base	RR ₁	RR ₂	RR ₃
p -valor	$8,67 \times 10^{-7}$	$8,67 \times 10^{-7}$	$9,60 \times 10^{-7}$	$8,67 \times 10^{-7}$
Estatística	465	465	464	465

Fonte: Autoria própria (2021)

Nas Tabelas 17 e 18 estão listados os p -valores dos testes que avaliam se há diferenças significativas nos desempenhos dos modelos obtidos. As hipóteses nula e alternativa têm a seguinte forma:

H_0 : Considerando a métrica R^2 , os desempenhos dos modelos comparados são equivalentes.

H_1 : Considerando a métrica R^2 , os desempenhos dos modelos comparados não são equivalentes.

As Tabelas 17 e 18 são estruturadas de modo que, o valor situado no cruzamento de uma linha e uma coluna representa a comparação dos modelos correspondentes àquela linha e coluna.

Tabela 17 – p -valores referentes aos testes que avaliam se há diferenças nos desempenhos dos modelos RLM em relação aos diferentes métodos de seleção de atributos

	Base	RR ₁	RR ₂	RR ₃
Base		$1,26 \times 10^{-6}$	$2,16 \times 10^{-4}$	$2,16 \times 10^{-4}$
SF ₁			$2,16 \times 10^{-4}$	0,63
SF ₂				0,63

Fonte: Autoria própria (2021)

Tabela 18 – p -valores referentes aos testes que avaliam se há diferenças nos desempenhos dos modelos RF em relação aos diferentes métodos de seleção de atributos

	Base	RR ₁	RR ₂	RR ₃
Base		0,12	0,26	$2,61 \times 10^{-4}$
SF ₁			0,43	$8,31 \times 10^{-4}$
SF ₂				$4,07 \times 10^{-5}$

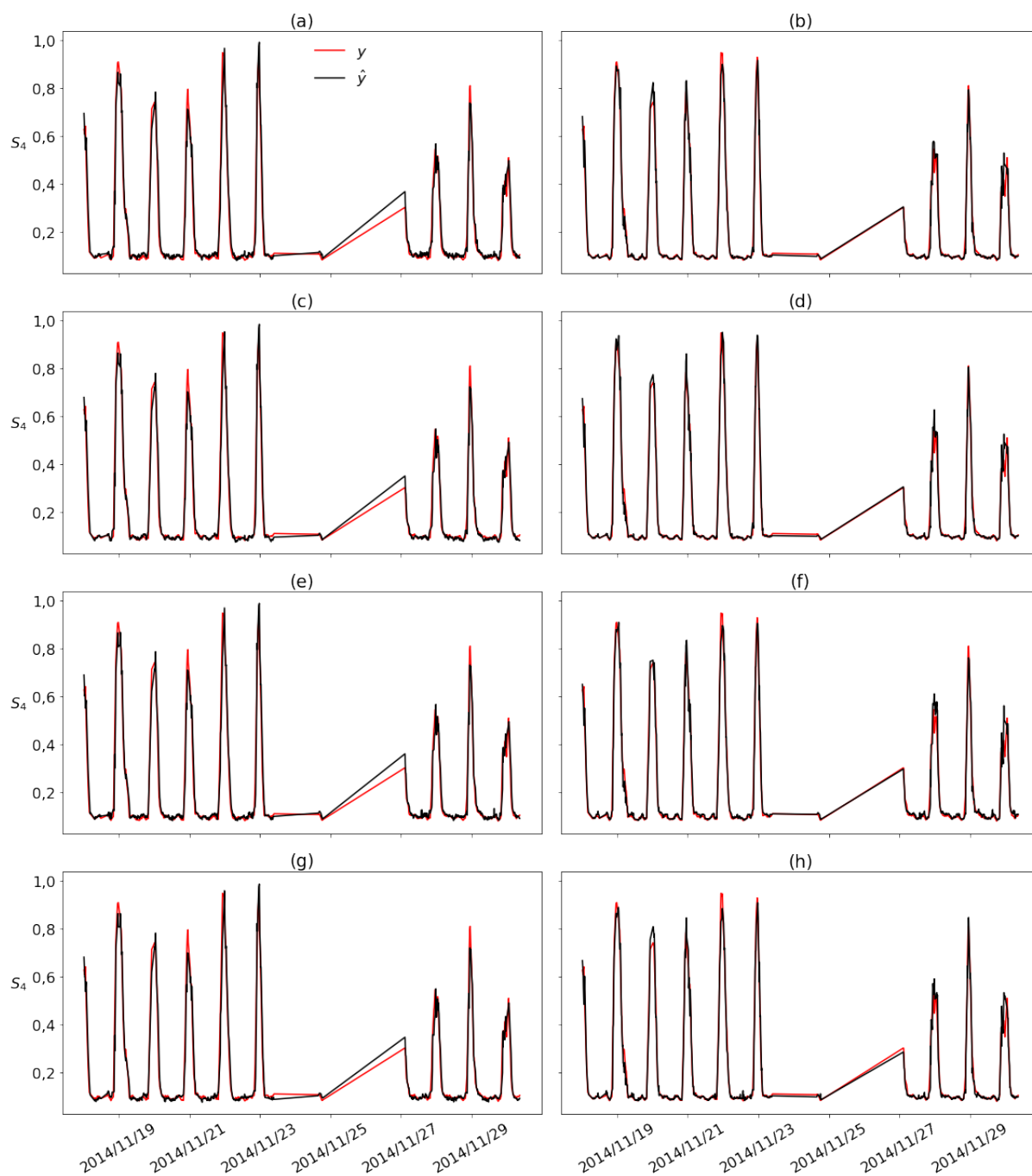
Fonte: Autoria própria (2021)

De acordo com os resultados presentes na Tabela 16, conclui-se que os dados utilizados contém evidência significativa de que o desempenho do algoritmo RF foi superior

aquele do modelo RLM. O que pode estar relacionado à maior complexidade e capacidade de tratar problemas de natureza não linear apresentada pelo algoritmo RF. Além disso, na Tabela 17, apenas na comparação entre os métodos RR_1 , RR_2 e RR_3 nota-se que os dados disponíveis não fornecem evidência suficiente para rejeitar a hipótese nula. Tal fato também está presente na Tabela 18, onde os dados considerados não fornecem evidência suficiente para a rejeição da hipótese nula de igualdade entre os modelos, exceto quando o método RR_3 está envolvido.

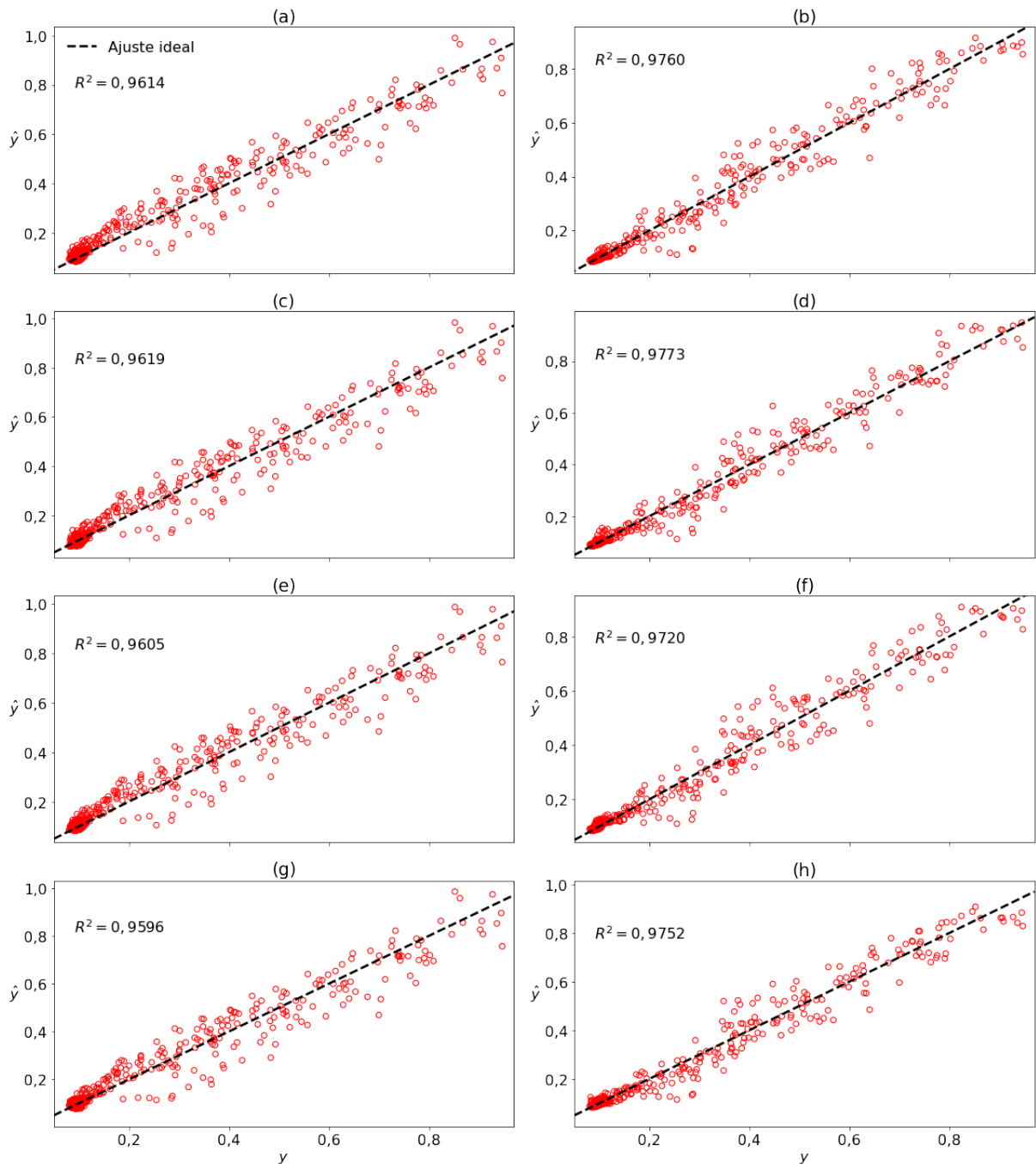
Nas Figuras 21 e 22 é possível ver os resultados obtidos por cada modelo no conjunto de testes. Os gráficos da esquerda são referentes aos modelos construídos usando o algoritmo RLM, já os da direita dizem respeito aos modelos obtidos por meio do algoritmo RF. Os gráficos da primeira linha são referentes aos modelos nos quais foram usadas todas as atributos disponíveis. Na segunda linha encontram-se os modelos construídos a partir dos atributos indicados pelo ranking RR_1 , na terceira aqueles derivados dos atributos determinados a partir do ranking RR_2 e assim por diante.

Figura 21 – Resultados obtidos no conjunto de testes



Fonte: Autoria própria (2021)

Figura 22 – Qualidade de ajustes aos dados do conjunto de testes



Fonte: Autoria própria (2021)

Analisando as Figuras 21 e 22, nota-se que visualmente não é possível identificar um modelo superior aos demais, com todos os resultados mostrando-se satisfatórios. Dessa forma, é possível ter uma maior flexibilidade quanto a escolha de qual modelo utilizar, considerando-se os atributos utilizados em cada um. Os coeficientes associados aos modelos RLM podem ser consultados no Apêndice A.

Quanto as estratégias de seleção de atributos, não foi possível indicar a melhor

estratégia devido à semelhança nos resultados obtidos, possivelmente isso é devido à quantidade reduzida de exemplos na base de dados. O que pode estar levando a algum grau de *overfitting* nos resultados analisados. Além disso, é importante ressaltar as limitações das metodologias empregadas, relacionadas principalmente ao fato das técnicas de ranqueamento não considerarem as relações internas entre os atributos, o que pode levar ao uso de atributos redundantes, que impactam negativamente o desempenho dos modelos, desse modo, é importante considerar outras estratégias de seleção.

5.2 Modelos de classificação

Nesta seção serão apresentados os resultados obtidos por meio dos modelos de classificação. Inicialmente será discutida a etapa de seleção de atributos e em seguida a avaliação dos modelos preditivos será abordada.

5.2.1 Seleção de atributos

Na Tabela 19, é possível observar os rankings de importância de atributos referentes aos modelos de classificação. Na Tabela 19, RC_1 indica o ranking de importância obtido por meio da métrica de Fisher, RC_2 indica o ranking de importância obtido por meio da técnica de informação mútua e RC_3 indica o ranking de importância obtido por meio do algoritmo RF.

Tabela 19 – Rankings de importância das atributos para a classificação

	Ranking de atributos
RC_1	$label_now - FF - h'F - CCH - yF2 - CSH - hmF2 -$ $fminF - CCD - fxI - D_M - h'F2 - foF2 - f(h'F2) -$ $\theta_E - fmin - f(h'F) - avg_cn0_l1 - CSD - dtec_3015 -$ $dtec_6045 - dtec_15tow - H_M - IMF - F_M - dtec_4530 -$ $\theta_A - P_P - IEF - B_z - tec - tec_45 - tec_30 - tec_15$
RC_2	$label_now - dtec_4530 - dtec_15tow - dtec_3015 -$ $dtec_6045 - CCH - FF - h'F - foF2 - F_M - CCD -$ $CSD - yF2 - H_M - D_M - fminF - CSH - fxI - f(h'F2) -$ $tec - h'F2 - f(h'F) - fmin - \theta_A - hmF2 - tec_30 - tec_15 -$ $tec_45 - \theta_E - B_z - IMF - P_P - IEF - avg_cn0_l1$
RC_3	$label_now - dtec_3015 - dtec_15tow - dtec_6045 -$ $dtec_4530 - h'F - yF2 - hmF2 - fxI - h'F2 - foF2 -$ $FF - H_M - tec_30 - CCH - tec - CCD - F_M - D_M -$ $tec_45 - tec_15 - avg_cn0_l1 - \theta_A - IMF - CSD -$ $P_P - CSH - IEF - f(h'F) - fminF - fmin - \theta_E - B_z -$ $f(h'F2)$

Fonte: Autoria própria (2021)

Ao observar os rankings dispostos na Tabela 19, percebe-se que as atributos d_tec são mais relevantes que os próprios parâmetros tec para identificar a ocorrência de cintilação. Esse resultado pode estar relacionado ao fato de que os atributos d_tec indicam a presença de gradientes no conteúdo eletrônico total da ionosfera que, em geral, relacionam-se diretamente à ocorrência de cintilação [100]. Em relação aos demais atributos, percebem-se resultados semelhantes aqueles observados nos modelos de regressão. O parâmetro $label_now$, que indica a ocorrência de cintilação no instante atual, figura como o mais relevante para os modelos preditivos, os atributos relacionados aos ventos solares não se

destacaram e aqueles obtidos por meio da ionossonda e do magnetômetro mostraram-se relevantes. Além disso, os parâmetros que modelam a influência temporal da cintilação também mostraram-se interessantes.

De posse dos rankings indicados na Tabela 19, a quantidade ótima de atributos para cada algoritmo foi determinada usando a metodologia descrita anteriormente e está indicada na Tabela 20.

Tabela 20 – Quantidade ótima de atributos para os modelos de classificação

Algoritmo	RC ₁	RC ₂	RC ₃
RLog	17	19	10
RF	11	18	18

Fonte: Autoria própria (2021)

5.2.2 Avaliação dos modelos preditivos

O desempenho dos modelos obtidos usando as quantidades de atributos indicadas na Tabela 20, foi avaliado por meio das métricas indicadas anteriormente usando o conjunto de treinamento e está indicado nas Tabelas 21 e 22. Novamente, "Base" indica o modelo construído a partir de todos os atributos disponíveis e RC_{*i*} denota o modelo obtido usando a quantidade de atributos correspondente ao ranking *i*. A acurácia será indicada por ACC, a precisão por PRE, a sensibilidade por SEN, a medida F1 por F1 e a especificidade por ESP.

Tabela 21 – Resultados obtidos por meio do algoritmo RLog

Modelo	ACC	PRE	SEN	F1	ESP
Base	0,9159 ± 0,0003	0,6125 ± 0,0006	0,9039 ± 0,0021	0,6592 ± 0,0008	0,9165 ± 0,0003
RC ₁	0,9147 ± 0,0005	0,6119 ± 0,0008	0,9065 ± 0,0018	0,6583 ± 0,0012	0,9152 ± 0,0005
RC ₂	0,9115 ± 0,0003	0,6085 ± 0,0005	0,9043 ± 0,0025	0,6529 ± 0,0008	0,9119 ± 0,0004
RC ₃	0,9184 ± 0,0003	0,6129 ± 0,0004	0,8915 ± 0,0026	0,6597 ± 0,0006	0,9199 ± 0,0005

Fonte: Autoria própria (2021)

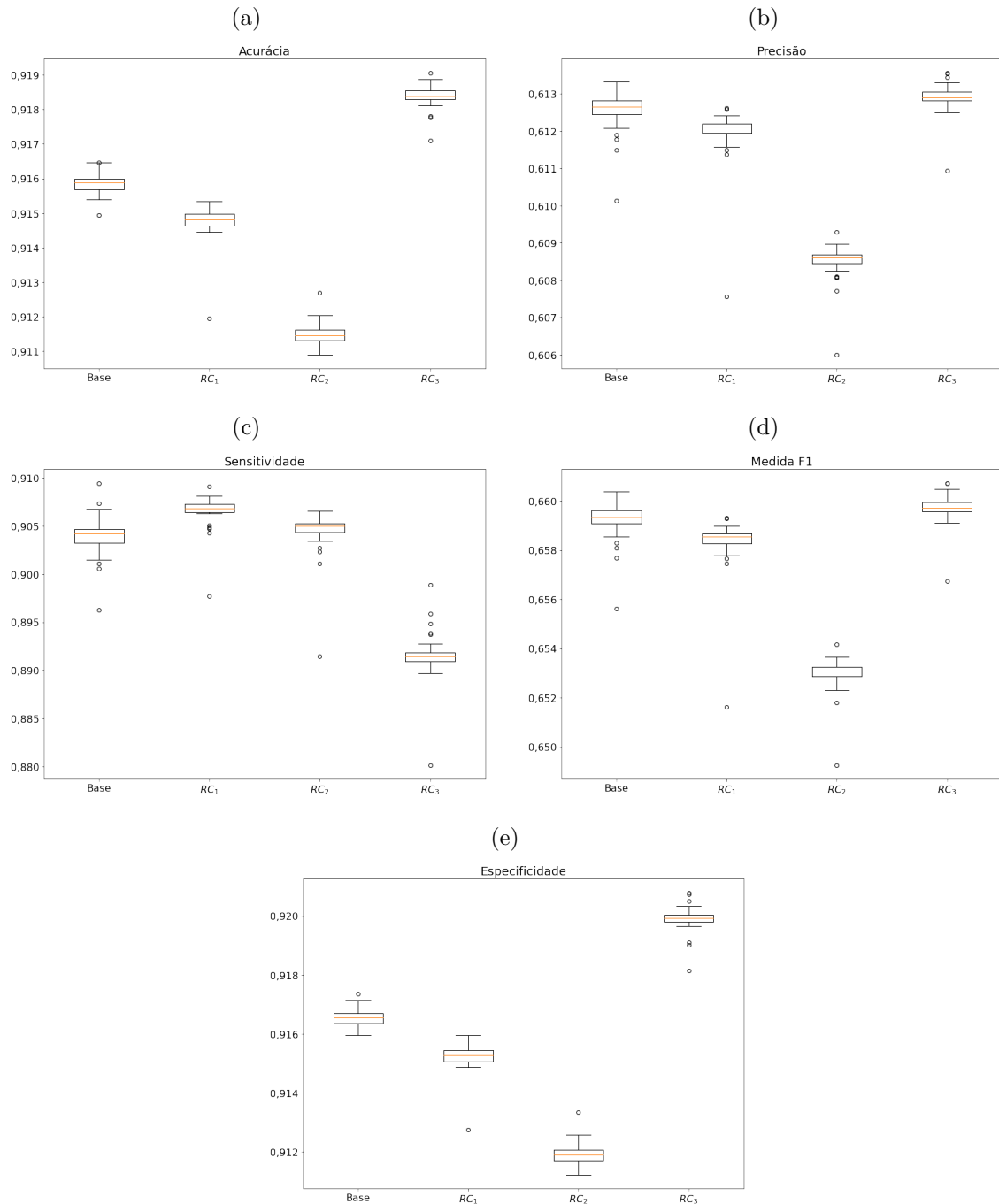
Tabela 22 – Resultados obtidos por meio do algoritmo RF

Modelo	ACC	PRE	SEN	F1	ESP
Base	0,9764 ± 0,0002	0,7713 ± 0,0016	0,8211 ± 0,0051	0,7937 ± 0,0024	0,9852 ± 0,0003
RC ₁	0,9744 ± 0,0005	0,7557 ± 0,0033	0,8420 ± 0,0041	0,7918 ± 0,0036	0,9820 ± 0,0003
RC ₂	0,9763 ± 0,0002	0,7705 ± 0,0016	0,8292 ± 0,0037	0,7965 ± 0,0020	0,9847 ± 0,0002
RC ₃	0,9754 ± 0,0004	0,7635 ± 0,0032	0,8241 ± 0,0033	0,7901 ± 0,0020	0,9840 ± 0,0005

Fonte: Autoria própria (2021)

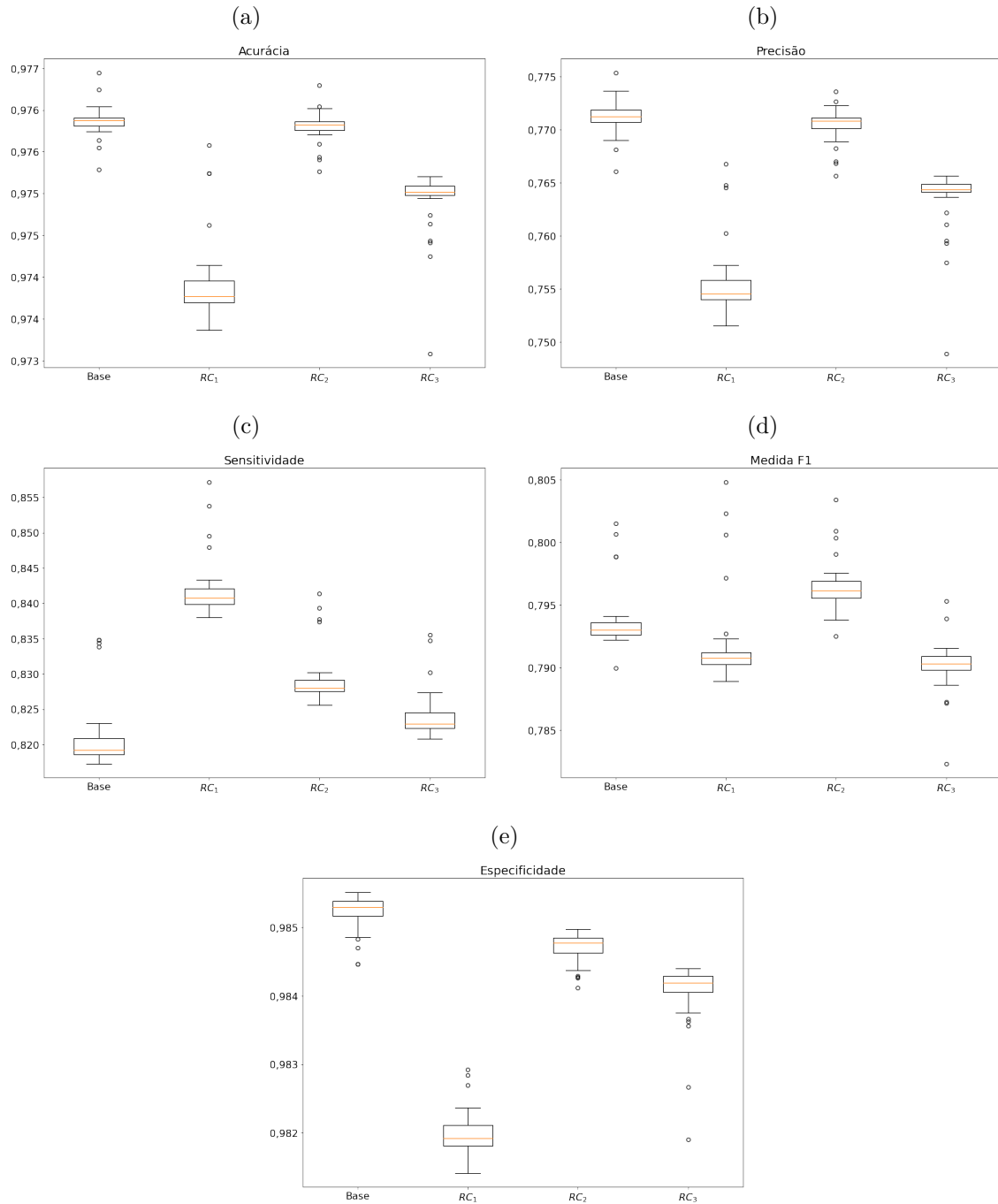
Os resultados dispostos nas Tabelas 21 e 22 podem ser observados graficamente nas Figuras 23 e 24, respectivamente, onde se percebe não haver uma diferença visual clara entre os resultados das diferentes métricas.

Figura 23 – Box plots referentes aos resultados da Tabela 21



Fonte: Autoria própria (2021)

Figura 24 – Box plots referentes aos resultados da Tabela 22



Fonte: Autoria própria (2021)

Além de computar as métricas de avaliação, os resultados dos testes de significância realizados estão dispostos a seguir. Por conta do desbalanceamento observado no conjunto de dados, a métrica de desempenho considerada nas comparações entre modelos será a medida F1, pois a mesma se mostra adequada em cenários de desbalanceamento de classes [90]. O primeiro teste realizado, referido como TH_1 , avaliou somente o desempenho dos modelos baseados no algoritmo Rlog. Foram consideradas as seguintes hipóteses:

H_0 : Considerando a medida F1, o desempenho dos modelos avaliados é equivalente.

H_a : Considerando a medida F1, o desempenho dos modelos avaliados não é equivalente.

Os p -valores correspondentes ao teste TH₁ estão indicados na Tabela 23.

Tabela 23 – p -valores referentes ao teste TH₁

	Base	RC ₁	RC ₂	RC ₃
Base		$2,70 \times 10^{-7}$	$2,48 \times 10^{-7}$	$4,53 \times 10^{-5}$
RC ₁			$2,48 \times 10^{-7}$	$2,48 \times 10^{-7}$
RC ₂				$2,48 \times 10^{-7}$

Fonte: Autoria própria (2021)

Avaliando os p -valores dispostos na Tabela 23, conclui-se que os dados utilizados fornecem evidência suficiente para rejeitar a hipótese H_o , ou seja, o desempenho dos modelos considerados não é equivalente. Dessa forma, foi realizado o teste TH₂, com as seguintes hipóteses:

H_o : Considerando a medida F1, o desempenho do modelo x é inferior aquele do modelo y .

H_a : Considerando a medida F1, o desempenho do modelo x é superior aquele do modelo y .

Os p -valores associados ao teste TH₂ estão indicados na Tabela 24, onde os modelos x estão dispostos nas linhas e os modelos y nas colunas.

Tabela 24 – p -valores referentes ao teste TH₂

	Base	RC ₁	RC ₂	RC ₃
Base		$1,35 \times 10^{-7}$	$1,24 \times 10^{-7}$	0,99
RC ₁			$1,24 \times 10^{-7}$	0,99
RC ₂				0,99

Fonte: Autoria própria (2021)

Os resultados dispostos na Tabela 24 indicam que os modelos obtidos usando o conjunto de atributos derivado do algoritmo RF foram superiores aos demais.

Agora serão apresentados os resultados dos testes de significância usados para avaliar o desempenho dos modelos baseados no algoritmo RF. Foram consideradas as seguintes hipóteses no teste TH₃:

H_o : Considerando a medida F1, o desempenho dos modelos random forest avaliados é equivalente.

H_a : Considerando a medida F1, o desempenho dos modelos random forest avaliados não é equivalente.

Os p -valores correspondentes ao teste TH₃ estão indicados na Tabela 25.

Tabela 25 – p -valores referentes ao teste TH_3

	Base	RC ₁	RC ₂	RC ₃
Base		$1,42 \times 10^{-5}$	$3,21 \times 10^{-7}$	$2,48 \times 10^{-7}$
RC ₁			$5,86 \times 10^{-7}$	0,02
RC ₂				$2,48 \times 10^{-7}$

Fonte: Autoria própria (2021)

Por meio dos resultados indicados na Tabela 25, conclui-se que os dados utilizados fornecem evidência suficiente para rejeitar a hipótese nula de equivalência entre os modelos. Dessa forma, foi realizado o teste TH_4 , com as seguintes hipóteses:

H_o : Considerando a medida F1, o desempenho do modelo random forest x é inferior aquele do modelo y .

H_a : Considerando a medida F1, o desempenho do modelo random forest x é superior aquele do modelo y .

Os p -valores associados ao teste TH_4 estão indicados na Tabela 26, onde os modelos x estão dispostos nas linhas e os modelos y nas colunas.

Tabela 26 – p -valores referentes ao teste TH_4

	Base	RC ₁	RC ₂	RC ₃
Base		$7,11 \times 10^{-6}$	0,99	$1,24 \times 10^{-7}$
RC ₁			0,99	0,001
RC ₂				$1,24 \times 10^{-7}$

Fonte: Autoria própria (2021)

Analisando a Tabela 26, nota-se que os dados considerados não foram capazes de fornecer evidência suficiente para rejeitar a hipótese nula em duas oportunidades. Por fim, foi realizado o teste TH_5 , comparando os modelos Rlog e RF, as hipóteses do teste TH_5 foram:

H_o : Considerando a medida F1, o desempenho do modelo RF é inferior aquele do modelo RLog.

H_a : Considerando a medida F1, o desempenho do modelo RF é superior aquele do modelo RLog.

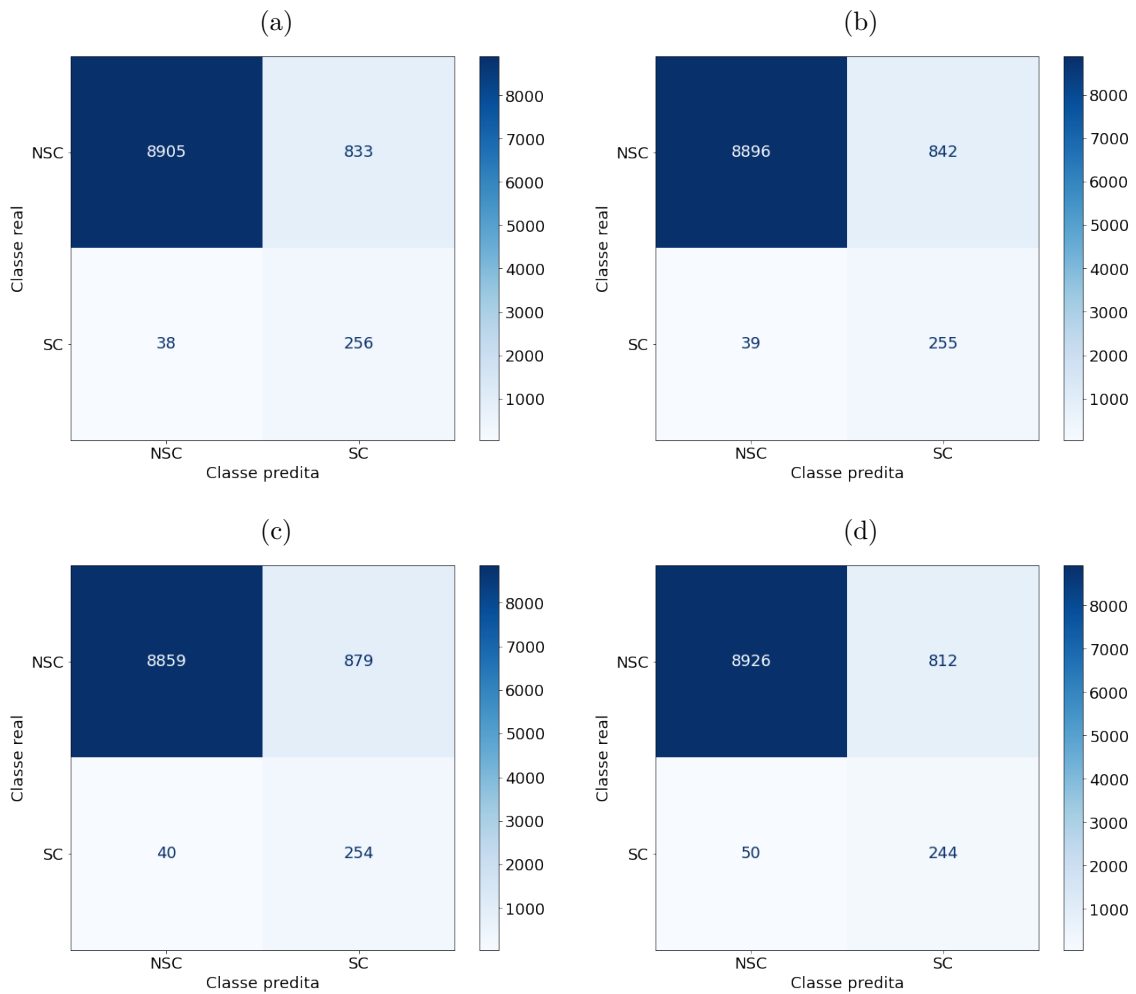
Tabela 27 – p -valores referentes ao teste TH_5

	Base	RC_1	RC_2	RC_3
p -valor	$1,24 \times 10^{-7}$	$1,24 \times 10^{-7}$	$1,24 \times 10^{-7}$	$1,24 \times 10^{-7}$

Fonte: Autoria própria (2021)

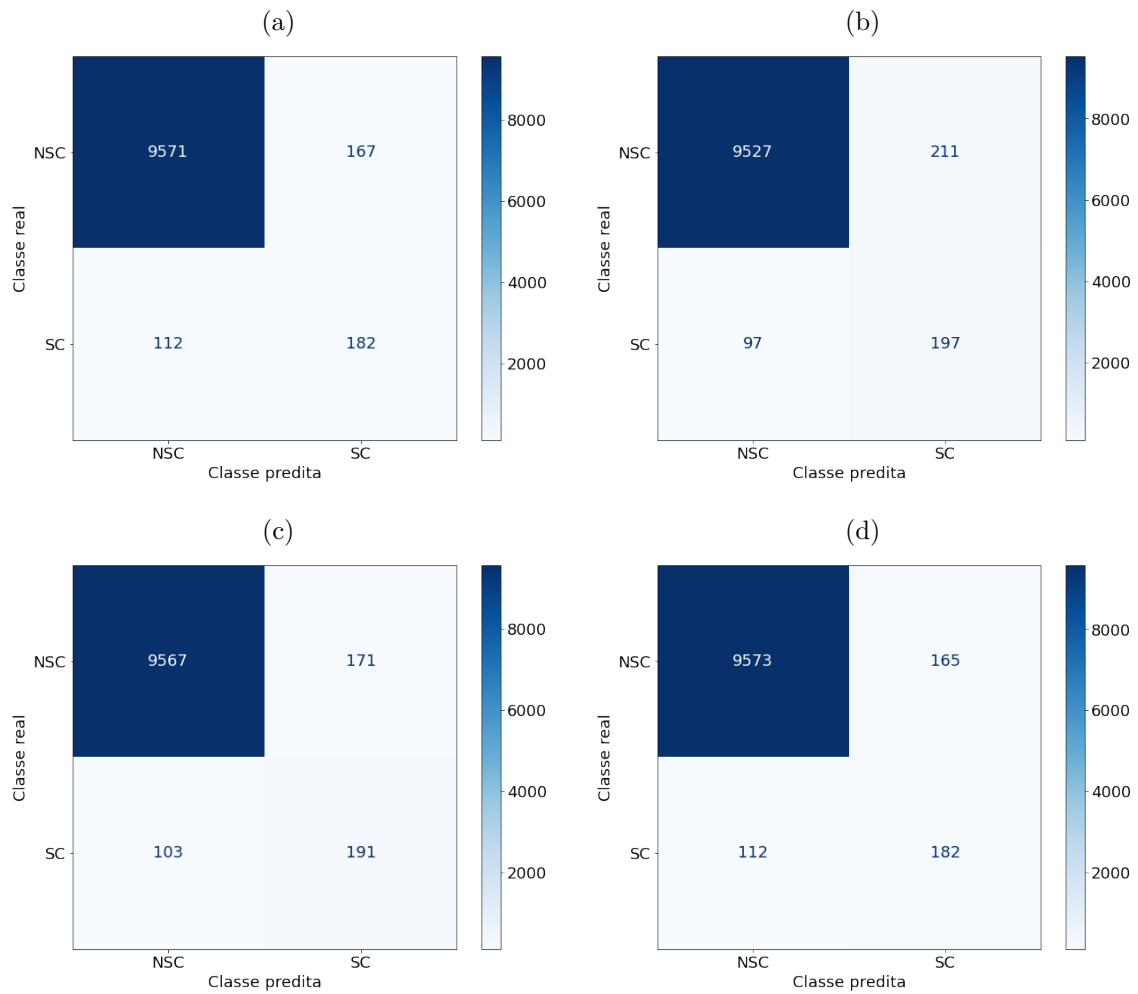
Os resultados dispostos na Tabela 27 indicam que há evidência suficiente para rejeitar a hipótese nula, ou seja, os modelos baseados no algoritmo RF mostraram-se superiores aqueles obtidos usando a RLog, quando se considera a medida F1. Esse comportamento pode estar relacionado à maior capacidade do algoritmo RF em modelar as não linearidades intrínsecas do fenômeno estudado. Os desempenhos dos modelos quando alimentados com os dados de teste podem ser avaliados nas Figuras 25 e 26, onde é possível observar as matrizes de confusão resultantes.

Figura 25 – Matrizes de confusão dos modelos RLog



Fonte: Autoria própria (2021)

Figura 26 – Matrizes de confusão dos modelos RF



Fonte: Autoria própria (2021)

As métricas de desempenho calculadas a partir dos resultados do conjunto de testes estão dispostas nas Tabelas 28 e 29.

Tabela 28 – Resultados obtidos por meio do algoritmo RLog no conjunto de testes

Modelo	ACC	PRE	SEN	F1	ESP
Base	0,9132	0,6154	0,8926	0,6618	0,9144
RC ₁	0,9122	0,6140	0,8904	0,6597	0,9135
RC ₂	0,9084	0,6098	0,8868	0,6533	0,9097
RC ₃	0,9141	0,6127	0,8733	0,6577	0,9166

Fonte: Autoria própria (2021)

Tabela 29 – Resultados obtidos por meio do algoritmo RF no conjunto de testes

Modelo	ACC	PRE	SEN	F1	ESP
Base	0,9722	0,7550	0,8009	0,7759	0,9828
RC ₁	0,9693	0,7364	0,8242	0,7727	0,9783
RC ₂	0,9727	0,7585	0,8160	0,7841	0,9824
RC ₃	0,9724	0,7565	0,8010	0,7768	0,9830

Fonte: Autoria própria (2021)

Analisando-se as Figuras 25 e 26 e as Tabelas 28 e 29, é possível notar que os modelos RLog apresentaram níveis de sensibilidade superiores aos modelos RF. Como essa métrica indica a proporção de exemplos pertencentes à classe SC classificadas corretamente, percebe-se que os modelos RLog são mais adequados para identificar exemplos que caracterizam a cintilação, no entanto, em uma perspectiva geral os modelos RF são superiores. Visto que nas demais métricas consideradas tais modelos obtiveram uma melhor performance em relação aos modelos RLog, essa característica pode estar relacionada à capacidade dos modelos RF de modelar relações não lineares presentes no conjunto de dados.

6 Conclusão

Os principais resultados deste estudo compreendem o desempenho satisfatório dos modelos preditivos desenvolvidos. O que evidencia ainda mais o potencial do uso de técnicas de aprendizagem de máquina em estudos relacionados à cintilação ionosférica. Além disso, foi possível realizar uma boa análise a respeito dos atributos usados, o que permitiu avaliar a importância de informações referentes às atividades geomagnética, solar e interplanetária, às dependências temporal e geográfica do fenômeno e ao estado da ionosfera no desempenho de modelos preditivos voltados à cintilação ionosférica.

As principais contribuições deste trabalho relacionam-se à indicação de atributos diferentes daquelas usualmente empregadas para desenvolver modelos preditivos com foco na cintilação ionosférica. Destaca-se o bom desempenho das informações coletadas a partir de magnetômetros da rede MagNet, parâmetros que ainda não haviam sido explorados desse modo. Além disso, as informações referentes aos receptores GNSS e às ionossondas foram explorados em maior profundidade. Esses aspectos podem ser úteis para direcionar o desenvolvimento de modelos futuros ainda melhores. Ademais, os modelos desenvolvidos podem ser úteis ao Centro Espacial de Alcântara, dado que atividades de lançamento podem ser afetadas pela cintilação ionosférica.

As limitações deste trabalho referem-se à existência de lacunas nas bases de dados usadas, devidas principalmente à existência de períodos nos quais alguma das fontes de dados usadas não tinham informações disponíveis. Além disso, os modelos desenvolvidos consideraram apenas duas cidades e um intervalo de tempo reduzido. Tais aspectos podem ser explorados em pesquisas futuras, nas quais é possível analisar: a incorporação de informações de outras bases de dados, voltadas para o estudo da ionosfera e do clima espacial, nos modelos preditivos; inserir informações coletadas por outras redes de sensores na base de dados nos períodos com ausência de exemplos; expandir os modelos desenvolvidos neste trabalho para outras cidades brasileiras; explorar ainda mais a dependência geográfica da cintilação por meio da consideração de informações coletadas em diferentes estações; a aplicação de técnicas de processamento de sinais para reduzir o ruído observado em alguns parâmetros da base de dados; considerar intervalos de coleta de dados maiores, de modo a desenvolver modelos mais abrangentes; usar dados satelitais de constelações voltadas para o estudo da ionosfera, como os satélites SWARM, para desenvolver modelos preditivos acerca da cintilação; considerar diferentes algoritmos de aprendizagem de máquina e até empregar estratégias de meta-aprendizagem para selecionar modelos melhores; considerar outras estratégias de seleção de atributos que explorem o relacionamento entre os próprios atributos, de modo a evitar uso de informação redundante na base de dados; desenvolver modelos que possam diferenciar as ocorrências de multicaminho e cintilação.

No geral, os objetivos propostos inicialmente foram alcançados e foi possível, além do desenvolvimento dos modelos preditivos em si, obter direcionamentos para o melhoramento de tais modelos, o que certamente contribuirá com o desenvolvimento científico futuro.

Referências

- 1 RAZANI, M. *Commercial Space Technologies and Applications: Communication, Remote Sensing, GPS, and Meteorological Satellites*. EUA: CRC Press, 2018. Citado na página 15.
- 2 SIA. *Summary of the 2021 State of the Satellite Industry Report*. 2021. Disponível: <<https://sia.org/news-resources/state-of-the-satellite-industry-report/>>. Acesso: 31/08/2020. Citado na página 15.
- 3 BOUSQUET, M. Satellite communications and space telecommunication frequencies. In: PELTON, J. N.; MADRY, S.; CAMACHO-LARA, S. (Ed.). *Handbook of Satellite Applications*. Cham: Springer International Publishing, 2017. p. 325–357. ISBN 978-3-319-23386-4. Disponível em: <https://doi.org/10.1007/978-3-319-23386-4_13>. Citado 2 vezes nas páginas 15 e 16.
- 4 ARBESSER-RASTBURG, B. Propagation effects on satellite communications and navigation systems. In: *International Conference on Applied Electromagnetics and Communications (ICECOM)*. Croácia: IEEE, 2016. p. 1–4. Citado na página 15.
- 5 IPPOLITO, L. J. Scintillation, bandwidth coherence, and other propagation factors. In: *Radiowave Propagation in Satellite Communications*. Dordrecht: Springer, 1986. p. 139–157. Citado na página 16.
- 6 ACHARYA, R. Ionosphere and ionospheric impairments. In: *Satellite Signal Propagation, Impairments and Mitigation*. Reino Unido: Elsevier, 2017. p. 87–125. Citado 2 vezes nas páginas 16 e 23.
- 7 ABDELGWAD, A. H.; SAID, T. M. Design of ground penetrating radar antenna for detecting soil contamination at l-band frequencies. *Journal of Microwaves, Optoelectronics and Electromagnetic Applications*, SciELO Brasil, v. 16, p. 853–866, 2017. Citado na página 16.
- 8 PELTON, J. N.; CAMACHO-LARA, S. Introduction to satellite navigation systems. In: PELTON, J. N.; MADRY, S.; CAMACHO-LARA, S. (Ed.). *Handbook of Satellite Applications*. Cham: Springer International Publishing, 2017. p. 723–734. Disponível em: <https://doi.org/10.1007/978-3-319-23386-4_10>. Citado na página 16.
- 9 BRASSAROTE, G. d. O. N.; SOUZA, E. M. de; MONICO, J. F. G. S 4 index: does it only measure ionospheric scintillation? *GPS Solutions*, Springer, v. 22, n. 1, p. 8, 2018. Citado na página 16.
- 10 PRIYADARSHI, S. A review of ionospheric scintillation models. *Surveys in geophysics*, Springer, v. 36, n. 2, p. 295–324, 2015. Citado 3 vezes nas páginas 16, 22 e 25.
- 11 ESA. *Space Weather study results*. 2016. Acesso: 28/06/2020. Disponível em: <https://esamultimedia.esa.int/docs/business_with_esas_Space_Weather_Cost_Benefit_Analysis_ESA_2016.pdf>. Citado na página 16.

- 12 KELLY, M. A. *et al.* Progress toward forecasting of space weather effects on uhf satcom after operation anaconda. *Space Weather*, Wiley Online Library, v. 12, n. 10, p. 601–611, 2014. Citado na página 16.
- 13 BASU, S. *et al.* Specification and forecasting of scintillations in communication/navigation links: current status and future plans. *Journal of atmospheric and solar-terrestrial physics*, Elsevier, v. 64, n. 16, p. 1745–1754, 2002. Citado 3 vezes nas páginas 16, 19 e 50.
- 14 CANDER, L. R. Forecasting fof2 and muf (3000) f2 ionospheric characteristics—a challenging space weather frontier. *Advances in Space Research*, Elsevier, v. 56, n. 9, p. 1973–1981, 2015. Citado na página 16.
- 15 SRIDHAR, M. *et al.* Ionospheric scintillation forecasting model based on nn-pso technique. *Astrophysics and Space Science*, Springer, v. 362, n. 9, p. 166, 2017. Citado 3 vezes nas páginas 17, 18 e 36.
- 16 PRIYADARSHI, S. Ionospheric scintillation modeling needs and tricks. In: *Satellites Missions and Technologies for Geosciences*. Online: IntechOpen, 2020. Citado 2 vezes nas páginas 17 e 36.
- 17 JIAO, Y.; HALL, J. J.; MORTON, Y. T. Automatic equatorial gps amplitude scintillation detection using a machine learning algorithm. *IEEE Transactions on Aerospace and Electronic Systems*, IEEE, v. 53, n. 1, p. 405–418, 2017. Citado 3 vezes nas páginas 17, 18 e 36.
- 18 MURPHY, K. P. *Machine learning: a probabilistic perspective*. EUA: MIT press, 2012. Citado 2 vezes nas páginas 17 e 25.
- 19 REZENDE, L. *et al.* Survey and prediction of the ionospheric scintillation using data mining techniques. *Space Weather*, Wiley Online Library, v. 8, n. 6, 2010. Citado 4 vezes nas páginas 17, 18, 36 e 40.
- 20 LIMA, G. de *et al.* Prediction of the level of ionospheric scintillation at equatorial latitudes in brazil using a neural network. *Space Weather*, Wiley Online Library, v. 13, n. 8, p. 446–457, 2015. Citado 2 vezes nas páginas 17 e 18.
- 21 MCGRANAGHAN, R. M. *et al.* New capabilities for prediction of high-latitude ionospheric scintillation: A novel approach with machine learning. *Space Weather*, AGU, v. 16, n. 11, p. 1817–1846, 2018. Citado 2 vezes nas páginas 17 e 18.
- 22 LINTY, N. *et al.* Detection of gnss ionospheric scintillations based on machine learning decision tree. *IEEE Transactions on Aerospace and Electronic Systems*, IEEE, v. 55, n. 1, p. 303–317, 2018. Citado 2 vezes nas páginas 18 e 25.
- 23 WU, A. J.; LIU, Y. Machine learning-based investigation of feature importance for high-latitude ionospheric scintillation forecasting. In: *Proceedings of the 2021 International Technical Meeting of The Institute of Navigation*. USA: ION, 2021. p. 637–647. Citado na página 18.
- 24 ATABATI, A. *et al.* Ionospheric scintillation prediction on s4 and roti parameters using artificial neural network and genetic algorithm. *Remote Sensing*, Multidisciplinary Digital Publishing Institute, v. 13, n. 11, p. 2092, 2021. Citado na página 18.

- 25 LUDWIG-BARBOSA, V. *et al.* Supervised detection of ionospheric scintillation in low-latitude radio occultation measurements. *Remote Sensing*, Multidisciplinary Digital Publishing Institute, v. 13, n. 9, p. 1690, 2021. Citado na página 18.
- 26 IEEE. IEEE Standard Definitions of Terms for Radio Wave Propagation. *IEEE Std 211-2018 (Revision of IEEE Std 211-1997)*, p. 1–57, 2019. Citado na página 21.
- 27 BORA, S. Ionosphere and radio communication. *Resonance*, Springer, v. 22, n. 2, p. 123–133, 2017. Citado na página 21.
- 28 KELLEY, M. C. *The Earth's ionosphere: plasma physics and electrodynamics*. EUA: Academic press, 2009. Citado na página 21.
- 29 DABAS, R. Ionosphere and its influence on radio communications. *Resonance*, Springer, v. 5, n. 7, p. 28–43, 2000. Citado na página 21.
- 30 HARALAMBOUS, H.; PAPADOPOULOS, H.; ECONOMOU, L. Using neural networks for predicting the likelihood of interference to groundwave users in the HF spectrum. In: *Proceedings of the 10th International Conference on Engineering Applications of Neural Networks (EANN 2007)*. Thessaloniki: CEUR-WS, 2007. p. 200–209. Citado na página 21.
- 31 RINO, C. *The theory of scintillation with applications in remote sensing*. EUA: John Wiley & Sons, 2011. Citado na página 22.
- 32 KINTNER, P. M.; LEDVINA, B. M.; PAULA, E. D. Gps and ionospheric scintillations. *Space weather*, Wiley Online Library, v. 5, n. 9, 2007. Citado na página 22.
- 33 GUO, K.; AQUINO, M.; VEETIL, S. V. Ionospheric scintillation intensity fading characteristics and gps receiver tracking performance at low latitudes. *GPS Solutions*, Springer, v. 23, n. 2, p. 1–12, 2019. Citado 2 vezes nas páginas 22 e 23.
- 34 VANI, B. C.; SHIMABUKURO, M. H.; MONICO, J. F. G. Visual exploration and analysis of ionospheric scintillation monitoring data: the ismr query tool. *Computers & Geosciences*, Elsevier, v. 104, p. 125–134, 2017. Citado 2 vezes nas páginas 22 e 42.
- 35 SPOGLI, L. *et al.* Assessing the gnss scintillation climate over brazil under increasing solar activity. *Journal of Atmospheric and Solar-Terrestrial Physics*, Elsevier, v. 105, p. 199–206, 2013. Citado 3 vezes nas páginas 22, 36 e 37.
- 36 ONOHARA, A. N. *et al.* *Demandas nacionais ao setor espacial*. Brasil: AEB, 2019. Citado na página 22.
- 37 KINTNER, P. M.; HUMPHREYS, T.; HINKS, J. *GNSS and ionospheric scintillation*. 2009. Acesso: 28/06/2020. Disponível em: <<https://www.insidegnss.com/auto/julyaug09-kintner.pdf>>. Citado na página 22.
- 38 RAY, S.; PAUL, A.; DASGUPTA, A. Equatorial scintillations in relation to the development of ionization anomaly. *Annales Geophysicae*, v. 24, n. 5, p. 1429–1442, 2006. Disponível em: <<https://angeo.copernicus.org/articles/24/1429/2006/>>. Citado na página 22.

- 39 ABDU, M. *et al.* South atlantic magnetic anomaly ionization: A review and a new focus on electrodynamic effects in the equatorial ionosphere. *Journal of Atmospheric and Solar-Terrestrial Physics*, Elsevier, v. 67, n. 17-18, p. 1643–1657, 2005. Citado na página 22.
- 40 WHEELON, A. D. *Electromagnetic Scintillation: Volume 1, Geometrical Optics*. [S.l.]: Cambridge University Press, 2001. Citado na página 25.
- 41 TIAN, Y. *et al.* Error analysis on ionospheric scintillation index S_4 measured by gnss receiver. *GPS Solutions*, Springer, v. 24, n. 3, p. 1–13, 2020. Citado na página 25.
- 42 LIMA, G. de *et al.* Correlation analysis between the occurrence of ionospheric scintillation at the magnetic equator and at the southern peak of the equatorial ionization anomaly. *Space Weather*, Wiley Online Library, v. 12, n. 6, p. 406–416, 2014. Citado na página 25.
- 43 ELANGO, G. A.; SUDHA, G.; FRANCIS, B. Weak signal acquisition enhancement in software gps receivers—pre-filtering combined post-correlation detection approach. *Applied Computing and Informatics*, Elsevier, v. 13, n. 1, p. 66–78, 2017. Citado na página 25.
- 44 CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. *Computers & Electrical Engineering*, Elsevier, v. 40, n. 1, p. 16–28, 2014. Citado 3 vezes nas páginas 25, 29 e 30.
- 45 REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: LIU, L.; OZSU, M. T. (Ed.). *Encyclopedia of Database Systems*. Boston, MA: Springer US, 2009. p. 532–538. ISBN 978-0-387-39940-9. Disponível em: <https://doi.org/10.1007/978-0-387-39940-9_565>. Citado 2 vezes nas páginas 26 e 27.
- 46 ARLOT, S.; CELISSE, A. *et al.* A survey of cross-validation procedures for model selection. *Statistics surveys*, The author, under a Creative Commons Attribution License, v. 4, p. 40–79, 2010. Citado 2 vezes nas páginas 26 e 28.
- 47 XU, Y.; GOODACRE, R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, Springer, v. 2, n. 3, p. 249–262, 2018. Citado na página 27.
- 48 JAMES, G. *et al.* *An introduction to statistical learning*. EUA: Springer, 2013. v. 112. Citado 2 vezes nas páginas 27 e 32.
- 49 ROBERTS, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, Wiley Online Library, v. 40, n. 8, p. 913–929, 2017. Citado na página 28.
- 50 BERGMEIR, C.; BENÍTEZ, J. M. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, Elsevier, v. 191, p. 192–213, 2012. Citado na página 28.
- 51 WU, J. *et al.* Hyperparameter optimization for machine learning models based on bayesian optimization. *Journal of Electronic Science and Technology*, Elsevier, v. 17, n. 1, p. 26–40, 2019. Citado 2 vezes nas páginas 28 e 29.

- 52 PROBST, P.; BOULESTEIX, A.-L.; BISCHL, B. Tunability: Importance of hyperparameters of machine learning algorithms. *J. Mach. Learn. Res.*, v. 20, n. 53, p. 1–32, 2019. Citado na página 28.
- 53 BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, JMLR. org, v. 13, n. 1, p. 281–305, 2012. Citado na página 29.
- 54 GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research*, v. 3, n. Mar, p. 1157–1182, 2003. Citado 2 vezes nas páginas 29 e 30.
- 55 BORISOV, A.; ERUHIMOV, V.; TUV, E. Tree-based ensembles with dynamic soft feature selection. In: *Feature Extraction*. Heidelberg: Springer, 2006. p. 359–374. Citado na página 29.
- 56 ROGERS, J.; GUNN, S. Identifying feature relevance using a random forest. In: SPRINGER. *International Statistical and Optimization Perspectives Workshop " Subspace, Latent Structure and Feature Selection"*. Heidelberg, 2005. p. 173–184. Citado 2 vezes nas páginas 29 e 55.
- 57 ROSS, B. C. Mutual information between discrete and continuous data sets. *PloS one*, Public Library of Science, v. 9, n. 2, p. e87357, 2014. Citado 2 vezes nas páginas 30 e 55.
- 58 KRASKOV, A.; STÖGBAUER, H.; GRASSBERGER, P. Estimating mutual information. *Physical review E*, APS, v. 69, n. 6, p. 066138, 2004. Citado 2 vezes nas páginas 30 e 55.
- 59 GU, Q.; LI, Z.; HAN, J. Generalized fisher score for feature selection. In: *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. Barcelona: AUAI Press, 2011. (UAI'11), p. 266–273. Citado na página 30.
- 60 LOWRY, R. *Concepts and Applications of Inferential Statistics*. USA: Online, 2020. Disponível: <<http://vassarstats.net/textbook/>>. Acesso: 30/06/2020. Citado na página 30.
- 61 BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 31.
- 62 HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer Science & Business Media, 2009. Citado na página 32.
- 63 NADEAU, C.; BENGIO, Y. Inference for the generalization error. *Machine learning*, Springer, v. 52, n. 3, p. 239–281, 2003. Citado na página 34.
- 64 TRAWIŃSKI, B. *et al.* Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *International Journal of Applied Mathematics and Computer Science*, Sciendo, v. 22, n. 4, p. 867–881, 2012. Citado 2 vezes nas páginas 34 e 35.
- 65 TRIOLA, M. F. *et al.* *Elementary statistics*. EUA: Pearson/Addison-Wesley Reading, 2010. Citado 2 vezes nas páginas 34 e 35.

- 66 PROL, F. dos S. *et al.* Tomographic imaging of ionospheric plasma bubbles based on gnss and radio occultation measurements. *Remote Sensing*, Multidisciplinary Digital Publishing Institute, v. 10, n. 10, p. 1529, 2018. Citado 2 vezes nas páginas 36 e 37.
- 67 CAMARGO, P. O. de; MONICO, J. F. G.; FERREIRA, L. D. D. Application of ionospheric corrections in the equatorial region for ll gps users. *Earth, Planets and Space*, Springer, v. 52, n. 11, p. 1083–1089, 2000. Citado na página 36.
- 68 JESUS, R. de *et al.* Morphological features of ionospheric scintillations during high solar activity using gps observations over the south american sector. *Journal of Geophysical Research: Space Physics*, Wiley Online Library, v. 125, n. 3, 2020. Citado na página 36.
- 69 JUANG, J.-C.; TSAI, Y.-F.; TSAI, C.-T. Design and verification of a magnetometer-based orbit determination and sensor calibration algorithm. *Aerospace Science and Technology*, v. 21, n. 1, p. 47–54, 2012. ISSN 1270-9638. Citado na página 37.
- 70 DENARDINI, C. *et al.* The embrace magnetometer network for south america: Network description and its qualification. *Radio Science*, AGU, v. 53, n. 3, p. 288–302, 2018. Citado na página 37.
- 71 KING, J.; PAPITASHVILI, N. Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data. *Journal of Geophysical Research: Space Physics*, Wiley Online Library, v. 110, n. A2, 2005. Citado na página 38.
- 72 HAPGOOD, M. *Geocentric systems*. 1997. Disponível: <https://www.mssl.ucl.ac.uk/grid/iau/extra/local_copy/SP_coords/geo_sys.htm>. Acesso: 20/08/2020. Citado na página 38.
- 73 REINISCH, B. W. *et al.* New digisonde for research and monitoring applications. *Radio Science*, v. 44, n. 1, 2009. Disponível em: <<https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008RS004115>>. Citado na página 38.
- 74 SOUZA, A. L. C. d.; CAMARGO, P. d. O. Comparison of gnss indices, ionosondes and all-sky imagers in monitoring the ionosphere in Brazil during quiet and disturbed days. *Boletim de Ciências Geodésicas*, SciELO Brasil, v. 25, n. SPE, 2019. Citado 2 vezes nas páginas 38 e 54.
- 75 INPE. *Ionosfera*. 2020. Acesso: 10/06/2020. Disponível em: <<http://www.dae.inpe.br/ionosfera/eires.php>>. Citado na página 38.
- 76 BIBL, K.; REINISCH, B.; KITROSSER, D. F. *General Description of the Compact Digital Ionospheric Sounder Digisonde 256*: Revised. 1981. Disponível em: <https://www.researchgate.net/publication/235130313_General_Description_of_the_Compact_Digital_Ionospheric_Sounder_Digisonde_256_Revised>. Citado na página 38.
- 77 REINISCH, B. W. New techniques in ground-based ionospheric sounding and studies. *Radio Science*, v. 21, n. 3, p. 331–341, 1986. Disponível em: <<https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/RS021i003p00331>>. Citado na página 38.

- 78 REINISCH, B. W. *et al.* The digisonde 256 ionospheric sounder. In: LIU, C. H. (Ed.). *World Ionosphere/ Thermosphere Study, WITS Handbook*. EUA: ICSU Scientific Committee on Solar-Terrestrial Physics (SCOSTEP), 1989. v. 2, p. 350 – 382. Citado na página 38.
- 79 KLIPP, T. d. S. *et al.* Ionosonde total electron content evaluation using international global navigation satellite system service data. In: COPERNICUS GMBH. *Annales Geophysicae*. Online, 2020. v. 38, n. 2, p. 347–357. Citado na página 39.
- 80 DEMYANOV, V. V.; SERGEEVA, M. A.; YASYUKEVICH, A. S. Gnss high-rate data and the efficiency of ionospheric scintillation indices. *Satellites Missions and Technologies for Geosciences*, IntechOpen, 2019. Citado na página 39.
- 81 REZENDE, L. F. C. de. *Mineração de dados aplicada à análise e predição de cintilação ionosférica*. Dissertação (Mestrado), 2009. Citado na página 40.
- 82 ZHU, Z.; LI, C. Study on real-time identification of gnss multipath errors and its application. *Aerospace Science and Technology*, Elsevier, v. 52, p. 215–223, 2016. Citado na página 40.
- 83 SIVAVARAPRASAD, G. *et al.* Performance evaluation of neural network tec forecasting models over equatorial low-latitude indian gnss station. *Geodesy and Geodynamics*, Elsevier, v. 11, n. 3, p. 192–201, 2020. Citado na página 41.
- 84 POOLE, A. W.; MCKINNELL, L.-A. On the predictability of f0f2 using neural networks. *Radio Science*, Wiley Online Library, v. 35, n. 1, p. 225–234, 2000. Citado na página 41.
- 85 MORAES, A. de O. *et al.* The variability of low-latitude ionospheric amplitude and phase scintillation detected by a triple-frequency gps receiver. *Radio Science*, v. 52, n. 4, p. 439–460, 2017. Disponível em: <<https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016RS006165>>. Citado na página 42.
- 86 AON, E. *et al.* Modeling of gps ionospheric scintillation using nonlinear regression technique. In: SPRINGER. *International Conference of Reliable Information and Communication Technology*. Alemanha, 2017. p. 180–188. Citado na página 42.
- 87 COKRLIC, M. *et al.* Ionospheric scintillation monitoring in a single station mode. In: *6th GNSS Vulnerabilities and Solutions Conference*. Croácia: The Royal Institute of Navigation, 2012. p. 71–78. Citado na página 42.
- 88 DIERENDONCK, A. V.; KLOBUCHAR, J.; HUA, Q. Ionospheric scintillation monitoring using commercial single frequency c/a code receivers. In: *Proceedings of ION GPS*. EUA: ION, 1993. v. 93, p. 1333–1342. Citado na página 42.
- 89 PAULA, E. R. de *et al.* Characteristics of the ionospheric irregularities over brazilian longitudinal sector. *Indian Journal of Radio & Space Physics*, v. 36, p. 268–277, 2007. Citado na página 43.
- 90 BRANCO, P.; TORGO, L.; RIBEIRO, R. P. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 49, n. 2, p. 1–50, 2016. Citado 2 vezes nas páginas 47 e 66.

- 91 SUN, Y.; WONG, A. K.; KAMEL, M. S. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, World Scientific, v. 23, n. 04, p. 687–719, 2009. Citado na página 47.
- 92 KING, G.; ZENG, L. Logistic regression in rare events data. *Political analysis*, Cambridge University Press, v. 9, n. 2, p. 137–163, 2001. Citado na página 47.
- 93 PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 48.
- 94 KUTIEV, I. *et al.* Solar activity impact on the earth's upper atmosphere. *Journal of Space Weather and Space Climate*, EDP Sciences, v. 3, p. A06, 2013. Citado na página 54.
- 95 CHERNIAK, I.; ZAKHARENKOVA, I.; KRANKOWSKI, A. Approaches for modeling ionosphere irregularities based on the tec rate index. *Earth, Planets and Space*, SpringerOpen, v. 66, n. 1, p. 1–5, 2014. Citado na página 54.
- 96 MOŠNA, Z.; ŠAULI, P.; SANTOLIK, O. Analysis of critical frequencies in the ionosphere. *WDS'Proceedings of Contributed Papers*, Citeseer, p. 172–177, 2008. Citado na página 54.
- 97 KAKA, R. *et al.* Variations of the critical frequency of the f2 layer, fof2 in west africa using ionosonde stations at Ouagadougou and Dakar. *Research Journal of Applied Science*, Medwell Journals, p. 474–480, 2012. Citado na página 54.
- 98 AMORIM, D. *et al.* Long-term study of medium-scale traveling ionospheric disturbances using oi 630 nm all-sky imaging and ionosonde over brazilian low latitudes. *Journal of Geophysical Research: Space Physics*, Wiley Online Library, v. 116, n. A6, 2011. Citado na página 54.
- 99 PRIYADARSHI, S. *et al.* The behaviors of ionospheric scintillations around different types of nightside auroral boundaries seen at the chinese yellow river station, svalbard. *Frontiers in Astronomy and Space Sciences*, Frontiers, v. 5, p. 26, 2018. Citado na página 54.
- 100 RODRIGUES, F. S. *et al.* Statistical analysis of gps ionospheric scintillation and short-time tec variations over northern europe. *Navigation*, Wiley Online Library, v. 51, n. 1, p. 59–75, 2004. Citado na página 63.
- 101 MICKUS, K. Chapter 5 - geophysical methods. In: MOHAMED, A.-M. O.; PALEOLOGOS, E. K.; HOWARI, F. M. (Ed.). *Pollution Assessment for Sustainable Practices in Applied Sciences and Engineering*. [S.l.]: Butterworth-Heinemann, 2021. p. 199–287. ISBN 978-0-12-809582-9. Citado na página 85.
- 102 BUDDEN, K. G. *The propagation of radio waves: The theory of radio waves of low power in the ionosphere and magnetosphere*. Reino Unido: Cambridge University Press, 1988. Citado na página 86.
- 103 BROWN, B. *HF propagation tutorial*. 2004. Disponível: <<http://www.astrosurf.com/luxorion/qs1-hf-tutorial-nm7m4.htm>>. Acesso: 26/01/2022. Citado na página 86.

APÊNDICE A – Coeficientes dos modelos RLM

Tabela 30 – Coeficientes do modelo RLM RR_1

Parâmetro	Coeficiente
1	0,720668
F_M	$-2,90000 \cdot 10^{-5}$
$foF2$	$-1,55000 \cdot 10^{-4}$
$fminF$	$-3,83700 \cdot 10^{-3}$
fxI	$-3,37000 \cdot 10^{-4}$
FF	$6,79000 \cdot 10^{-3}$
$hmF2$	$2,11000 \cdot 10^{-4}$
$yF2$	$-2,55000 \cdot 10^{-4}$
CET	$-1,21000 \cdot 10^{-4}$
$f(h'F)$	$6,11900 \cdot 10^{-3}$
$f(h'F2)$	$-1,33200 \cdot 10^{-3}$
CCH	$-9,56500 \cdot 10^{-3}$
$ROTI$	$1,73000 \cdot 10^{-3}$
$MA-S4$	$8,62885 \cdot 10^{-1}$

Fonte: Autoria própria (2021)

Tabela 31 – Coeficientes do modelo RLM RR_2

Parâmetro	Coeficiente
1	0,42083
H_M	$-1,40000 \cdot 10^{-5}$
$foF2$	$-1,45000 \cdot 10^{-4}$
fxI	$-1,41000 \cdot 10^{-4}$
$yF2$	$-9,50000 \cdot 10^{-5}$
TEC	$-4,01000 \cdot 10^{-4}$
CSD	$2,05580 \cdot 10^{-2}$
CCD	$1,79600 \cdot 10^{-3}$
CSH	$-1,75010 \cdot 10^{-2}$
CCH	$-1,51000 \cdot 10^{-2}$
$ROTI$	$1,76000 \cdot 10^{-3}$
$MA-S4$	$8,56763 \cdot 10^{-1}$

Fonte: Autoria própria (2021)

Tabela 32 – Coeficientes do modelo RLM RR_3

Parâmetro	Coeficiente
1	-0,014375
$h'F2$	$4,60000 \cdot 10^{-5}$
$h'F$	$5,40000 \cdot 10^{-5}$
CSH	$-4,77400 \cdot 10^{-3}$
$ROTI$	$1,80500 \cdot 10^{-3}$
$MA-S4$	$8,63395 \cdot 10^{-1}$

Fonte: Autoria própria (2021)

Tabela 33 – Coeficientes do modelo RLM Base

Parâmetro	Coeficiente
1	-1,823794
F_M	$1,10000 \cdot 10^{-4}$
H_M	$-5,00000 \cdot 10^{-5}$
D_M	$-1,19690 \cdot 10^{-2}$
IMF	$2,56000 \cdot 10^{-4}$
B_z	$1,98700 \cdot 10^{-3}$
P_P	$-7,30000 \cdot 10^{-5}$
IEF	$6,10600 \cdot 10^{-3}$
$foF2$	$-1,37000 \cdot 10^{-4}$
$fminF$	$-5,80600 \cdot 10^{-3}$
$fmin$	$-2,10000 \cdot 10^{-5}$
$h'F2$	$3,4000 \cdot 10^{-5}$
$h'F$	$1,90000 \cdot 10^{-5}$
fxI	$-3,87000 \cdot 10^{-4}$
FF	$3,94100 \cdot 10^{-3}$
$hmF2$	$7,00000 \cdot 10^{-5}$
$yF2$	$-1,73000 \cdot 10^{-4}$
CET	$-2,07000 \cdot 10^{-4}$
$f(h'F)$	$6,39100 \cdot 10^{-3}$
$f(h'F2)$	$-1,96300 \cdot 10^{-3}$
CSD	$2,67950 \cdot 10^{-2}$
CCD	$3,94000 \cdot 10^{-4}$
CSH	$-1,03040 \cdot 10^{-2}$
CCH	$-1,28600 \cdot 10^{-2}$
$ROTI$	$1,70300 \cdot 10^{-3}$
$MA-S4$	$8,57695 \cdot 10^{-1}$
θ_A	$4,00000 \cdot 10^{-6}$
θ_E	$2,54000 \cdot 10^{-4}$

Fonte: Autoria própria (2021)

APÊNDICE B – Detalhamento das bases de dados

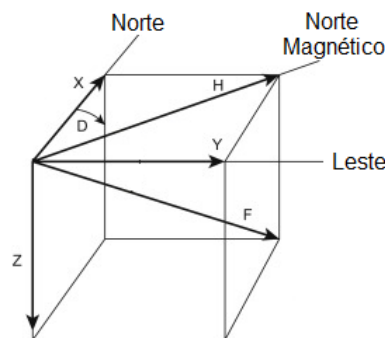
B.1 Definições de atributos

Nesta seção são detalhados alguns conceitos relacionados aos dados provenientes de magnetômetros e ionossondas.

B.1.1 Magnetômetro

Os parâmetros F_M , H_M e D_M são definidos de acordo com a Figura 27, indicados por F , H e D , respectivamente.

Figura 27 – Parâmetros do campo magnético terrestre.



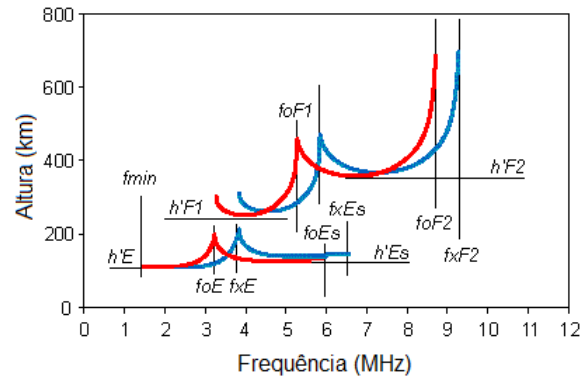
Fonte: Adaptado de [Mickus \[101\]](#)

B.1.2 Ionossonda

- Frequência crítica - Máxima frequência refletida de volta para a Terra por uma determinada camada da ionosfera.
- Conteúdo eletrônico total - Quantidade de elétrons presente em uma coluna vertical com área de seção transversal unitária, que estende-se desde o satélite até o receptor em solo.
- Altura virtual - Altura na qual uma onda incidente é refletida em direção à Terra por uma determinada camada da ionosfera.
- FF - Diferença entre os valores das frequências f_{xI} e f_{xF2} .

Diversos parâmetros fornecidos por ionossondas são derivados a partir de ionogramas. Na Figura 28 é possível observar como algumas dessas informações são determinadas. Um descrição mais detalhada pode ser observada em Budden [102].

Figura 28 – Exemplo de ionograma



Fonte: Adaptado de Brown [103]

B.2 Base de dados usada para regressão

Tabela 34 – Descrição da base de dados MAG usada nos modelos de regressão

Parâmetro	Unidade	Média	Desvio Padrão	Valor máximo	Valor mínimo
F_M	nT	26256,4946	25,1394	26334,7000	26181,5000
H_M	nT	26008,6570	28,9392	26085,8000	25932,2000
D_M	°	-21,0166	0,0381	-20,8998	-21,0991

Fonte: Autoria própria (2021)

Tabela 35 – Descrição da base de dados SW usada nos modelos de regressão

Parâmetro	Unidade	Média	Desvio Padrão	Valor máximo	Valor mínimo
IMF	nT	7,0478	2,3612	15,5700	0,6100
B_z	nT	0,2972	3,6810	12,7700	-10,1800
P_P	nPa	2,1336	1,2068	11,2700	0,3000
IEF	mV/m	-0,1192	1,5333	4,5300	-4,9600

Fonte: Autoria própria (2021)

Tabela 36 – Descrição da base de dados ION usada nos modelos de regressão

Parâmetro	Unidade	Média	Desvio Padrão	Valor máximo	Valor mínimo
$foF2$	MHz	4,2964	3,1109	9,9620	0,0000
$fminF$	MHz	3,5009	1,3399	11,0000	1,1000
$fmin$	MHz	2,8006	1,0768	9,6000	0,2000
$h'F2$	km	257,1155	59,5895	558,0000	92,0000
$h'F$	km	255,0303	60,1054	558,0000	92,0000
fxI	MHz	4,4208	3,1086	9,9500	0,0000
FF	MHz	0,2262	0,2292	3,0000	0,0500
$hmF2$	km	368,6544	68,2635	620,0000	223,5000
$yF2$	km	128,7009	56,6081	296,8000	10,7000
TEC	TECU ⁵	44,5989	29,5174	99,5000	0,4000
$f(h'F)$	MHz	3,6077	1,2953	9,6000	0,2000
$f(h'F2)$	MHz	3,6387	1,3218	9,6000	0,2000

Fonte: Autoria própria (2021)

Tabela 37 – Descrição da base de dados GPS usada nos modelos de regressão

Parâmetro	Unidade	Média	Desvio Padrão	Valor máximo	Valor mínimo
$ROTI$	TECU/s	3,4622	12,4077	174,7900	0,0000
θ_E	°	38,4856	8,2969	87,0000	30,1000
θ_A	°	195,9758	88,5711	359,0000	0,0000
S_4	—	0,1446	0,1916	1,6900	0,0500

Fonte: Autoria própria (2021)

Tabela 38 – Descrição da base de dados MAG usada nos modelos de classificação

Parâmetro	Unidade	Média	Desvio Padrão	Valor máximo	Valor mínimo
F_M	nT	26120,6665	403,4080	26334,7000	24072,4000
H_M	nT	25789,8839	432,8204	26085,8000	23703,6000
D_M	°	-20,5289	0,2813	-20,3302	-21,6723

Fonte: Autoria própria (2021)

Tabela 39 – Descrição da base de dados SW usada nos modelos de classificação

Parâmetro	Unidade	Média	Desvio Padrão	Valor máximo	Valor mínimo
IMF	nT	5,4414	2,7317	32,5700	0,3200
B_z	nT	-0,0742	3,0580	26,8000	-22,4900
P_P	nPa	2,1688	1,3956	29,0500	0,2500
IEF	mV/m	0,0409	1,3539	13,1800	-13,7300

Fonte: Autoria própria (2021)

B.3 Base de dados usada para classificação

Tabela 40 – Descrição da base de dados ION usada nos modelos de classificação

Parâmetro	Unidade	Média	Desvio Padrão	Valor máximo	Valor mínimo
$foF2$	MHz	6,2564	2,5527	15,0500	0,0000
$fminF$	MHz	2,9240	1,0777	11,0000	1,0000
$fmin$	MHz	1,9538	0,7188	9,6000	0,2000
$h'F2$	km	273,5282	68,2843	840,0000	92,0000
$h'F$	km	230,9671	54,6777	760,0000	92,0000
fxI	MHz	6,5200	2,6137	15,4500	0,0000
FF	MHz	0,1775	0,1932	3,0000	0,0500
$hmF2$	km	314,2713	54,6167	667,4000	141,9000
$yF2$	km	103,2448	53,9764	445,2000	4,7000
$f(h'F)$	MHz	3,0029	1,0922	9,6000	0,2000
$f(h'F2)$	MHz	3,3507	1,3872	9,6000	0,2000

Fonte: Autoria própria (2021)

⁵ 1 TECU = 10^{16} elétrons/m²

Tabela 41 – Descrição da base de dados GNSS usada nos modelos de classificação

Parâmetro	Unidade	Média	Desvio Padrão	Valor máximo	Valor mínimo
<i>tec</i>	TECU	40,3669	26,3352	170,7900	-1,1780
<i>tec_15</i>	TECU	40,3576	26,3601	168,4390	-5,4720
<i>tec_30</i>	TECU	40,3676	26,3727	169,1690	-4,3480
<i>tec_45</i>	TECU	40,4043	26,3500	172,4820	-3,9480
<i>dtec_15tow</i>	TECU	0,0005	0,0895	2,8300	-1,5740
<i>dtec_3015</i>	TECU	-0,0004	0,0909	1,6080	-4,1470
<i>dtec_4530</i>	TECU	-0,0007	0,0934	2,3300	-2,0700
<i>dtec_6045</i>	TECU	-0,0005	0,0907	3,2550	-1,4440
<i>avg_cn0_l1</i>	dB-Hz	46,5762	2,1451	54,0000	41,2000
θ_E	°	37,3535	7,4056	88,0000	31,0000
θ_A	°	199,0739	107,1768	359,0000	0,0000

Fonte: Autoria própria (2021)