


Departamento de
Meteorologia
Universidade Federal do Rio de Janeiro
Centro de Ciências Matemáticas e da Natureza
Instituto de Geociências

MODELO DE PREVISÃO DE EVENTOS CONVECTIVOS BASEADO EM INTELIGÊNCIA ARTIFICIAL PARA A ROTA AÉREA RIO DE JANEIRO – SÃO PAULO

CAROLINE MENEGUSSI SOARES

Rio de Janeiro - RJ

Setembro de 2020

MODELO DE PREVISÃO DE EVENTOS CONVECTIVOS BASEADO EM INTELIGÊNCIA ARTIFICIAL PARA A ROTA AÉREA RIO DE JANEIRO – SÃO PAULO

CAROLINE MENEGUSSI SOARES

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Meteorologia do Instituto de Geociências do Centro de Ciências Matemáticas e da Natureza da Universidade Federal do Rio de Janeiro (PPGM – IGEO – CCMN – UFRJ), como parte dos requisitos necessários à obtenção do grau de Mestre em Meteorologia.

Orientadores: Gutemberg Borges França, *PhD*.
Manoel Valdonel de Almeida, *DSc*.

Rio de Janeiro
Setembro de 2020

MODELO DE PREVISÃO DE EVENTOS CONVECTIVOS BASEADO EM INTELIGÊNCIA ARTIFICIAL PARA A ROTA AÉREA RIO DE JANEIRO – SÃO PAULO

CAROLINE MENEGUSSI SOARES

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM METEOROLOGIA DO INSTITUTO DE GEOCIÊNCIAS DO CENTRO DE CIÊNCIAS MATEMÁTICAS E DA NATUREZA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO (PPGM-IGEO-CCMN-UFRJ) COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA À OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS (ÁREA: METEOROLOGIA).

PROF. GUTEMBERG BORGES FRANÇA (ORIENTADOR)

DR. MANOEL VALDONEL DE ALMEIDA (ORIENTADOR)

PROF. DR. HUGO ABI KARAM

PROF. DR. WALLACE FIGUEIREDO MENEZES

DR. HAROLDO FRAGA DE CAMPOS VELHO

RIO DE JANEIRO, RJ – BRASIL

SETEMBRO DE 2020

Soares-Menegussi, Caroline

Modelo de previsão de eventos convectivos baseado em inteligência artificial para a rota aérea rio de janeiro – São Paulo / Caroline Menegussi Soares. - Rio de Janeiro: UFRJ/PPGM/IGEO/CCMN, 2020.

XXIV, 94 p.: il.; 29,7 cm.

Orientador: Gutemberg Borges França

Manoel Valdonel de Almeida

Dissertação - (Mestrado) UFRJ/PPGM/IGEO/CCMN Programa de Pós-graduação em Meteorologia, 2020.

Referências Bibliográficas: p. 87-94.

1. Eventos Meteorológicos Convectivos. 2. Inteligência Computacional. 3. Descargas Atmosféricas. 4. Meteorologia Aeronáutica. I. França, Gutemberg Borges *et al*. II. Universidade Federal do Rio de Janeiro, PPGM/IGEO/CCMN, Programa de Pós-graduação em Meteorologia. III. Título.

AGRADECIMENTOS

A Deus, por me ajudar a ultrapassar todos os obstáculos encontrados ao longo deste período.

Agradeço aos meus pais Lenilza e Moji pela força, amor e apoio incondicional.

Aos meus doces avós Laura e José (*in memoriam*), com todo amor do mundo e gratidão.

Aos meus queridos professores e orientadores Gutemberg e Valdonel pelas correções e ensinamentos que me permitiram apresentar um melhor desempenho no meu processo de formação.

Aos meus amigos do Laboratório de Meteorologia Aplicada e de vida, Ana Carolina, Iago, Vinicius, Mayara, Fabrício e Jimmy, pela ajuda, pelos momentos de felicidade e amizade, principalmente nas horas difíceis.

Agradeço também à Cátedra de Meteorologia Aeronáutica, e ao Departamento de Controle do Espaço Aéreo (DECEA) por me proporcionar a oportunidade de trabalhar com este tema, na forma de autor e bolsista da Cátedra de Meteorologia Aeronáutica. Os créditos deste trabalho são devidos ao DECEA, que me possibilitou a bolsa de pesquisa, por meio da Organização Brasileira para o Desenvolvimento Científico e Tecnológico do Controle do Espaço Aéreo (CTCEA).

E por fim, meus agradecimentos a Eletrobrás Furnas pela gentileza em fornecer os dados de descargas atmosféricas utilizados neste trabalho.

RESUMO

Este propõe modelos de previsão de 6-8h dos eventos meteorológicos convectivos, baseado em inteligência computacional, para duas áreas I e II correspondendo à rota aérea Rio de Janeiro – São Paulo. Os modelos têm como dados de entrada (input) cincos índices de instabilidade atmosféricos extraídos de dados de sensoriamento remoto do satélite GOES-R, considerando os períodos de janeiro a março de 2018 e 2019. A caracterização espaço-temporal dos eventos meteorológicos convectivos foi realizada através de dados de descargas atmosféricas de 2001 a 2019 e o estudo climatológico definiu o período de previsão dos modelos. Cinco algoritmos de inteligência computacional foram treinados considerando mais 200 experimentos visando à previsão de eventos de meteorológicos convectivos e a sua severidade. Os resultados indicam que os algoritmos denominados *Multilayer Perceptron* e *Simple Logistic* foram os de melhores desempenhos para a previsão de 8h e 6h, visto que os valores médio, entre parênteses, das estatísticas de desempenho estão mais próximo do valor ideal, e são: a probabilidade de detecção (0,86 e 0,94), taxa da falso alarme (0,14 e 0,08), viés (1,01 e 1,01), *F-measure* (0,86 e 0,94) e KAPPA (0,72 e 0,85), nas áreas I e II, respectivamente. Resultados dos modelos de previsão (algoritmo *Random Forest* de melhor desempenho) da severidade de 6h e 8h dos eventos são apresentados e os valores das estatísticas são superiores as anteriores para a área II. Resultados de *hand cast* (estudo de caso) do mês de abril de 2019 revelam que os modelos (algoritmos treinados) foram capazes de capturar dos dados o conteúdo físico e assim identificar 96% dos dias eventos convectivos severos ou não, com 6h e 8h de antecedência para área de estudo e talvez possa ser utilizada para aperfeiçoar a previsão operacional deste eventos na região de estudo.

Palavras-chave: eventos meteorológicos convectivos, inteligência computacional, descargas atmosféricas

SHORT TERM FORECAST OF CONVECTIVE EVENTS USING ARTIFICIAL INTELLIGENCE FOR THE ROUTE RJ – SP

ABSTRACT

It proposes 6-8h models to forecast convective meteorological events, based on machine learning, for the Rio de Janeiro - São Paulo air route, which was divided into two areas (I and II). The models input data are the five atmospheric instability indexes derived from remotely sensed data by GOES-R during January-March 2018 and 2019, respectively. Lightning discharge data were used to characterize the spatial-temporal behavior of convective events during the period from 2001 to 2019 and its climatological study defined the forecast periods of the proposed models. More than 200 experiments were carried out using five computational intelligence algorithms aiming at predicting convective or non-convective meteorological events and their severity. The results indicate that the Multilayer Perceptron and Simple Logistic algorithms presented better average values, in parentheses, of the performance statistics which are the detection probability (0.86 and 0.94), false alarms rate (0.14 and 0.08), bias (1.01 and 1.01), f-measure (0.86 and 0.94) and KAPPA (0.72 and 0.85) of 8h and 6h predicted for areas I and II, respectively. The 6h and 8h severity forecast models (Random Forest algorithm) of the events are presented and the values of the statistics are higher than the previous ones for area II. Hand cast (case study) results for April 2019 revealed that the suggested 6h and 8h prediction models (areas I and II) were able to capture the physical content from the data and identified 96% of severe convective events or not for both areas. The models' results may be used to improve the forecast of severe convective event during warm season in aforementioned air route.

Keywords: convective meteorological events, aviation meteorology, computational intelligence, atmospheric discharge

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABI	*Advanced Baseline Imager*
AM	Aprendizado de máquina
B	*Bias*
CAPE	*Convective Available Potencial Energy*
CB’s	*cumulonimbus*
CEMIG	Companhia Energética de Minas Gerais
CENIPA	Centro de Investigação e Prevenção de Acidentes Aeronáuticos
CIMAER	Centro Integrado de Meteorologia Aeronáutica
COPEL	Companhia Paranaense de Energia
CT	*Cross Totals*
DA	Descargas Atmosféricas
DECEA	Departamento de Controle do Espaço Aéreo
DTCEA	Destacamento de Controle do Espaço Aéreo
EMC	Eventos Meteorológicos Convectivos
GOES	Satélite Ambiental Operacional Geoestacionário
IA	Inteligência Artificial
INPE	Instituto Nacional de Pesquisas Espaciais
IS	Índice *Showalter*
KNN	*Nearest Neighbors*
IL	Índice de Levantamento
K	Índice K
MLP	*Multilayer Perceptron*
NASA	*National Aeronautics and Space Administration*

NCC	Nível de Condensação Convectiva
NOAA	*National Oceanic and Atmospheric Administration*
OACI	Organização de Aviação Civil Internacional
OMM	Organização Meteorológica Mundial
POD	Probabilidade de Detecção ou Previsão Correta do Evento
POFD	Probabilidade de Falsa Detecção ou alarme falso
RAMS	*Regional Atmospheric Modeling System*
RINDAT	Rede Integrada Nacional de Detecção de Descargas Atmosféricas
RMRJ	Região Metropolitana do Rio de Janeiro
RMS	*Root Mean Square*
SBGL	Aeroporto Internacional do Rio de Janeiro – Galeão
SBMT	Aeroporto do Campo de Marte – São Paulo
SIMEPAR	Sistema Meteorológico do Paraná
SISCEAB	Sistema de Controle do Espaço Aéreo
SVM	*Support Vector Machine*
TT	*Total Totals*
VFDT	*Very Fast Decision Tree*
VT	*Vertical Totals*
WEKA	*Waikato Enviroment for Knowledge Analysis*

1. INTRODUÇÃO

A previsão local de tempestades é uma das tarefas mais desafiadoras na previsão do tempo devido à sua alta variabilidade espaço-temporal. As condições do tempo interferem no cotidiano de todos, mas poucas atividades humanas são tão dependentes das condições da atmosfera quanto a navegação aérea. Vários sistemas e fenômenos meteorológicos têm sido apontados como os responsáveis pelos acidentes e incidentes aeronáuticos ocorridos no mundo (CENIPA, 2018), (GULTEPE *et al.*, 2019); o que significa que a compreensão da atmosfera torna possível a condução de voos mais seguros.

Nas últimas décadas, observou-se um aumento significativo dos desastres naturais, muitos deles são induzidos por fenômenos atmosféricos, como as chuvas. Os recentes estudos relacionados às mudanças climáticas apontam o aumento do número de eventos extremos, como por exemplo, mais episódios de chuvas concentradas, ou mais secas (NUNES, 2015).

Os Eventos Meteorológicos Convectivos (EMC) estão entre as principais causas das catástrofes naturais que impactam diretamente a sociedade e o meio ambiente. Portanto é de fundamental importância a caracterização desses fenômenos e consequentemente o conhecimento termodinâmico do mesmo para aumentar a assertividade de suas previsões.

A Meteorologia Aeronáutica é a área da meteorologia destinada à aviação que se propõe a contribuir para a garantia dos padrões de segurança, de economia e de eficiência dos voos.

Durante a Conferência de Chicago, em novembro de 1944, na qual teve origem a Organização de Aviação Civil Internacional (OACI), foi estabelecido que os países membros mantenham um serviço meteorológico com o intuito de fornecer aos usuários as informações sobre as condições do tempo necessárias à segurança das operações aéreas (DOC9750, 2016).

A Organização Meteorológica Mundial (OMM) é uma instituição das Nações Unidas que auxilia a OACI na elaboração de normas e de procedimentos específicos de meteorologia para a aviação e no treinamento de profissionais.

No Brasil, a atividade de Meteorologia Aeronáutica é de competência do Comando da Aeronáutica e desenvolvida pelo Sistema de Controle do Espaço Aéreo (SISCEAB), sob a responsabilidade do Departamento de Controle do Espaço Aéreo (DECEA) e do Centro Integrado de Meteorologia Aeronáutica (CIMAER).

1.1 MOTIVAÇÃO E OBJETIVO

O voo requer o conhecimento das condições meteorológicas reinantes, em rota, nos aeródromos de partida e destino (alternativas) visando decisão da realização, ou não, do voo. Para tanto, é necessário saber se essas condições de tempo sofrerão variações significativas, como, de formação de gelo, turbulência, trovoadas (nuvens *cumulonimbus*, CB's), que possam acarretar possíveis desvios de rota e por consequência maior consumo de combustível, no teto, na visibilidade, no vento de rajada (ou surgimento de tesoura de vento). Desta forma, o controle do tráfego requer previsões assertivas sobre o horário previsto para início dessas variações e o período previsto para sua duração.

Segundo a OACI (DOC9750, 2016) o tráfego aéreo global dobrou de tamanho uma vez a cada 15 anos desde 1977 e continuará a fazê-lo. O crescente aumento no número de aeronaves em operação exige dos órgãos de controle a otimização do espaço aéreo, que com a devida segurança, deve executar a diminuição do espaçamento entre aeronaves em todos estágios do voo. Nesse contexto, as informações meteorológicas passam a ser cada vez mais decisivas.

Em particular, a rota aérea Rio de Janeiro - São Paulo, denominada tubulão e representada na Figura 2.1 pela área em cor vermelha, é a mais movimentada do continente sul-americano e a quinta com maior número de voos domésticos do mundo, visto que no ano de 2017, foram registradas 39.325 viagens entre Rio de Janeiro e São Paulo, uma média de 107 por dia (disponível em: https://g1.globo.com/economia/noticia/rota-aerea-mais-movimentada-das-americas-esta-no-brasil.ghtml). Segundo Bender (2018), a região entre o Rio de Janeiro e São Paulo é fortemente impactada pelos EMC (conforme apresentado neste trabalho na seção de resultados) principalmente nos meses mais quentes do ano. Estes eventos são

frequentemente responsáveis por desvios indesejáveis das aeronaves ocasionando em prejuízos significativos a aviação comercial.

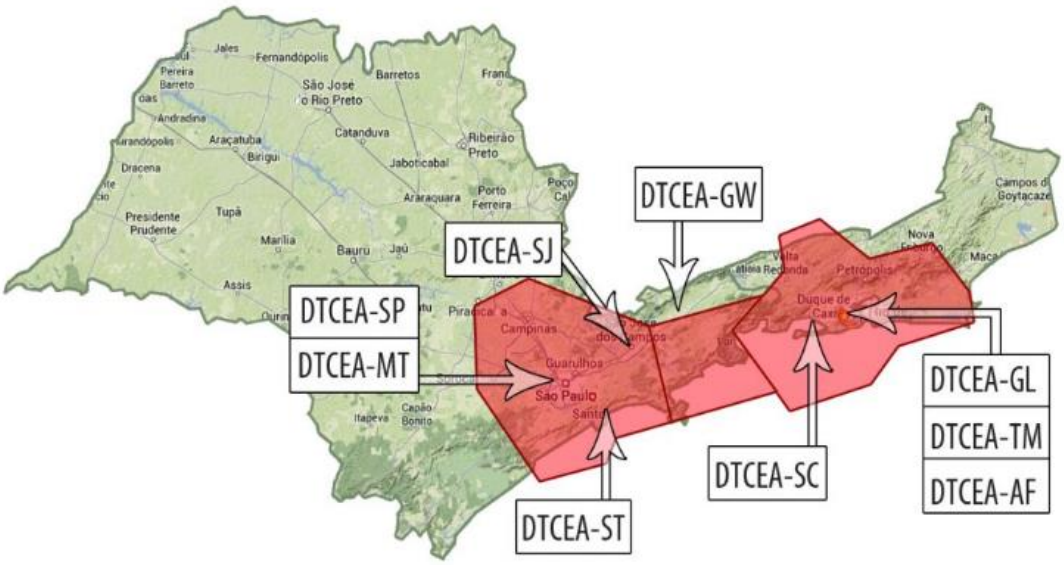


Figura 1.1: A área em vermelho compreende o "tubulão". Disponível em: https://www.decea.gov.br/?i=quem-somos&p=espaco-aereo-brasileiro. Os retângulos, em branco, representam a localização aproximadas dos nove Destacamento de Controle do Espaço Aéreo (DTCEA) na área do Tubulão (rota RJ-SP).

Segundo França *et al*. (2016), no Brasil a previsão meteorológica é realizada, ainda, pelos meteorologistas de forma subjetiva integrando as informações diversas disponíveis. Além disso, sabe-se que no Brasil ainda não existe modelos objetivos de previsão de eventos convectivos para áreas específicas. Sendo assim, para melhor atender as necessidades operacionais do órgão de controle do tráfego aéreo nacional, o CIMAER necessita de ferramentas preditivas objetivas que possam auxiliar os meteorologistas a elaborar suas previsões assertivas para locais específicos.

Portanto, o objetivo deste trabalho divide-se em duas partes, a saber:

a) Estudo das ocorrências dos eventos convectivos na rota Rio - São Paulo; e

b) Avaliar o uso de Inteligência Artificial (IA) como ferramenta de previsão de eventos convectivos e sua e severidade, considerando como variáveis preditivas aos índices termodinâmicos determinados dos perfis atmosféricos extraídos dos sondadores a bordo de plataforma orbital.

2. REVISÃO BIBLIOGRÁFICA

Desde a década de 1900, o impacto dos processos atmosféricos na aviação é reconhecido (GULTEPE *et al*., 2019). Dines (1917) afirmou que a dependência do aviador no meteorologista é que este será capaz de prever vento e nevoeiro, e em menor medida nuvens, na rota, e desta forma ele poderá realizar o voo.

A partir da necessidade de se desenvolver predições de curto prazo, diversos trabalhos associados à previsão do tempo vêm sendo desenvolvidos, como, Wilson (1966), Wilk e Gray (1970) que sugeriram abordagens de previsão fundamentando-se em extrapolações de dados de radar meteorológico limitando-se à interpretação subjetiva para gerar previsões de tempestades.

O crescimento demográfico, avanço computacional e a disponibilidade de dados remotos e *in situ*, permitiram que os modelos numéricos de previsão do tempo fornecessem previsões de curto prazo acuradas. Bender (2018) estudou a relação do crescimento populacional, consequente ilha de calor, e o potencial aumento da atividade convectiva e da severidade das tempestades na Região Metropolitana de São Paulo, utilizando simulações com o modelo BRAMS. O estudo conclui que o aumento da mancha urbana é capaz de aumentar a quantidade de chuva, e que o aumento da área com construções altas tende a causar redução da precipitação.

Mccann (1992) discutiu uma previsão de tempestade significativa de 3 a 7 horas, desenvolvida com redes neurais. Duas redes aprenderam a prever tempestades significativas a partir de campos de índices baseados na convergência de umidade da superfície. Desse modo, inferiu que essas redes são sensíveis aos padrões que os especialistas em meteorologia reconhecem como ocorrendo antes de fortes tempestades. O seu estudo também mostrou o potencial impacto das redes neurais em previsões significativas de trovoadas.

Mueller *et al*. (2003) propôs um sistema de previsão de até 1h em locais de tempestades utilizando-se de dados de superfície, radar, dados de satélite, e modelagem numérica, considerando os estágios da tempestade.

Em relação às aplicações na aviação, Isaac *et al*. (2006, 2011, 2014) apresentaram uma sequência de trabalhos que resultaram em um refinado sistema de previsão para a aviação usando dados de modelos numéricos, observações de superfície, radar, satélite e um radiômetro de micro-ondas para gerar projeções de até aproximadamente 6h para os principais aeroportos no Canadá.

Nascimento (2005) realizou uma discussão sobre a questão da previsão de tempo severo no Brasil, fazendo uma descrição atualizada de parâmetros atmosféricos úteis no auxílio à identificação de ambientes favoráveis à ocorrência de tempestades convectivas severas. Para o cálculo dos parâmetros atmosféricos fez uso de saídas de modelos de mesoescala. Ele mostrou que há indícios de que a ocorrência de tempestades severas no Brasil não é tão rara como tipicamente considerada, e o fato de que alguns índices de tempo severo originalmente concebidos para as latitudes médias do hemisfério norte podem ser úteis para a previsão de sistemas convectivos no Brasil e servir de base conceitual para a elaboração de índices mais adequados para as regiões tropicais do país.

De acordo com Nascimento (2005), a detecção de raios é essencial para o *nowcasting* e monitoramento do tempo. Uma chuva forte, com grande atividade elétrica e grande de intensidade em um grande centro urbano, por exemplo, ocasiona grandes inundações.

As Descargas Atmosféricas (DA) também chamadas de raios ou descargas elétricas, em sua grande maioria, ocorrem durante tempestades com chuvas e ventos intensos em nuvens CB's, mas também podem ocorrer em tempestades de neve, areia, erupções vulcânicas e em nuvens que não sejam de tempestade, mesmo que em menor intensidade.

Paulucci (2017) realizou um estudo climatológico, utilizando descargas atmosféricas, sobre as características espaço-temporais para a Região Metropolitana do Rio de Janeiro (RMRJ). Ele concluiu que regiões com maior ocorrência de tempestades possui grande atividade elétrica. Seus resultados também mostraram que os raios ocorrem predominantemente no verão com máximo no mês de fevereiro.

Da mesma forma, Hermsdorff (2018) realizou um estudo sobre as descargas atmosféricas e sua previsão através de Inteligência Artificial, especificamente árvores de decisão, na RMRJ, com o proposto de melhor entender este fenômeno e aumentar previsibilidade de tempestades convectivas profundas. Foram utilizados dados de descargas atmosféricas da rede RINDAT e dados de sondagem da estação do Aeroporto Internacional do Rio de Janeiro – Galeão (SBGL). Foram selecionadas variáveis termodinâmicas através de análises e correlações. Os resultados indicaram árvores de decisão com taxa de acertos geral acima de 80% e taxa de erro geral abaixo de 20% nos eventos.

Grossman (2010) analisou a contribuição dos índices de instabilidade, com o intuito de determinar os locais favoráveis à formação de tempestades intensas que deram

origem a fortes chuvas. Foram realizadas simulações numéricas de quatro tempestades que atingiram o estado do Rio de Janeiro e causaram precipitação intensa, sendo um desses casos, uma tempestade severa com descargas atmosféricas e granizo sobre o município do Rio de Janeiro. Foi verificado que para os quatro eventos, o comportamento dos índices de instabilidade caracterizou os locais que eram favoráveis à formação de tempestades quando comparados com imagens de satélite e radar.

França *et al*. (2016) apresentaram um novo modelo, baseado em técnicas de rede neural, para produzir a curto prazo e em locais específicos, previsões de instabilidade significativa para voos na área terminal do Aeroporto do Galeão, Rio de Janeiro, Brasil. O estudo mostrou que o modelo proposto pode captar o conteúdo físico dentro do conjunto de dados e seu desempenho é bastante encorajador para a primeira e a segunda horas a transmitir eventos significativos de instabilidade na área de estudo. Mais recentemente, França *et al.* (2018) também mostraram um modelo a curto prazo para perfis de vento de baixo nível no Aeroporto Internacional de Guarulhos, São Paulo, Brasil.

Seguindo esta mesma base, Almeida *et al*. (2020) submeteram um modelo de previsão para eventos convectivos meteorológicos na área de controle de terminais do Aeroporto Internacional do Galeão, Rio de Janeiro, Brasil, usando inteligência artificial, sonorização e sensoriamento remoto dados descarregados atmosféricos. Os resultados mostraram que os classificadores Random Forest e a rede RBF foram os melhores modelos para a detecção e severidade de tempestades, respectivamente, com estatísticas de avaliação de POD (0,82 e 0,75), BIAS (1,0 e 0,99) e FAR (0,18 e 0,24). Sua análise mostrou que o modelo tem desempenho crescente para eventos de alto impacto, e este desempenho diminui gradualmente à medida que os eventos se tornam mais fracos e mais frequentes.

Dessa forma, implica-se que os eventos extremos juntamente com a ocorrência de descargas elétricas têm alta correlação, de forma que sua formação é influenciada principalmente por características meteorológicas locais e da sua interação com a topografia da região de desenvolvimento.

3. FUNDAMENTAÇÃO TEÓRICA

Esta seção discorre sucintamente sobre evento meteorológico convectivo, os índices termodinâmicos de instabilidade atmosférica − primordialmente utilizada pelos meteorologistas como referência na previsão de EMC − e os princípios da inteligência computacional enfatizando os algoritmos utilizados na construção dos modelos preditivos de curto prazo de EMC utilizados neste trabalho.

3.1 EVENTOS METEOROLÓGICOS CONVECTIVOS

Brooks *et al.* (2003), define como tempestades severas, eventos com ocorrência de granizo, rajadas de ventos e ocorrência de tornados, a partir de observações e variáveis meteorológicas associadas com a intensidade convectiva. Dessa forma, os EMC são tempestades capazes de gerar fenômenos com grande impacto social e econômico, e uma concepção de que são tempestades com correntes ascendentes e descendentes extremamente intensas , capazes de gerar e suportar (em suspensão) granizo gigante, gerar rajadas de vento destrutivas, e amplificar processos de estiramento de vórtices em baixos níveis (NASCIMENTO, 2005).

Ainda segundo Nascimento (2005), eventos de vendavais, granizos e tornados, ainda que relativamente raros em comparação com outros sistemas meteorológicos, representam grande ameaça para atividades importantes como a defesa civil, aviação, agricultura e transmissão e distribuição de energia elétrica. Deste modo, identificar com antecedência de várias horas, condições favoráveis à formação de sistemas convectivos severos é essencial para a difusão apropriada de alertas e antecipação de estratégias que eliminem ou reduzam o impacto negativo destes fenômenos meteorológicos.

3.2 ÍNDICES ATMOSFÉRICOS

Os índices de instabilidade termodinâmicos são construídos baseado no modelo conceitual do comportamento da atmosfera considerando os dados registrados nos perfis atmosféricos (temperatura, umidade e vento) obtidos por radiossondagens (no Brasil, realizadas em alguns aeroportos duas vezes ao dia, às 12 e 00 Z, respectivamente). Estes, com mencionado, são parâmetros preditivos, para área de representatividade do perfil de

radiossondagem, de ocorrência de célula (ou complexos) convectiva e este, normalmente, associado à ocorrência de chuva, granizo, descargas atmosféricas e ventos fortes (ou de rajada). Mourão e Menezes (2006) mostraram que índices de instabilidade, simulados pelo *Regional Atmospheric Modeling System* (RAMS), são correlacionados com o comportamento das tempestades sobre a região do estado do Rio de Janeiro,

Atualmente existem índices termodinâmicos sugeridos por diversos autores e nas cinco subseções subsequente são apresentados o conjunto de índices selecionados, como preditores dos algoritmos de inteligência computacional, na construção dos modelos de previsão de EMC neste estudo.

3.2.1 Índice *K*

Este é definido conforme a Equação 3.1 (GEORGE, 1960):

$$K = (T_{850} + TD_{850}) - (T_{700} - TD_{700}) - T_{500}. \quad (3.1)$$

Indica o potencial para ocorrência de tempestades baseando-se na taxa de variação vertical da temperatura, no conteúdo de umidade na baixa troposfera e na extensão da camada úmida. Onde T e TD referem-se à temperatura do bulbo seco e a temperatura do ponto de orvalho, respectivamente. Os números subscritos indicam o nível de pressão (em hPa) correspondente. Na Tabela 3.1 estão apresentados seus os intervalos de valores de *K* e as possibilidade de atividade convectiva conforme estabelecido por George (1960).

Tabela 3.1 - Valores de K para Tempestades (GEORGE, 1960).

Índice K	**Possibilidade de Tempestades**
K < 20	Sem Atividade Convectiva
20 < K < 25	Tempestades isoladas
25 < K < 30	Tempestades muito isoladas
30 < K < 35	Tempestades esparsas
K > 35	Muitas Tempestades

3.2.2 Índice Showalter

O índice Showalter (IS) é o excesso de temperatura de uma parcela de ar em relação ao ambiente em 500 hPa. A temperatura da parcela é obtida a partir de seu levantamento com início em 850 hPa pela adiabática seca até o nível de condensação por levantamento (NCL) calculado a partir desse mesmo nível e, em seguida, trazida pela adiabática saturada até o nível de 500 hPa. Valores negativos de *IS* indicam a possibilidade de convecção (BLUESTEIN, 1993). Este foi definido por Showalter (1947) pela Equação 3.3:

$$IS = T_{500} - T_{500_{pl850hPa}} . \quad (3.3)$$

A Tabela 3.3 ilustra os valores de Showalter favoráveis à formação de tempestades (SHOWALTER, 1947).

Tabela 3.2 - Valores de índice Showalter (SHOWALTER, 1947).

Índice Showalter (IS)	Possibilidade de Tempestades
IS > +3	Sem Atividade Convectiva
+1 ≤ IS ≤ +3	Possíveis Pancadas de Chuva/Tempestades Isoladas
-2 ≤ IS ≤ +1	Tempestades Prováveis
-6 ≤ IS ≤ -2	Possibilidade de Tempestades Severas
IS ≤ -6	Tempestades Severas Prováveis/Possibilidade de Tornados

3.2.3 Índice de Levantamento

O *IL* foi sugerido por Galway (1956) e é representado conforme Equação 3.2:

$$LI = T_{500} - Tp_{500}\,[°C], \quad (3.2)$$

onde, T representa a temperatura do ambiente e Tp é a temperatura da parcela em 500 hPa. No *IL* a parcela é levantada da superfície. Este é definido na Tabela 3.2:

Tabela 3.3 - Valores do Índice de Levantamento (IL) (GALWAY, 1956).

Índice de Levantamento	Possibilidade de Tempestades
0 ≤ IL ≤ +3	Estável. Possível convecção fraca na presença de forte levantamento ou algum mecanismo forçante.
-3 ≤ IL ≤ 0	Marginalmente Instável
-6 ≤ IL ≤ -3	Moderadamente Instável
-9 ≤ IL ≤ -6	Muito Instável
IL < -9	Extremamente Instável

O *IL* se caracteriza pela diferença de temperatura entre uma parcela de ar levantada adiabaticamente - isto é, sem trocar calor com as vizinhanças - a um nível de livre convecção (definir em nota de rodapé) e a temperatura do ambiente. Esse levantamento é importante para uma instabilidade inicial na camada mais baixa da atmosfera, onde menores valores indicam maior instabilidade atmosférica.

3.2.4 Índice *Total Totals*

O *TT* é descrito matematicamente como na Equação (3.4) por Miller (1972):

$$TT = T_{850} + TD_{850} - 2(T_{500})\,. \tag{3.4}$$

Onde T é a temperatura do bulbo seco e TD a temperatura do ponto de orvalho. O índice *TT* é a soma de dois índices de estabilidade (MILLER, 1972): o *Vertical Totals* (VT), dado pelo *lapse rate* entre os níveis de 850 e 500 hPa, e o *Cross Totals* (CT), dado pela diferença do nível de umidade em 850 hPa e a temperatura em 500 hPa. Este índice é usado para avaliar o entranhamento de ar frio na troposfera média, sendo importante para previsão de eventos severos associados a queda de granizo. A Tabela 3.4 exemplifica os valores de *TT* para tempestades.

Tabela 3.4 - Valores de TT para Tempestades (MILLER, 1972).

Índice Total Totals (TT)	Potencial para Tempestades
40 ≤ TT ≤ 45	Chuvas isoladas, algumas poucas moderadas
46 ≤ TT ≤ 47	Esparsas, pouca chuva intensa
48 ≤ TT ≤ 49	Chuva esparsa moderada, algumas intensas; algumas isoladas severas
50 ≤ TT ≤ 51	Chuva intensa esparsa, algumas severas; tornados isolados
52 ≤ TT ≤ 55	Chuva intensa esparsa a numerosas, poucas a esparsas, alguns tornados
TT > 55	Chuva intensa numerosa, pancadas de chuvas esparsas, tornados esparsos

3.2.5 *Convective Available Potencial Energy* (CAPE)

O CAPE é utilizado para avaliar o potencial da atmosfera para desenvolver convecção em função do aquecimento da superfície. Esse índice quantifica a máxima energia disponível para a ascensão de uma parcela de ar de acordo com a teoria da parcela. Nascimento (2005) relata que, para as Planícies americanas, os valores de CAPE entre 1000 e 2500 Jkg^{-1} são considerados altos; valores acima de 2500 Jkg^{-1} indicam instabilidade acentuada e, acima de 4000 Jkg^{-1}, instabilidade extrema, como exemplificado na Tabela 3.5.

Tabela 3.5 - Valores de CAPE para Tempestades (HOUZE, 1993).

CAPE (Jkg^{-1})	Potencial para Tempestades Severas
1000 < CAPE ≤ 2500	Alto Potencial
2500 < CAPE ≤ 4000	Potencial para instabilidade acumulada
CAPE > 4000	Potencial para instabilidade extrema

Matematicamente pode ser definido por (HOUZE, 1993) na equação 3.5:

$$\text{CAPE} = g \int_{NCC}^{NE} \frac{T_{vp} - T_{va}}{T_{va}} dz \,, \tag{3.5}$$

onde *NCC* é o nível de condensação convectiva, *NE* o nível de equilíbrio, T_{vp} a temperatura virtual da parcela e T_{va} a temperatura virtual do ambiente.

3.4 INTELIGÊNCIA ARTIFICIAL

Inteligência artificial é uma área interdisciplinar que envolve, por exemplo, Neurociências, Matemática, Estatística, Física, Ciência da Computação e Engenharia, focada em desenvolver máquinas e/ou modelos que simulem comportamentos inteligentes, tais como percepção visual, processamento de linguagem natural, reconhecimento de padrões, a realização de diagnósticos e/ou prognósticos de eventos futuros etc.

É um termo geral no sentido de que sob a sua abrangência existem diversas áreas da ciência como Sistemas especialistas, *Machine Learning* (Aprendizado de Máquinas), Robótica, Algoritmos Genéticos, Lógica *Fuzzy*, automação...etc.

Aprendizado de máquina (AM) – É a área dentro da Inteligência artificial em que se faz uso de algoritmos em uma máquina (computador) para que a mesma aprenda, a partir de dados históricos, a realizar tarefas, como prognóstico, classificação e previsão. Aprendizado de máquinas se refere a uma vasta gama de modelos ou técnicas, baseados em diversos algoritmos e em dados históricos. Durante os seus desenvolvimentos, os modelos adquirem conhecimento a partir do ambiente (dados de entrada). Essa aquisição de conhecimento ocorre a partir do treinamento do algoritmo de aprendizagem do modelo que está sendo usado.

Os dados, em geral, estão organizados em: a) Eventos (casos, exemplos) – são as linhas de dados em uma planilha; b) Atributos ou preditores dos eventos; c) Classe ou atributo alvo.

Existem, basicamente, dois tipos de dados: a) qualitativos ou nominais e b) Quantitativos ou numéricos (valores numéricos reais).

Fundamentalmente, pode-se agrupar as tarefas de AM em dois tipos: Preditivas e Descritivas. A seguir, são apresentadas algumas as tarefas de aprendizado de máquinas:

Classificação – a finalidade de uma classificação é prever ou definir a classe de um evento. Em geral, um problema de classificação ocorre quando a variável de saída é categórica. Esta é a tarefa de **AM** mais utilizada nas pesquisas. Por exemplo, ela é usada

para prevenir fraude, prever doenças ou classificar a gravidade de uma doença, prever a classe de um evento meteorológico futuro (nevoeiro ou não-nevoeiro), classificar o risco de incêndios (por exemplo, em um determinado dia o risco será alto, moderado ou baixo), classificar ou prever o tipo de doença que atinge uma plantação...etc.

Regressão – tem a finalidade fazer a previsão de um atributo alvo com valores numéricos reais, como, por exemplo, a previsão de valores de parâmetros meteorológicos, tais como visibilidade, temperatura, velocidade do vento, descargas atmosféricas...etc.

Agrupamento ou **clusterização** – Esta tarefa, a partir de uma população, procura criar grupos de acordo com as características ou atributos dos elementos dessa população (baseado em uma determinada métrica).

Existem diversas técnicas de aprendizado de máquina, como, por exemplo: Redes neurais artificiais; Árvore de decisão; *Random forest*; Aprendizagem bayesiana; Aprendizado baseado em exemplos; Agrupamento não-supervisionado e Máquina de Vetor de Suporte (SVM).

De acordo com Liu *et al.* e Brownlee (2016) existem dois tipos principais de aprendizado da máquina, a aprendizagem supervisionada e a aprendizagem não supervisionada:

- Na aprendizagem supervisionada existe um conjunto prévio de dados inseridos na máquina e as sugestões que serão dadas ao usuário devem ser parecidas com os dados registrados. As informações são usadas para prever um resultado esperado pelo usuário ou para fazer a classificação de elementos usados. A maioria do aprendizado de máquina prático usa aprendizado supervisionado. O aprendizado supervisionado é onde você tem variáveis de entrada (X) e uma variável de saída (Y) (Equação 3.6) e usa um algoritmo para aprender a função.

$$Y = f(X). \tag{3.6}$$

O objetivo é aproximar tão bem a função de mapeamento que, quando você possui novos dados de entrada (X), pode prever as variáveis de saída (Y) para esses dados. Os problemas de aprendizado supervisionado podem ser agrupados em problemas de regressão e classificação:

  - Classificação: Um problema de classificação ocorre quando a variável de saída é uma categoria, como vermelho ou azul ou doença e nenhuma doença.
  - Regressão: Um problema de regressão ocorre quando a variável de saída é um valor real.
- Na aprendizagem não supervisionada não existe um resultado específico esperado. Dessa forma, não é possível prever os resultados do cruzamento das informações. Nesse tipo de aprendizagem os dados são agrupados e os resultados mudam de acordo com as variáveis. No aprendizado não supervisionado existe apenas os dados de entrada (X) e nenhuma variável de saída correspondente. O objetivo do aprendizado não supervisionado é modelar a estrutura ou distribuição subjacente nos dados para aprender mais sobre os dados. Isso é chamado aprendizado não supervisionado, porque, diferentemente do aprendizado supervisionado, não há respostas corretas. Os algoritmos são deixados por conta própria para descobrir e apresentar a estrutura dos dados. Os problemas de aprendizado não supervisionado podem ser agrupados em problemas de *clustering* e associação.
  - *Clustering*: Um problema de *clustering* é onde deseja-se descobrir os agrupamentos inerentes nos dados;
  - Associação: um problema de aprendizado de regras de associação é onde deseja-se descobrir regras que descrevem grandes partes de seus dados.

Em Brownlee (2016) os algoritmos de aprendizado de máquina são descritos como o aprendizado de uma função de destino (f) que melhor mapeia as variáveis de entrada (X) para uma variável de saída (Y), como visto na Equação 3.6. Esta é uma tarefa de aprendizado geral para fazer previsões no futuro (Y), dados novos exemplos de variáveis de entrada (X). A função (f) assim como sua forma, não é conhecida. Há também o erro (e) que é independente dos dados de entrada (X) (Equação 3.7).

$$Y = f(X) + e. \tag{3.7}$$

Esse erro pode ser pode ser devido ao fato de não haver atributos suficientes para caracterizar de maneira adequada o melhor mapeamento de X para Y. Esse erro é

chamado de erro irredutível, porque, por mais que seja bom em estimar a função de destino (f), não podemos reduzir esse erro.

Como já foi mencionado, algoritmos de aprendizado de máquina são técnicas para estimar a função de destino (f) com a finalidade de prever a variável de saída (Y), a partir das variáveis de entrada (X). Diferentes algoritmos de aprendizado de máquina fazem suposições diferentes sobre a forma e estrutura da função que está sendo estimada, bem como a melhor configuração de otimização. Portanto, é importante testar um conjunto de algoritmos diferentes em um problema de aprendizado de máquina, de modo a procurar encontrar a melhor abordagem para estimar a estrutura da função ótima.

Suposições podem simplificar bastante o processo de aprendizado, mas também podem limitar o que pode ser aprendido.

Algoritmos que simplificam a função para uma forma conhecida são chamados algoritmos paramétricos de aprendizado de máquina. Um modelo de aprendizado que resume os dados com um conjunto de parâmetros de tamanho fixo é chamado de modelo paramétrico.

Geralmente, a forma funcional assumida é uma combinação linear das variáveis de entrada e, devido a isto, os algoritmos paramétricos de aprendizado de máquina também são chamados de algoritmos lineares de aprendizado de máquina. Alguns exemplos de algoritmos paramétricos de aprendizado de máquina incluem:

- Regressão logística;
- Análise Discriminante Linear;
- Perceptron.

Vantagens dos algoritmos de aprendizado de máquina paramétricos:

- são mais fáceis de se entender e de interpretar os resultados;
- tem uma aprendizagem mais rápida;
- não exigem, necessariamente, muitos dados de treinamento e podem apresentar bons resultados, mesmo que o ajuste aos dados seja imperfeito.

Limitações dos algoritmos paramétricos de aprendizado de máquina:

- são altamente restritos à forma funcional especificada;
- são mais adequados a apenas problemas mais simples.

Algoritmos que não fazem suposições fortes sobre a forma da função de mapeamento são chamados algoritmos de aprendizado de máquina não paramétricos. Ao não fazer suposições fortes, eles são livres para aprender qualquer forma funcional a partir dos dados de treinamento. Os métodos não paramétricos são bons quando há uma grande quantidade de dados e nenhum conhecimento prévio sobre eles. Os métodos não paramétricos buscam melhor ajustar os dados de treinamento na construção da função de mapeamento. Como tal, eles são capazes de ajustar um grande número de formas funcionais. Um modelo não paramétrico de fácil compreensão é o algoritmo *k-nearest*, que faz previsões com base nos *k* padrões de treinamento mais semelhantes para um novo exemplo de dados. O método não faz suposições sobre a forma da função de mapeamento, exceto os padrões que estão próximos. Exemplos de algoritmos de aprendizado de máquina não paramétricos são: Redes Neurais Artificiais; Árvore de decisão; *Naive Bayes* (Aprendizagem bayesiana) e Máquina de Vetor de Suporte (SVM).

Vantagens dos algoritmos não paramétricos de aprendizado de máquina:

- Flexibilidade: Capaz de ajustar muitas formas funcionais;
- Potência: sem suposições (ou suposições fracas) sobre a função subjacente;
- Desempenho: pode resultar em modelos de previsão de alto desempenho.

Limitações dos algoritmos não paramétricos de aprendizado de máquina:

- Necessita de muitos dados de treinamento para estimar a função de mapeamento;
- Treinamento lento, pois, em geral, possuem muitos parâmetros a serem treinados;

Ainda segundo Brownlee (2016), ajustes excessivos e inadequados podem levar a um desempenho ruim do modelo. O problema mais comum é o ajuste excessivo. O excesso de ajustes leva a uma avaliação dos algoritmos nos dados de treinamento diferente da avaliação dos dados de teste. Duas técnicas são usadas para avaliar algoritmos de aprendizado de máquina para limitar o excesso de ajustes:

- Técnica de reamostragem para estimar a precisão do modelo;
- Reter um conjunto de dados de teste.

A técnica de reamostragem mais usada é a *cross-validation k-fold*[1]. Ela permite que o modelo seja treinado e testado *k-times* em diferentes subconjuntos de dados de treinamento e cria uma estimativa média do desempenho do modelo de aprendizado de máquina sobre dados não trabalhados anteriormente.

Um conjunto de dados de teste é simplesmente um subconjunto dos dados de treinamento que é retido dos algoritmos de aprendizado de máquina. Depois de selecionar e ajustar os algoritmos de aprendizado de máquina no conjunto de dados de treinamento, os modelos aprendidos são avaliados no conjunto de dados de teste para obter uma medida do desempenho final do modelo com dados não vistos.

A implementação desses algoritmos, possui baixo custo computacional, e o método permite seu uso tanto em pesquisa como em um ambiente operacional. O WEKA (*Waikato Enviroment for Knowledge Analysis*), uma plataforma de software livre, desenvolvido pela Universidade de Waikato, Nova Zelândia, será utilizado neste trabalho. O WEKA trabalha com a linguagem Java, assim é possível utilizá-la em diferentes sistemas operacionais.

O WEKA contém ferramentas para pré-processamento de dados, classificação, regressão, agrupamento, mineração de regras de associação e visualização de dados. Dentre as técnicas que podem ser utilizadas, encontram-se: Árvore de Decisão, Classificador Bayesiano, SVM, Conjunto de Regras, Regressão Logística e Linear, MLP e o KNN.

A seguir serão apresentadas as técnicas de aprendizados de máquinas utilizadas neste estudo, assim como os classificadores selecionados e correspondentes a cada aprendizado.

Foi utilizado um conjunto de classificadores da versão WEKA 3.7.12 (HALL *et al.*, 2009). O teste inicial foi realizado com todos os 56 classificadores, porém a seleção foi baseada nos que obtiveram melhor desempenho.

A seguir serão apresentados alguns fundamentos teóricos de algumas das técnicas de aprendizado de maquinas utilizadas.

---

[1] O método do *cross-validation k-fold* consiste em dividir o conjunto total de dados em *k* subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir daí, um subconjunto é utilizado para teste e os *k-1* restantes são utilizados para estimação dos parâmetros, fazendo-se o cálculo da acurácia do modelo. Este processo é realizado *k* vezes alternando de forma circular o subconjunto de teste (KOHAVI, 1995).

3.4.1 Aprendizado baseado em Redes Neurais Artificiais

Uma rede neural é uma ferramenta de inteligência artificial que se destaca no reconhecimento de padrões.

As principais características das Redes Neurais Artificiais (RNA) são a capacidade de aprender e distribuir o aprendizado. Graças a essas características, elas têm sido amplamente aplicadas em diversas áreas de ciências naturais. Os princípios das RNA podem ser encontrados em diversos livros-texto, como por exemplo, Fausett (1994) e Bishop (1995).

A RNA mais amplamente utilizada em diversas aplicações é a *Multilayer Perceptron* (**MLP**) alimentada adiante.

Multilayer Perceptron (MLP) ou rede múltiplas camadas é uma rede alimentada adiante (*feedforward*). Tipicamente, esse tipo de RNA consiste em um conjunto de unidades sensoriais que constituem a camada de entrada; uma ou mais camadas ocultas e uma camada de saída de nós computacionais. O sinal de entrada se propaga para frente por meio da rede, camada por camada (Haykin, 2002). Possui aprendizado supervisionado e tem o objetivo de estimar o erro na camada de saída e retro propagar este no sentido saída-entrada (*backpropagation*), para que seja feito o ajuste dos pesos de todas as camadas, este processo é chamado de retro propagação do erro. Segundo Affonso *et al*. (2010), o algoritmo de *backpropagation* é um tipo de aprendizado supervisionado no qual um valor de saída é gerado, o erro é calculado e, em seguida, seus valores são retro propagados para entrada, os pesos são ajustados e os valores são novamente calculados, até que o conjunto de dados de saídas tenha o menor erro considerado aceitável.

3.4.2 Aprendizado baseado em Árvores de Decisão

As árvores de decisão possuem uma compreensão intuitiva de sua estrutura, diferente de redes neurais, por exemplo, cuja estrutura é mais difícil de interpretar (uma caixa-preta). Além disso, as árvores de decisão podem identificar os atributos mais importantes de um conjunto de dados e excluir os irrelevantes. Essa capacidade de ser seletiva aumenta sua legibilidade humana e pode proporcionar uma melhor compreensão sobre o que é mais significativo em um conjunto de dados (HERMSDORFF, 2018). A

seguir são apresentados os algoritmos utilizados neste trabalho baseados em árvores de decisão.

- ***Hoeffding Tree*** é um algoritmo incremental de indução de árvores de decisão de fluxo que, por meio de um conceito denominado Limite de *Hoeffding*, disponibiliza garantias de desempenho. Esse limite consiste em conceder certo nível de confiança no melhor atributo para dividir a árvore. Foi implementado no algoritmo de árvore de decisão muito rápida (VFDT – *Very Fast Decision Tree*), que inclui aprimoramentos tais como, limitação de nós, introdução de parâmetros de desempate e remoção de atributos ruins. (HULTEN *et al*., 2001).
- ***J48*** é um algoritmo de aprendizagem supervisionada, sendo uma versão adaptada do algoritmo C4.5 de Quinlan (1993). Esse algoritmo percorre todos os atributos de modo a identificar aquele que apresenta o maior ganho de informação (ou seja, maior contribuição para o resultado na árvore). Após identificar esse atributo, ele é definido como um nó na árvore (ou a raiz, caso seja a primeira iteração). Uma das vantagens da aplicação deste algoritmo na tomada de decisão é que o mesmo se mostra adequado para os procedimentos, envolvendo as variáveis qualitativas contínuas e discretas presentes nas bases de dados, permitindo a construção de árvores de decisão que classifica e apresenta em suas ramificações os atributos de maior relevância.
- ***Random Forest***: uma técnica para classificação de dados do tipo aprendizagem supervisionada. É um algoritmo proposto por Breiman (2001). Consiste em uma técnica de agregação de classificadores do tipo árvore de decisão, construídos de forma que sua estrutura seja composta de maneira aleatória. Para determinar a classe de um evento, o método combina o resultado de várias árvores de decisão, por meio de um mecanismo de votação. Ao final cada árvore dá uma classificação, ou um voto para uma classe. A classificação final é dada pela classe que recebeu o maior número de votos entre todas as árvores da floresta. Ou seja, o algoritmo *Random Forest* faz aprendizado por agrupamento, gerando de forma aleatória um *ensemble* de árvores de decisão para obter um resultado melhor e mais estável do que com apenas uma árvore. Possui a característica de “dividir para conquistar”, e isto possibilita ao mesmo algumas características que se destacam de outras técnicas, como:

- Algoritmo mais poderoso comparado com o de apenas uma árvore de decisão;
- Possui boa taxa de acerto quando testado em diferentes conjuntos de dados;
- Evitam sobre ajuste (*overfitting*);
- Menos sensíveis a ruídos;
- Classificação aleatória das árvores sem intervenção humana.

3.4.3 Aprendizado baseado em regressão

Aqui são apresentadas duas formas desse tipo de aprendizado.

Regressão logística - é outra técnica utilizado em aprendizado de máquina. É um dos métodos usado para problemas de classificação binária. Sua denominação é em razão da função usada no núcleo do método, a função logística, também chamada de função sigmoide. Foi desenvolvida para descrever propriedades do crescimento populacional em ecologia, com a possibilidade de aumentar rapidamente, atingindo a capacidade de carga do ambiente. Apresenta uma curva em forma de S que em que qualquer número real é mapeado em um valor entre 0 e 1 (BROWNLEE, 2016).

As suposições feitas pela regressão logística sobre a distribuição e interação nos dados são praticamente as mesmas que as feitas em regressão linear. É usada uma linguagem probabilística e estatística precisa. Ela prevê a probabilidade de um evento pertencer à classe padrão, que pode ser do tipo 0 ou 1. A regressão logística não assume nenhum erro na variável de saída (y).

Apresenta distribuição gaussiana; a regressão logística é um algoritmo linear (com transformação não linear na saída). De modo que assume-se uma relação linear entre as variáveis de entrada e a saída. As transformações de dados de suas variáveis de entrada que melhor expressem essa interação linear podem resultar em um modelo mais preciso.

Como a regressão linear, os resultados do modelo podem se sobrepor se houver várias entradas altamente correlacionadas (BROWNLEE, 2016):

Simple Logistic - é uma regressão logística linear que modela as probabilidades da classe posterior Pr(G = j|X = x) para as classes *j* por meio de funções lineares em *x* e garante que elas somam 1 (um) e permaneça em [0, 1]. Algoritmos de otimização numérica que se

aproximam da solução de máxima verossimilhança iterativamente são usados para encontrar as estimativas. Um desses métodos iterativos é o algoritmo *LogitBoost*. Em cada iteração, ele ajusta um regressor de mínimos quadrados a uma versão ponderada dos dados de entrada com uma variável de destino transformada. Acelerando a indução em árvores do modelo de função logística, apenas do atributo que resulta no menor erro quadrado, chega-se a um algoritmo que executa a seleção automática de atributos. Usa a técnica de reamostragem *cross-validation k-fold* para determinar o melhor número de iterações (SUMNER *et al*., 2005).

3.4.4 Redes Bayesianas

A rede bayesiana é uma outra ferramenta disponível para aprendizado de máquinas. John e Langley (1995) mostram que elas são úteis em tarefas de desempenho, como diagnóstico: lidam explicitamente com questões de incerteza e ruído, que são problemas centrais em qualquer indução de tarefa. Apesar das suposições simplificadoras do classificador bayesiano, diversos experimentos mostraram que ele é competitivo em relação a algoritmos de indução muito mais sofisticados.

O classificador bayesiano fornece uma abordagem simples, com clara semântica, para representar, usar e aprender o conhecimento probabilitístico. O método foi projetado para uso em ambientes supervisionados, tarefas de indução, nas quais o objetivo de desempenho é prever com precisão a classe de eventos de teste e, em quais eventos de treinamento incluem informações da classe. Pode-se ver esse classificador como uma forma especializada de rede bayesiana, porque depende de duas suposições simplificadoras importantes. Em particular, pressupõe que os atributos preditivos sejam condicionalmente independentes, dada a classe, e postula que nenhum atributo oculto ou latente influencia o processo de previsão (JOHN e LANGLEY, 1995).

Um exemplo de um classificador bayesiano é o ***Naïve Bayes***.

Naïve Bayes - É um classificador probabilístico baseado no teorema de Bayes. Os valores estimados são escolhidos com base na análise dos dados de treinamento. Neste classificador, cada uma das características contribui independentemente para a classificação. Os classificadores *Naïvies Bayes* são úteis para conjuntos de dados muito grandes. Apesar de sua simplicidade, esse classificador proporciona bons resultados para problemas complexos. Além disso, eles exigem pequena quantidade de dados de

treinamento para prever os parâmetros de classificação (ZHANG, 2004) (KUMAR & SAHOO, 2012).

4. DADOS E MÉTODO

4.1 ÁREAS DE ESTUDO

A área de estudo é um retângulo com área de 160.908 km^2 (Figura 4.1), correspondente à rota Rio-São Paulo representada pela posição geográfica dos aeroportos Antônio Carlos Jobim (Rio de Janeiro) e Guarulhos (São Paulo). E assim, considerando a representatividade do perfil atmosférico em um círculo de raio de 150 km, segundo a Organização Meteorológica Mundial (OMM), a área foi dividida aqui, para fins de estudo, em duas áreas I e II, conforme mostrado na Figura 4.2. As áreas I e II são, respectivamente, limitadas pelas coordenadas geográficas de latitudes de -24.022 à -22.785 e longitudes -46.430 à -44.854 e latitudes de -24.022 à -22.785 e longitudes -44.854 à -43.300.

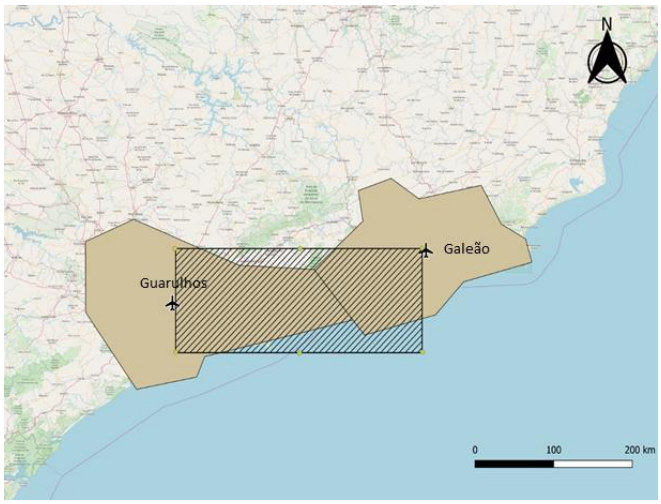


Figura 4.1 - Área de estudo.

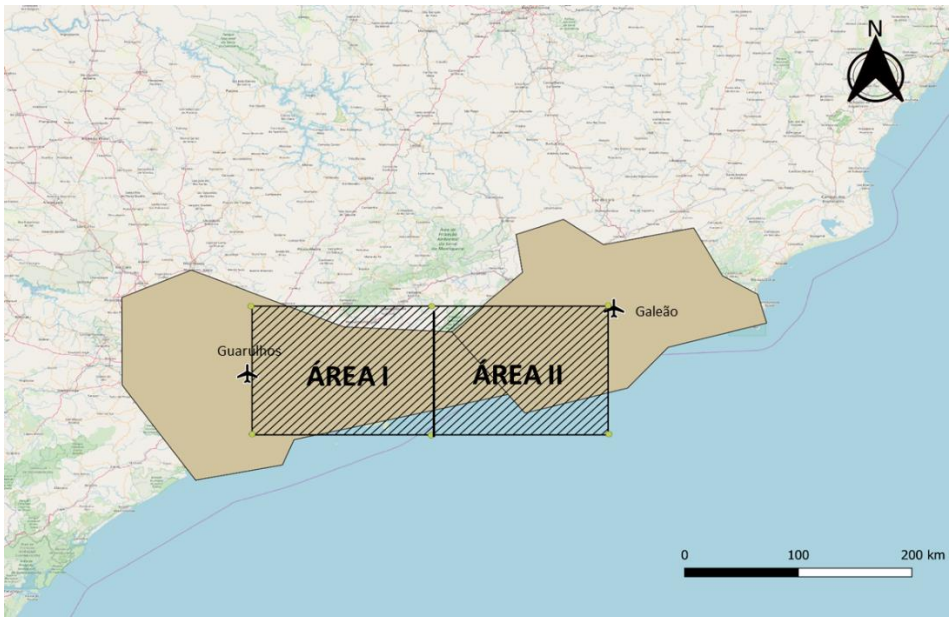


Figura 4.2 - Representação das áreas I e II.

4.2 DADOS

Os dados utilizados se resumem a DA, responsáveis para caracterização dos EMC na fase de maturação e os índices termodinâmicos calculados a partir dos perfis atmosféricos extraídos do sondador atmosférico a bordo de satélite. Na Tabela 4.1 é apresentado sucintamente as características sobre os dados e nas duas seções subsequentes são detalhadas os aspectos técnicos dos dados.

Tabela 4.1 - Dados utilizados no estudo.

Fonte	Frequência	Informação	Período
GOES-R	15 min	Índices termodinâmicos (https://www.ncdc.noaa.gov/airs-web/search)	JAN/FEV/MAR 2018 e 2019
RINDAT	300 ns	Descargas atmosféricas (http://www.rindat.com.br)	2001-2019

4.2.1 Dados de descargas atmosféricas

A Descargas Atmosféricas são registrados continuamente desde 1996, segundo informações disponibilizadas em http://simepar.br/rindat/internas/institucional.shtml), através da Rede Integrada Nacional de Detecção de Descargas Atmosféricas (RINDAT). A RINDAT é um consórcio entre a Companhia Paranaense de Energia (COPEL), Companhia Energética de Minas Gerais (CEMIG) e FURNAS Centrais Elétricas S.A

(ELETROBRÁS FURNAS). Na Figura 4.3 é apresentada a distribuição geográfica dos sensores da rede no território Brasileiro e as instituições, incluindo as cooperadas, proprietária de sensores acoplados a RINDAT, como o Instituto Nacional de Pesquisas Espaciais (INPE) e Sistema Meteorológico do Paraná (SIMEPAR).

A Figura 4.4 representa esquematicamente a detecção das descargas atmosféricas pelos sensores para os tipos de sensores utilizados na RINDAT, processamento, armazenagem dos dados. A RINDAT é capaz de localizar as DA com precisão média de 0,5 km e 2 km metros dentro do perímetro definido pela posição das estações remotas de recepção e de temporização de descargas com resoluções de até 300 nano segundos. Atualmente o sistema é composto de 4 centrais processamento de dados localizadas em Belo Horizonte, Curitiba, Rio de Janeiro e São José dos Campos. Informações detalhadas sobre a RINDAT podem ser encontrada no seguinte endereço: http://www.rindat.com.br/.

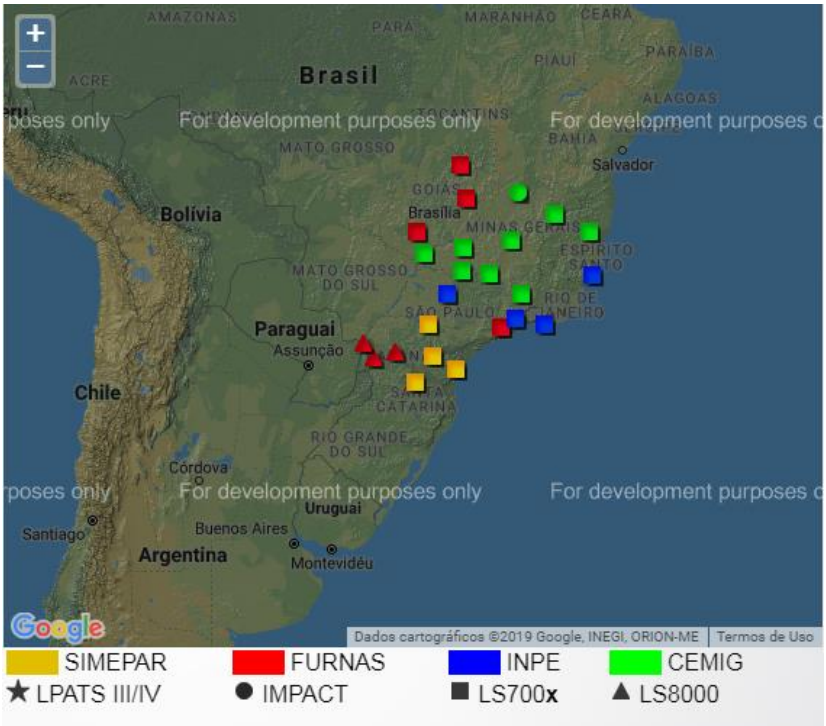


Figura 4.3 - Distribuição no território brasileiro dos sensores da RINDAT e instituição o qual pertence o sensor (Fonte: RINDAT). Maiores informações sobre o sensor disponível em: https://www.vaisala.com/sites/default/files/documents/LS8000-Datasheet-B210422EN-I-LoRes.pdf.

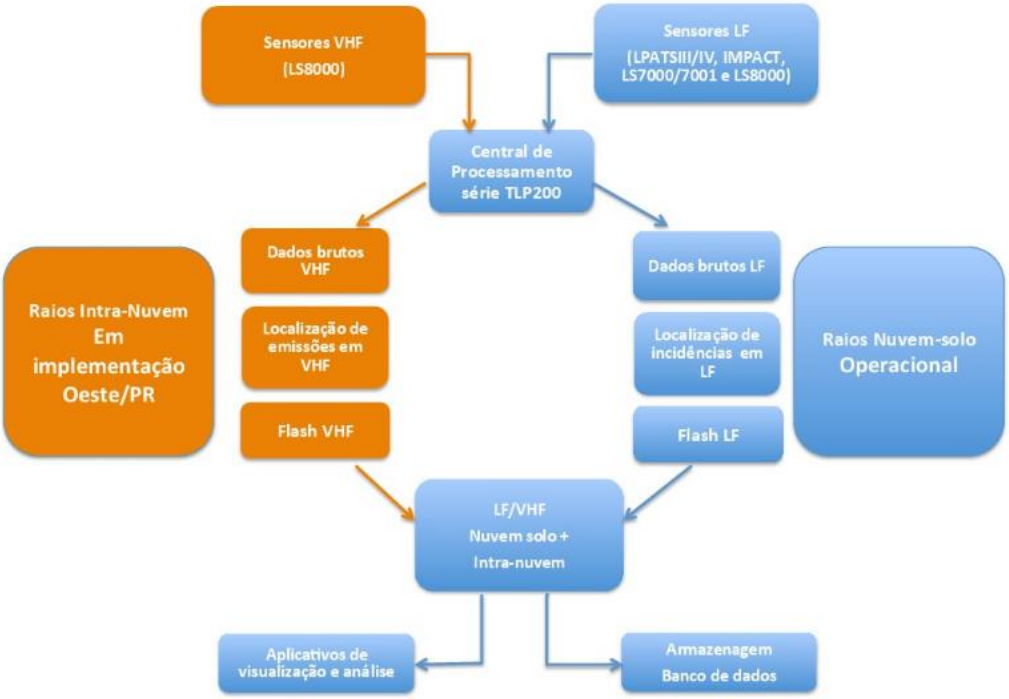


Figura 4.4 - Representação do sistema detecção de descargas atmosféricas, transmissão, processamento e armazenamento dos dados da RINDAT. Fonte: RINDAT.

4.2.2 Dados de índices termodinâmicos – GOES-R

A série de Satélite Ambiental Operacional Geoestacionário (GOES) começou formalmente 16 de outubro de 1975 com o lançamento do GOES-A ou GOES I. Esta série é um esforço conjunto entre a *National Aeronautics and Space Administration* (NASA) e a Administração Nacional Oceânica e Atmosférica (NOAA) para desenvolvimento satélites meteorológicos geoestacionários e forma idealizados 19 (GOES A - U) e com exceção do GOES-Q (que não foi construído) e GOES-T – U planejado para serem lançados 2021 e 2024, respectivamente, os demais foram lançados com sucesso (ver em https://www.nasa.gov/content/goes-overview/index.html).

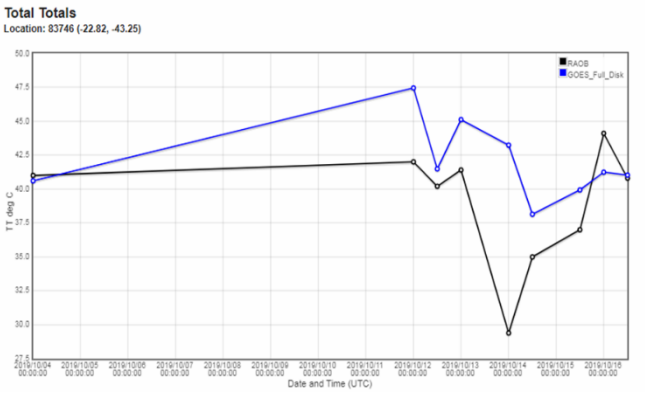
Atualmente o satélite ativo da série é GOES-R (GOES-16) que foi lançado em 18 de dezembro de 2017 e está posicionado a 75,2°W permitindo, assim, uma visão centrada sobre as Américas.

O GOES-R possui um radiômetro denominado *Advanced Baseline Imager* (ABI) com 16 canais (ou banda espectrais) cujo são distribuídos:

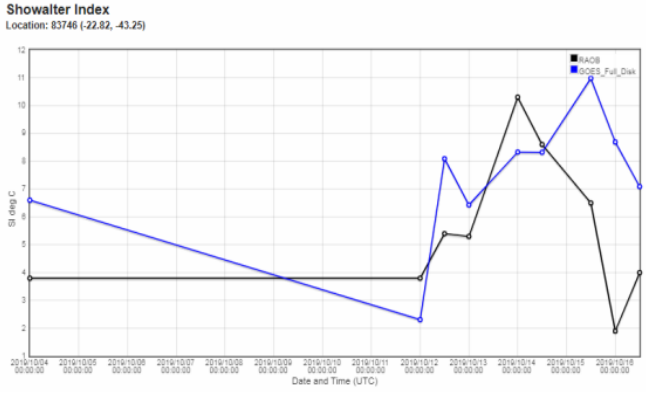
- dois no espectro visível centrados nos comprimentos de onda de 0,47 μm e 0,64 μm (resolução espacial de 1 km e 0,5 km), respectivamente;
- quatro no infravermelho, de resolução de 2 km, próximo centrados em 0,865 μm, 1378 μm, 1,61 μm e 2,25 μm;
- e dez nos infravermelhos, de resolução de 2 km, e centrados em 3,9, 6,9, 6,95, 7,38, 8,5, 9,61, 10,35, 11,2, 12,3 e 13,3 μm.

O grupo de trabalho denominado GOES-R *algorithm* (www.star.nesdis.noaa.gov/goesr/documentation_ATBDs.php) fornece para toda visada do radiômetro ABI, considerando radiação de céu claro, os perfis de temperatura e umidade, índices de água precipitável total e a instabilidade atmosférica através da estimativa de cinco índices de instabilidade a saber: índice levantado (LI), energia potencial disponível convectiva (CAPE), índice de total totals (TT), índice Showalter (IS) e índice K (K), acima mencionados na seção 3.2. Detalhes sobre os experimentos de validação das estimativas de perfis atmosféricos *versus* dados observacionais calculados a partir de dados de radiossondagens, para localidades do território da América do Norte pode ser encontrado em: soundingval.ssec.wisc.edu/station_plots. Considerando que os índices de instabilidades que são variáveis preditoras neste trabalho.

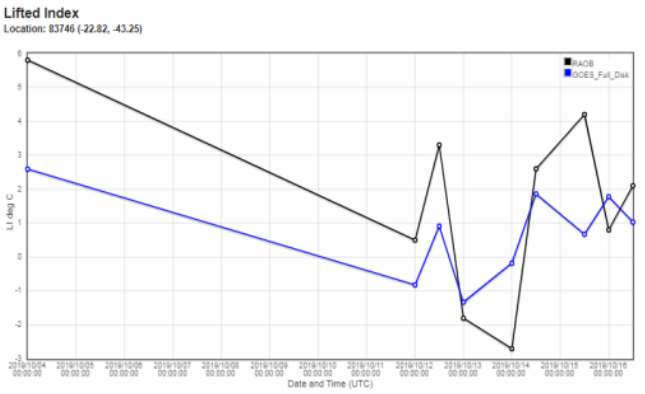
Visando a certificação do uso dos índices GOES-R de instabilidades para região sudeste do Brasil, CAPE disponíveis) estimados dos dados ABI-GOES R *versus* da radiossondagem do Galeão às 12Z para o período de 4 a 10 de outubro de 2019. Na Figura 4.5 (a-e) são apresentados os comportamentos dos cinco índices utilizados, ou seja, (a) *Total Totals*, (b) *Showalter Index*, (c) *Lifted Index*, (d) *K Index*, (e) CAPE. A Tabela 4.2 são apresentados os valores dos vieses, desvio padrão e RMS (*root mean square*) para os índices de instabilidade (TT, IS, LI, K e CAPE) entre os ABI-GOES-R *versus* a radiossondagem do Galeão e o GOES-R. Os valores estatísticos são considerados relativamente baixo e, assim, plausível para uso em áreas desprovidas de radiossondagens na região de interesse deste estudo (rota Rio-São Paulo).



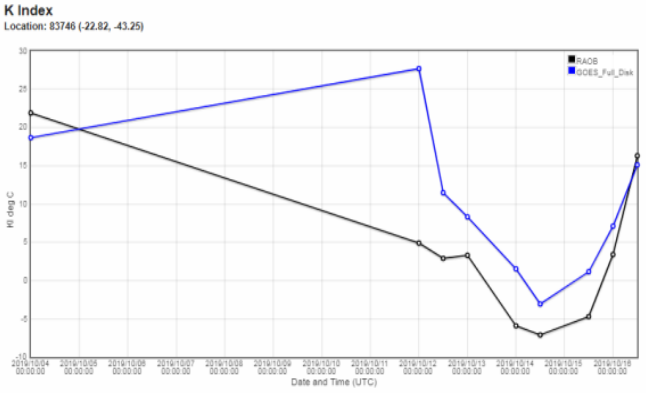
(a)



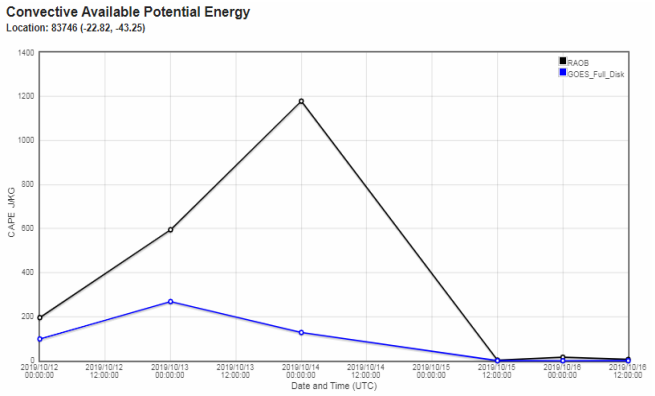
(b)



(c)



(d)



(e)

Figura 4.5 – a-e representam os plotes diários de 12 Z dos índices Total Totals, *Showalter*, *Lifted Index*, K *Index* e CAPE disponíveis, estimados do dados ABI-GOES R (linha azul) versus da radiossondagens do Galeão (RAOB, linha preta) às 12 Z para o período de 4 a 10 de outubro de 2019.

Tabela 4.2 - Estatísticas das diferenças entre os índices ABI-GOES-R e radiossondagens da estação do Galeão às 12Z para o período de 4 a 10 de outubro de 2019.

	CAPE	K	LI	SI	TT
viés	-149,885	4,994	-1,016	1,128	2,695
Desvio Padrão	323,849	6,746	1,656	2,428	3,658
RMS	356,852	8,393	1,943	2,677	4,543

Fonte: http://soundingval.ssec.wisc.edu/station_plots.

4.3 MÉTODO

A seguir será detalhada a metodologia desenvolvida para esse estudo. Esta metodologia está esquematizada através do diagrama de blocos conforme a Figura 4.6, onde são mostrados os passos metodológicos desse estudo.

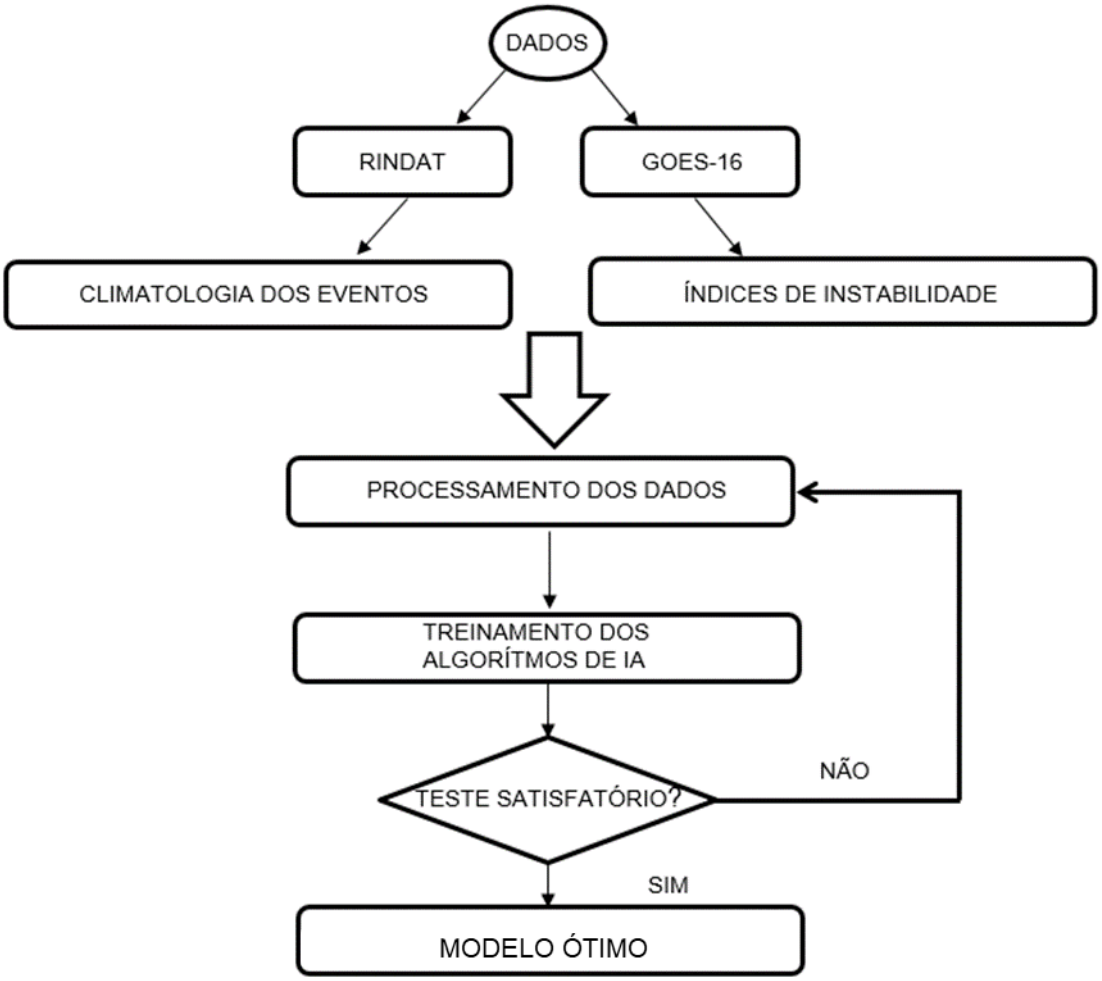


Figura 4.6 - Passos do método.

O método divide-se em três partes com treze passos, a saber:

a) Processamento de dados

i. Análise de consistência dos índices de instabilidade ABI-GOES R na região de estudo;
ii. Cálculo da média dos índices de instabilidade ABI-GOES R dos horários das 12Z, 13Z, 14Z, 15Z e 16Z para áreas estudadas conforme definidas na seção 4.1, para período dos dados, respectivamente;
iii. Análise do comportamento espaço-temporal das descargas atmosféricas com o pacote computacional, software livre, denominado QGIS (versão 3.14.0, disponível em https://qgis.org/en/site/). Define-se o período diário de maior probabilidade de ocorrência de EMC (aqui definido como período de previsão de EMC);
iv. Organização dos registros diário dos índices termodinâmicos médios (variáveis preditoras ou *input* dos modelos a serem desenvolvidos), para cada área, e ocorrência, "SIM", ou "NÃO" de DA (preditante ou *output* dos modelos), nos horários das 12Z, 13Z, 14Z, 15Z e 16Z, conforme um exemplo na Tabela 4.4. A quantidade de DA nos registros é utilizado para classificar a ocorrência (sim ou não) de EMC e sua severidade;
v. Divisão aleatória dos registros em dois conjuntos denominados treinamento (70%) e teste (30%).

Tabela 4.4 - Exemplo do registro diário para um dado horário.

DATA	CAPE	LI	KI	SI	TT	DA	OCORRÊNCIA
01/01/2019	0.31	0.75	28.15	2.69	40.52	244	SIM

b) Treinamento e teste dos algoritmos de previsão de EMC

vi. Experimentos com os algoritmos de inteligência computacionais disponíveis no pacote WEKA (descrito na seção 3.4), considerando o período de previsão definido no passo ii, avaliados via correlação cruzada de dez amostras do conjunto de treinamento, usando as métricas definidas

na seção 4.3.4. Avaliam-se os algoritmos considerando o conjunto de teste, e selecionam-se aqueles de melhor desempenho;

vii. Experimentos com os algoritmos selecionados no passo vi, variando artificialmente os pesos do conjunto de treinamento, para cinco horários definidos em ii, considerando registros de *output*, (vide Tabela 4.4) de "SIM" e "NÃO" em 50% e 50%, 60% e 40%, 70% e 30%, 80% e20%, respectivamente, com avaliação (a exemplo do passo vi) via correlação cruzada de dez amostras do conjunto de treinamento;

viii. Avaliação dos algoritmos treinados considerando o conjunto de teste utilizando as métricas definidas na seção 4.3.4 e define-se o(s) algoritmo(s) ótimo para previsão de EMC;

c) Treinamento e teste dos algoritmos de previsão da severidade dos EMC

ix. Organização do conjunto de treinamento e teste para severidade extraindo registros aleatoriamente nas proporções de 70% e 30% cujo os *output* são "SIM", respectivamente;

x. Divisão do conjunto de treinamento, definido em ix, considerando a distribuição dos EMC versus quantidade de DA (sabe-se que os EMC com baixa quantidade de DA são os mais frequentes) variando de 10 em 10% até 90% os registros e, assim, constroem-se nove os conjuntos de treinamentos onde output será severo ou "SIM" (ou contrário "NÃO") se quantidade de DA do EMC estiver acima do percentil estabelecido para cada conjunto de treinamento;

xi. Treinamento do(s) algoritmo(s) utilizando os nove conjuntos de treinamento definido em x e estabelece-se o limiar de DA de melhor desempenho;

xii. Utilizando-se o limiar DA estabelecido em xi, divide-se a conjunto de teste de severidade; e

xiii. Teste e avaliação do(s) algoritmo(s) e define-se o(s) ótimo para previsão de EMC severo.

4.3.1 Métricas de avaliação

Os algoritmos classificatórios são avaliados nas fases de treinamento e teste considerando as seguintes estatísticas abaixo relacionadas:

Considerando a Tabela 4.5 (tabela de contingência), é possível definir as seguintes estatísticas utilizadas.

Tabela 4.5 – Tabela de contingência 2x2.

		Evento	**Não-evento**	**Total**
Previsto	**Evento**	a	c	a + c
	Não-evento	b	d	b + d
		a + b	c + d	n = a + b + c + d
		Observado		

Fonte: Adaptada de Wilks (2006).

Onde os termos internos da Tabela 4.6 representam:

- a é o número de acertos do evento;
- b é o número de eventos previstos, mas que não foram observados;
- c é o número de eventos que foram observados e não foram previstos;
- d é o número de acertos do não-evento;
- (a+c) é o total de eventos observados;
- (b+d) é o total de não-eventos observados;
- (a+b) é o total de eventos previstos;
- (c+d) é o total de não-eventos previstos;
- n é o tamanho do conjunto ou amostra.

a) POD – Probabilidade de Detecção ou Previsão Correta do Evento (ou verdadeiro-positivo); fornece a taxa de previsão correta do evento desejado. É descrito na Equação 4.1.

$$POD = \frac{a}{a+c}. \tag{4.1}$$

b) POFD ou FAR – Probabilidade de Falsa Detecção ou alarme falso (ou falso-positivo). Corresponde à razão do número de alarme falso dividido pelo total de não eventos observados, como na Equação 4.2.

$$FAR = \frac{b}{b+d}. \tag{4.2}$$

O FAR quanto mais próximo a zero melhor é o resultado.

c) B – Bias é comparação da previsão com a observação. É a razão entre o número de previsões do evento pelo número total de eventos observados. Uma previsão não tendenciosa exibe B = 1. Um valor maior que 1 indica uma previsão superestimada e menor que 1, uma previsão subestimada. Matematicamente é descrito como na Equação 4.3.

$$B = \frac{a+b}{a+c}. \tag{4.3}$$

d) Coeficiente KAPPA - é uma maneira de mensurar o desempenho do processo de classificação. Esse coeficiente pode ser definido como uma medida de associação utilizada para descrever e testar o grau de concordância na classificação, ou seja, sua confiabilidade e precisão (KOTZ & JOHNSON, 1983).

De acordo com Cohen (1960), no cálculo do Coeficiente KAPPA é assumido que:

- As unidades são independentes;
- As classes são independentes e mutuamente exclusivas;
- O classificador e os pontos de referência operam de forma independente.

Segundo Viera e Garrett (2005), o cálculo se baseia na diferença entre o valor observado em comparação com o valor esperado, ou seja, avalia o nível de concordância entre dois conjuntos de dados. O coeficiente KAPPA é uma medida padronizada para se situar em uma escala de -1 a 1, onde 1 é perfeita concordância, 0 é exatamente o que seria esperado por acaso, e valores negativos sugerem que a concordância encontrada foi menor daquela esperada por acaso. O Coeficiente KAPPA pode ser calculado pela Equação 4.4.

$$k = \frac{p_0 - p_e}{1 - p_e}, \tag{4.4}$$

onde:

p_0 é a taxa de aceitação relativa (Equação 4.5)

$$p_0 = \frac{\mathrm{a} + \mathrm{d}}{\mathrm{a} + \mathrm{b} + \mathrm{c} + \mathrm{d}}, \tag{4.5}$$

e p_e é a taxa hipotética de aceitação (Equação 4.6)

$$p_e = \frac{[(a+b)(a+c)]+[(c+d)(b+d)]}{(a+b+c+d)^2}. \quad (4.6)$$

A Tabela 4.6 mostra os valores de interpretação do valor de KAPPA obtido.

Tabela 4.6 - Interpretação do Coeficiente KAPPA.

Coeficiente KAPPA	Nível de Concordância
< 0	Não existe concordância
0,01-0,20	Concordância mínima
0,21-0,40	Concordância razoável
0,41-0,60	Concordância moderada
0,61-0,80	Concordância substancial
0,81-1,0	Concordância perfeita

Fonte: VIERA e GARRETT (2005).

e) F-*Measure* - É uma medida da precisão de um teste. Ele considera a precisão e a revocação do teste para calcular a pontuação. O F-*Measure* é a média harmônica da precisão e revocação[2] (Equação 4.7), onde atinge o seu melhor valor em 1 (precisão perfeita) (SASAKI, 2007)

$$F = \frac{2P.POD}{P + POD}, \quad (4.7)$$

onde *R* ou POD e *P* representam revocação e precisão, respectivamente. E *P* pode ser descrito como na Equação (4.8)

$$P = \frac{a}{a + b}. \quad (4.8)$$

Com o objetivo de padronizar esses valores, será utilizado a seguir 1-FAR. Dessa forma, as métricas estatísticas de avaliação com valores iguais a 1 são as consideradas ideais.

[2] Na classificação binária, a revocação é chamada de sensibilidade.

5. RESULTADOS

Como apresentado na seção 4.3, a seguir serão mostrados os resultados em cada passo metodológico mencionado.

5.1 CLIMATOLOGIA DOS RAIOS

Na Figura 5.1 (a)-(d) é apresentada a variação espaço-sazonal, por meio de mapa de calor, da densidade das DA para área de estudo. A Figura 5.2 e Figura 5.3 representam, respectivamente, o histograma sazonal e a média mensal das descargas atmosféricas para período de 2001 a 2019. Pode-se observar que o verão e o inverno são, respectivamente, os períodos de maior (58,8% dos eventos) e menor (4,0% dos eventos) ocorrências de DA (conforme observado Paulucci *et al*., 2017) e que há, durante o verão, duas regiões preferencialmente de maior densidade de DA (conforme Figura 5.2 (b)), que são a região próxima ao município do Resende (devido possivelmente ao efeito orográfico) e a grande São Paulo (devido à ilha calor).

Na Figura 5.4 é representado o número de ocorrência de DA *versus* horário do dia. Nota-se que a atividade convectiva se inicia a partir das 15 horas e tem seu pico às 18 h.

Considerando o estudo realizado por Paulucci *et al*. (2017) sobre a ocorrência de descargas atmosféricas na região metropolitana do Rio de Janeiro, pode-se induzir que os eventos convectivos severos e duráveis podem começar no período da tarde e se estender ao longo da noite, e também no poço. o trabalho desenvolvido Almeida *et al*. (2020) de previsão do CME aplicado à região mencionada definiu como período de previsão o intervalo das 18h às 0 e, e aqui se assume que 66,0% do EMC estão registrados.

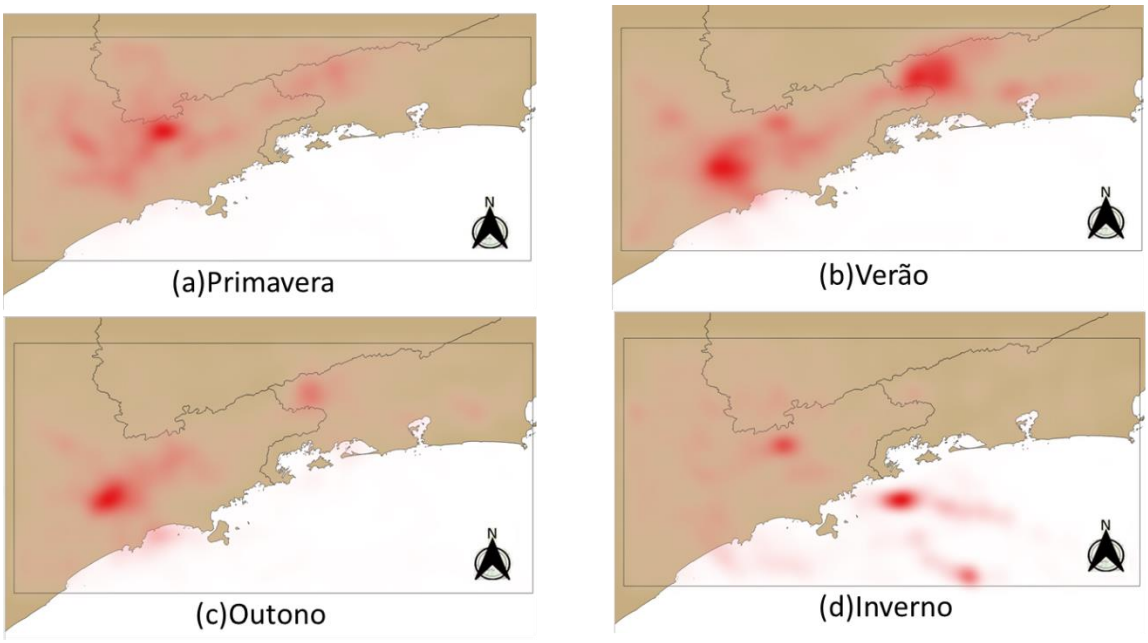


Figura 5.1 - Mapa de calor sazonal das descargas atmosféricas para área de estudo para (a) primavera, (b) verão, (c) outono, (d) inverno para o período de 2001 a 2019.

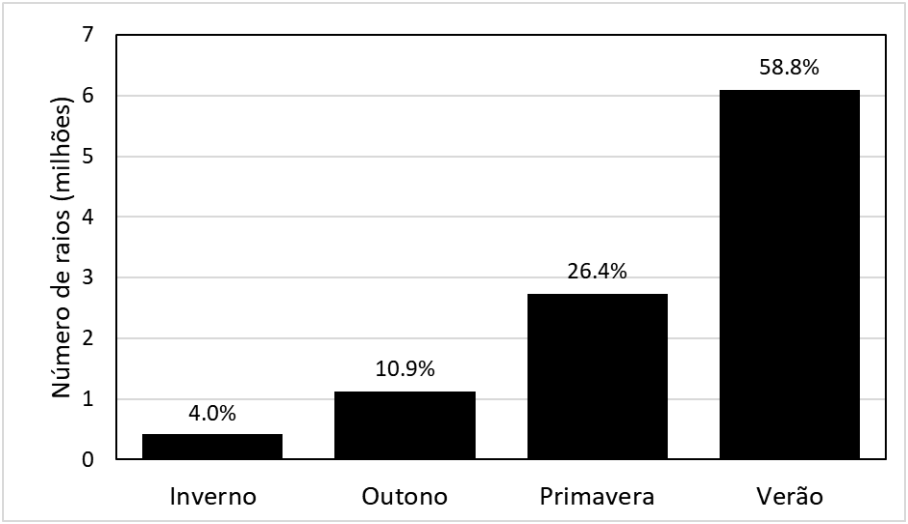


Figura 5.2 - Número de DA (milhões) por estação do ano.

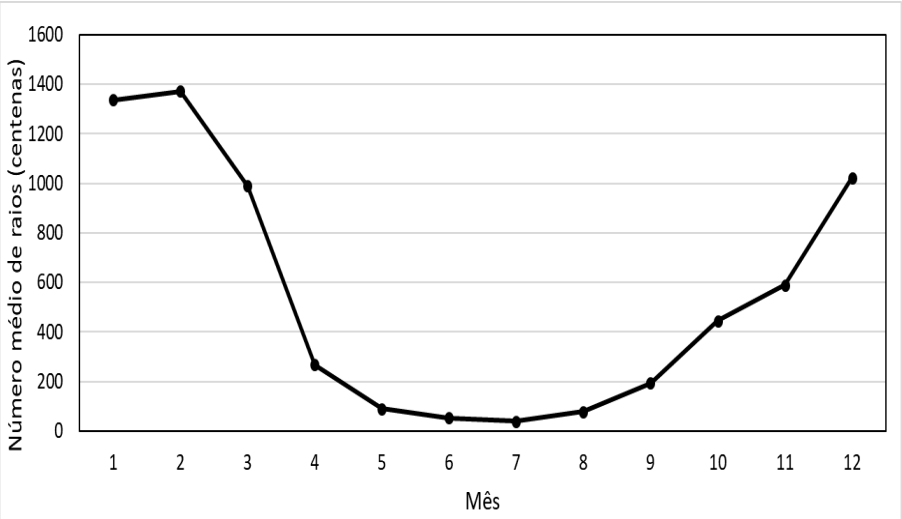


Figura 5.3 - Número médio de DA (centenas) por mês.

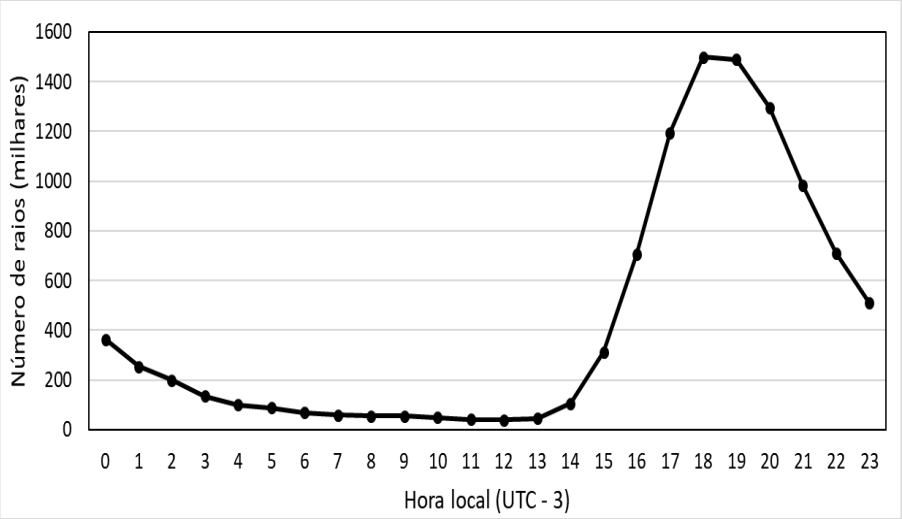


Figura 5.4 - Número de DA (milhares) por hora local.

A Figura 5.5 representa a distribuição de 409 EMC por seus respectivos intervalos de AD nos períodos de janeiro a março de 2018 e 2019 e 79,4% e 20,6% do CME são compostos de AD superior ao limite de 130 (classificado aqui como EMC severo) e iguais ou abaixo deste limite de eventos não graves, respectivamente.

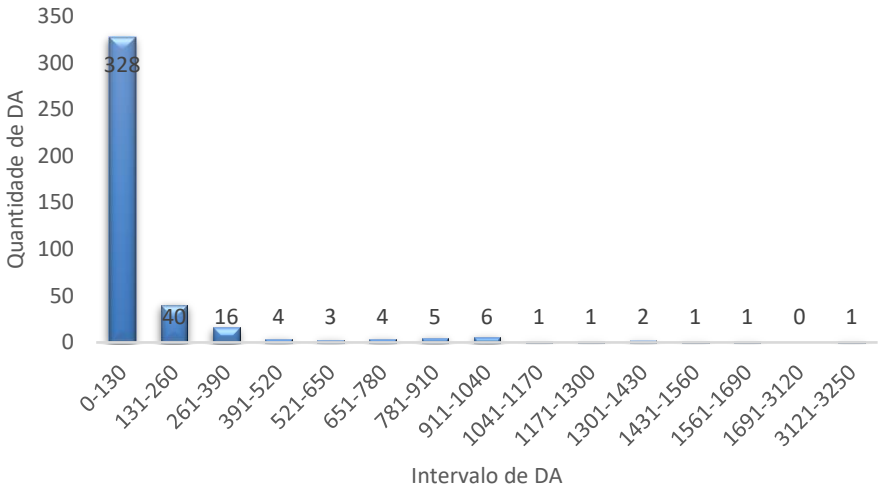


Figura 5.5 - Distribuição de 409 eventos meteorológico convectivos e seus intervalos de descargas atmosféricas para os períodos de janeiro a março de 2018 e 2019.

5.2 COMPARAÇÃO ENTRE OS ÍNDICES TERMODINÂMICOS DAS ÁREAS I E II

Nas Figuras 5.6 à 5.10 são apresentados os valores dos os índices termodinâmicos utilizados (CAPE, LI, KI, SI e TT) das áreas de estudo I (em cor azul) e II (em cor laranja) para os períodos de janeiro a março de 2018. Nota-se que os comportamentos dos índices são similares para ambas as áreas, no entanto, há diferenças significativas, em alguns dias, justificando, assim, a divisão da área da rota aérea (Rio - SP) em duas iguais, conforme especificado anteriormente.

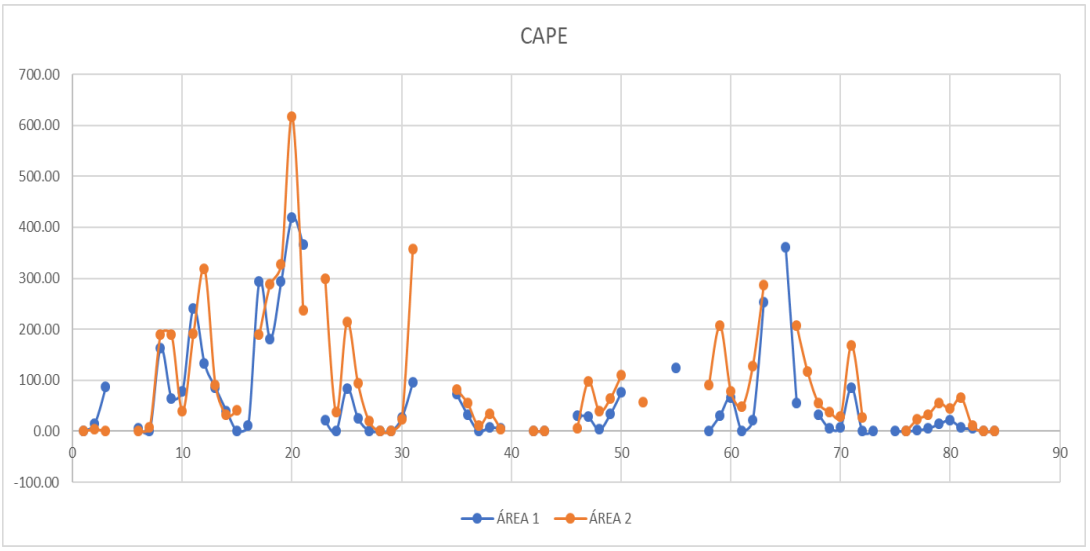


Figura 5.6 – Valores CAPE médio para a área I (azul) e área II (laranja) no período de janeiro a março de 2019.

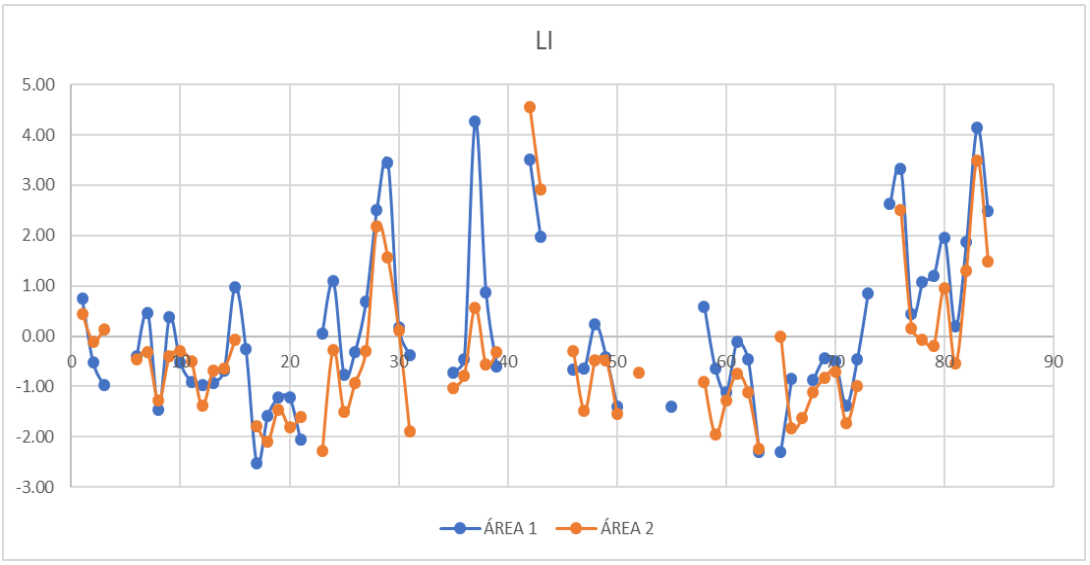


Figura 5.7 – Valores LI médio para a área I (azul) e área II (laranja) no período de janeiro a março de 2019.

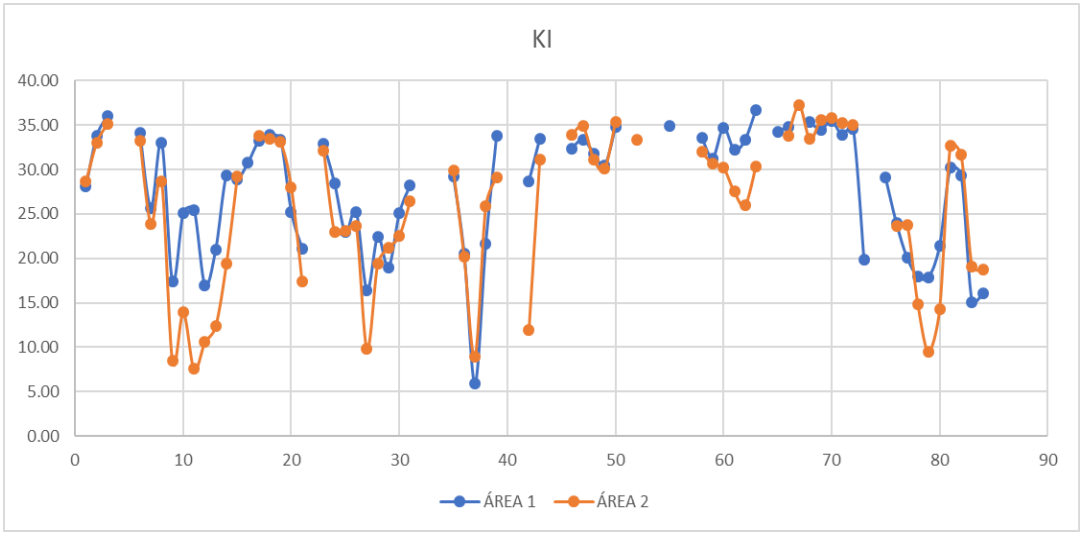


Figura 5.8 – Valores KI médio para a área I (azul) e área II (laranja) no período de janeiro a março de 2019.

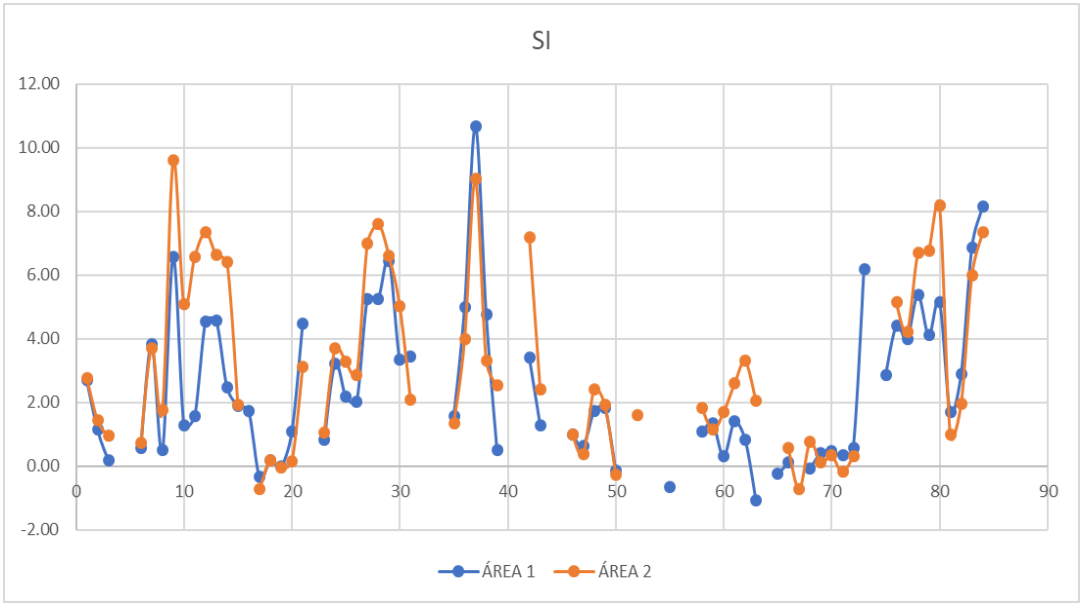


Figura 5.9 – Valores SI médio para a área I (azul) e área II (laranja) no período de janeiro a março de 2019.

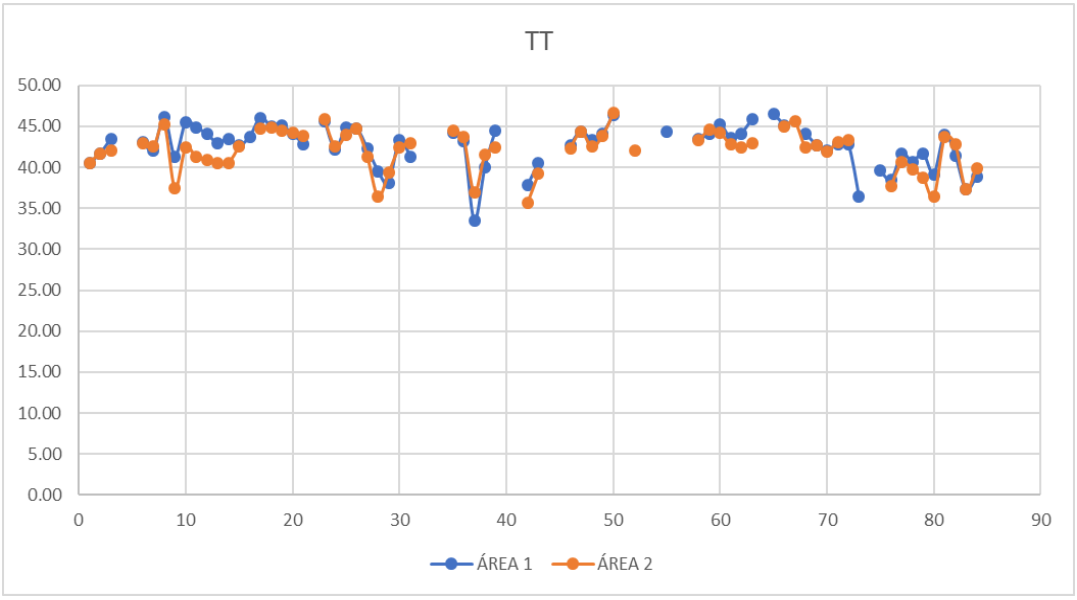


Figura 5.10 – Valores TT médio para a área I (azul) e área II (laranja) no período de janeiro a março de 2019.

5.2 TREINAMENTO E TESTE

É importante salientar que na mineração dos dados, cinquenta e seis algoritmos disponíveis no pacote WEKA foram inicialmente utilizados e devido ao desempenho, cinco algoritmos foram selecionados como mencionado na sessão 3.4.

Conforme método descrito em 4.3, os cinco algoritmos selecionados foram treinados (via correlação cruzada com divisão da população dos dados em dez amostras aleatórias) considerando cinco configurações de conjuntos de treinamento, a saber: (1) dado originais; (2) balanceado artificialmente com 50% e 50%; (3) 60% e 40%; (4) 70% e 30%; (5) 80% e 20% de valores de sim (EMC) e não (não-EMC), respectivamente, para previsão com dados de média de índice de instabilidades (*input*), por área, para às 12Z, 13Z, 14Z, 15Z e 16Z.

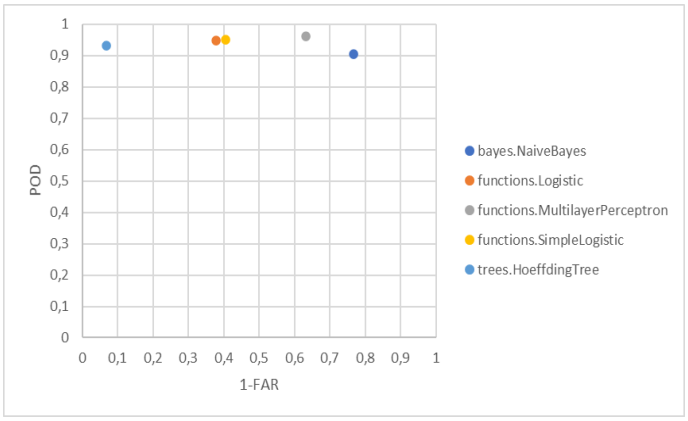
Em sequência são apresentados os resultados dos treinamentos (com 70% da população dos dados e teste (realizado com 30% dos dados) por horário de dados de entrada conforme horários e estudo de caso (*hand cast* do mês de abril de 2019).

5.2.1 12Z

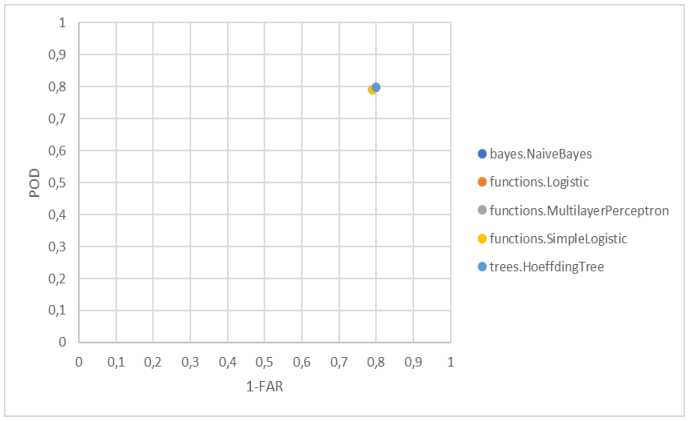
5.2.1.1 Área I

Na Figura 5.11 (a)-(e) é, respectivamente, apresentado os valores das estatísticas de POD *versus* 1-FAR (para correlação cruzada, conforme descrito em 4.3) para os cinco conjuntos de treinamento. Considerando que os valores ideais das estatísticas utilizadas é um, observa-se que os resultados de treinamentos são similares – com exceção daqueles utilizando com os dados na forma original (1) – e o de melhor desempenho é o denominado *Simple Logistic* para os dados na configuração de treinamento (2).

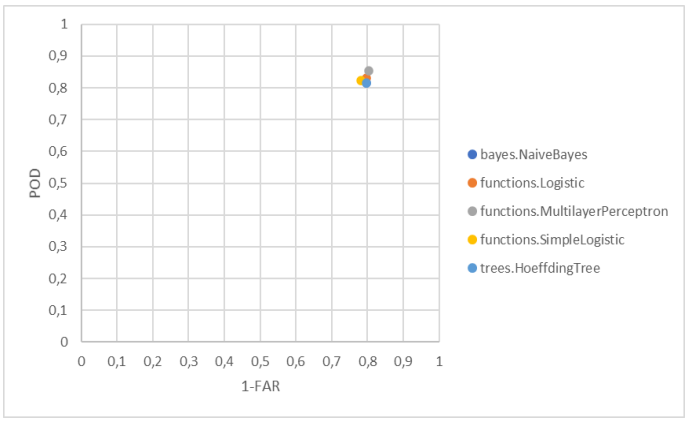
O mesmo se aplica para o teste, que quando balanceados mostram um melhor desempenho. A melhor performance também foi do classificador *Simple Logistic* de configuração (2) (Figura 5.12 (b)), chegando a 85% de POD e 76% de 1-FAR.



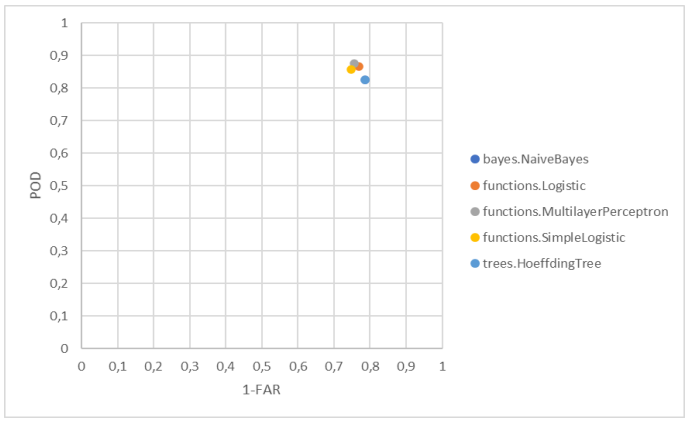
(a)



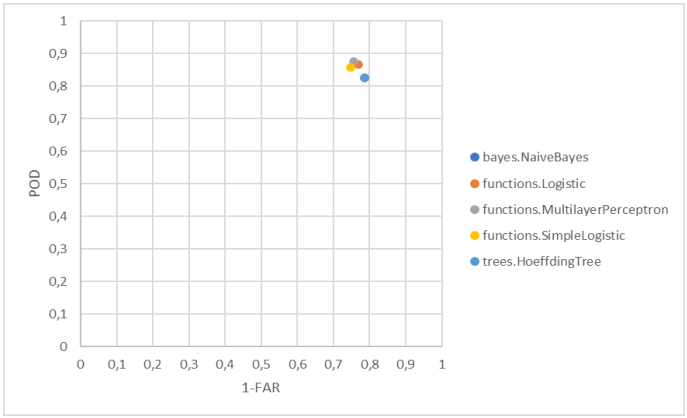
(b)



(c)

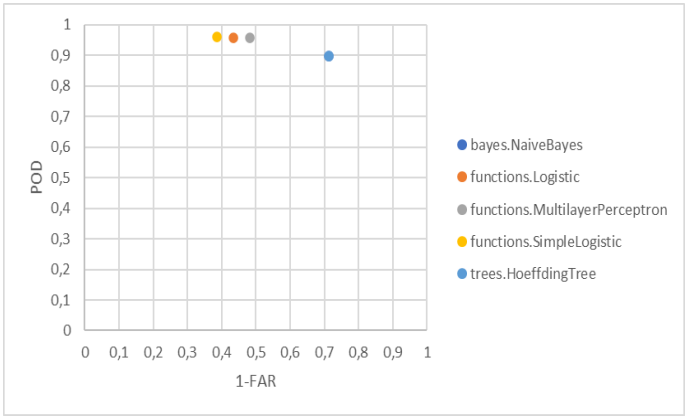


(d)

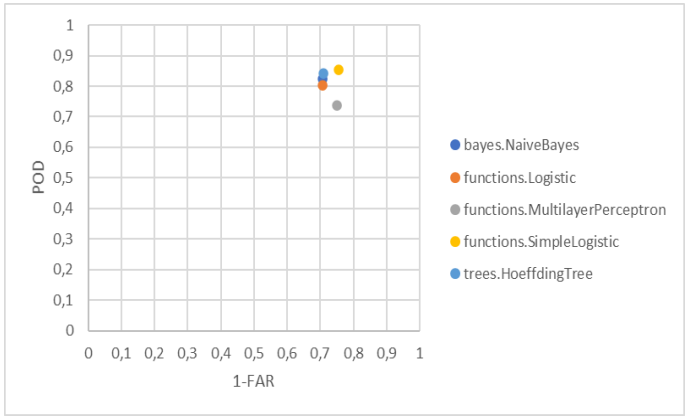


(e)

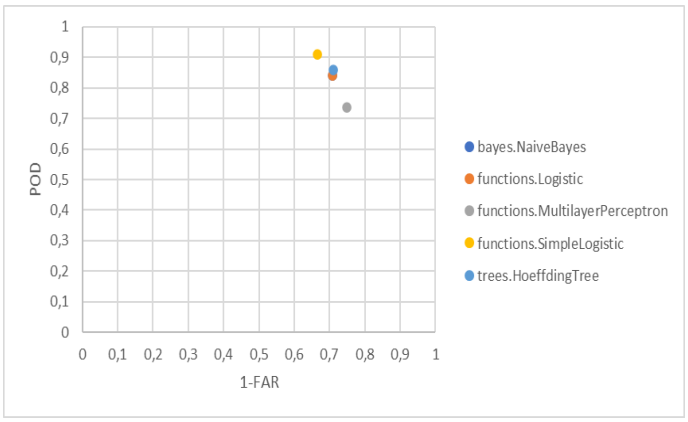
Figura 5.11 - Valores de POD *versus* 1-FAR de treinamento para cinco configurações de dados, para previsão às 12Z, considerando os cinco algoritmos utilizados na área I.



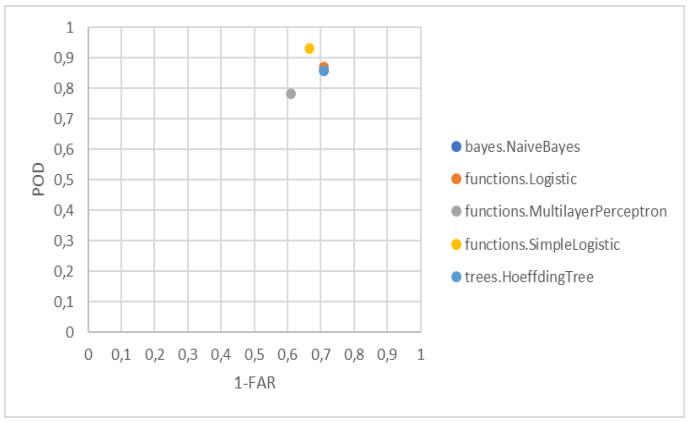
(a)



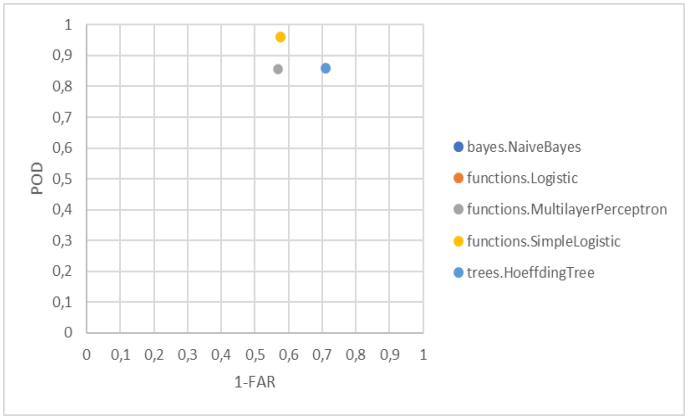
(b)



(c)



(d)



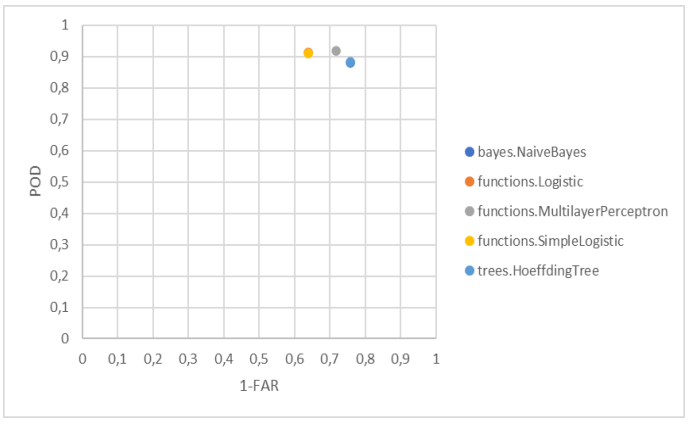
(e)

Figura 5.12 - Valores de POD *versus* 1-FAR de teste para cinco configurações de dados, para previsão às 12Z, considerando os cinco algoritmos utilizados na área I.

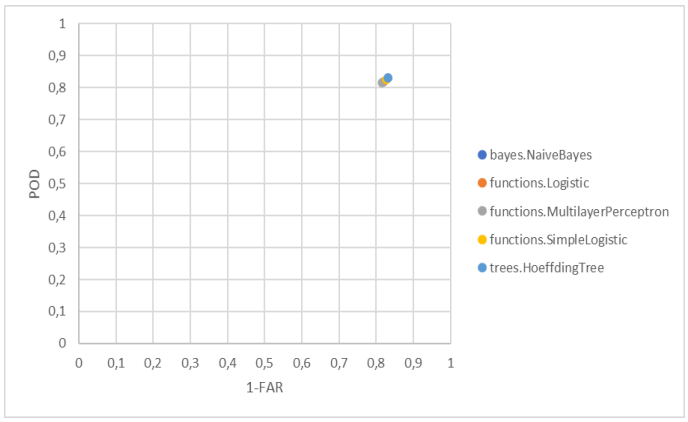
5.2.1.2 Área II

As Figuras 5.13 e 5.14 representam a área II para o treinamento e teste, respectivamente. Da mesma forma que na área I, os modelos com a forma original (1) mostraram performances inferiores aos dados balanceados. O melhor treinamento foi dos classificadores *Naive Bayes* e *Hoeffding Tree* com 83% de POD e 1-FAR para os dados balanceados de configuração (2) (Figura 5.13(b)). É notório que os classificadores para a área II, apresentam melhor desempenho. Isso se mantém para os demais horários, como será mostrado.

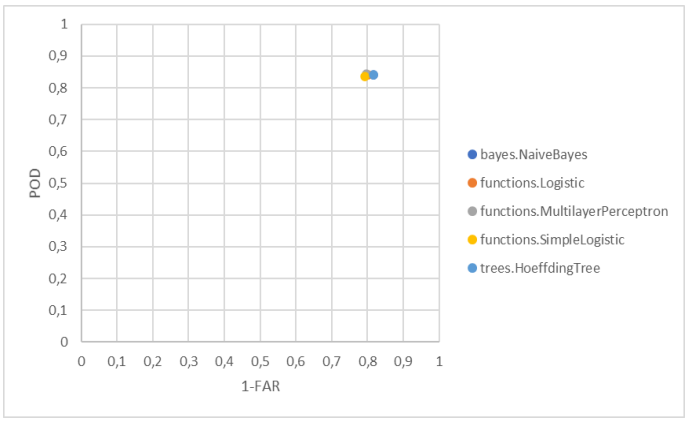
Para o teste (Figura 5.14) o melhor resultado foi do classificador *Simple Logistic* (83% de POD e 1-FAR) para dados balanceados com configuração (2) (Figura 5.14(c)).



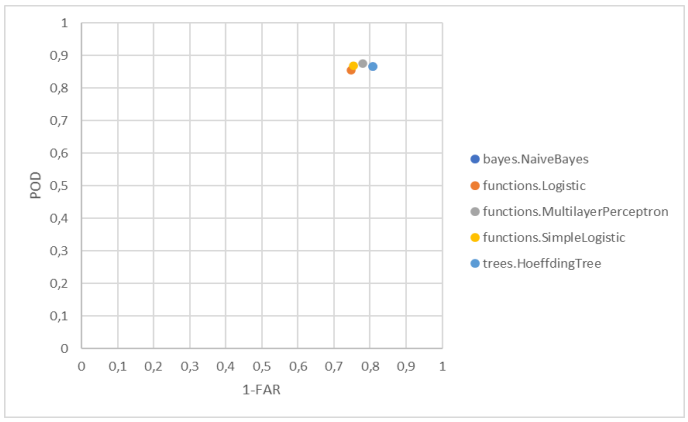
(a)



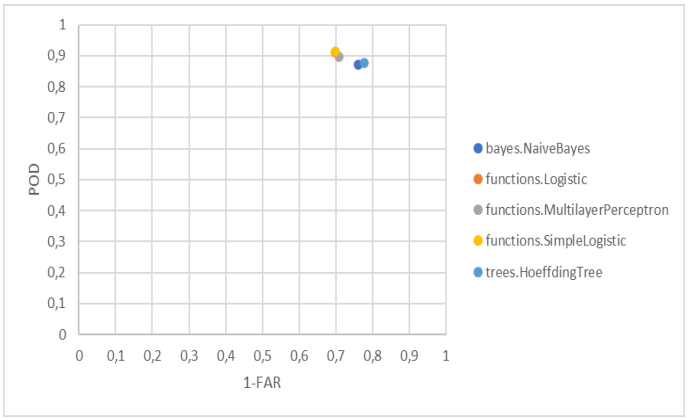
(b)



(c)

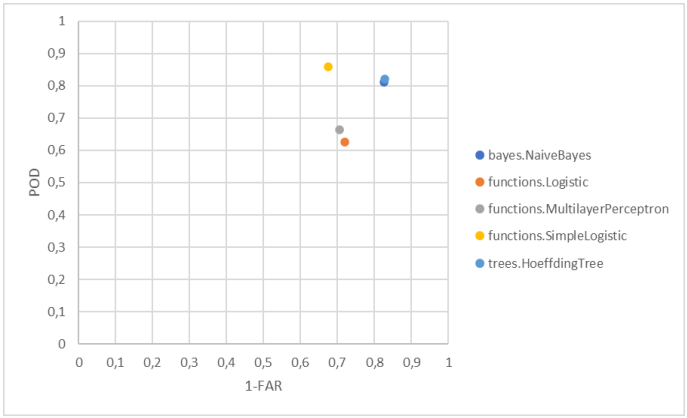


(d)

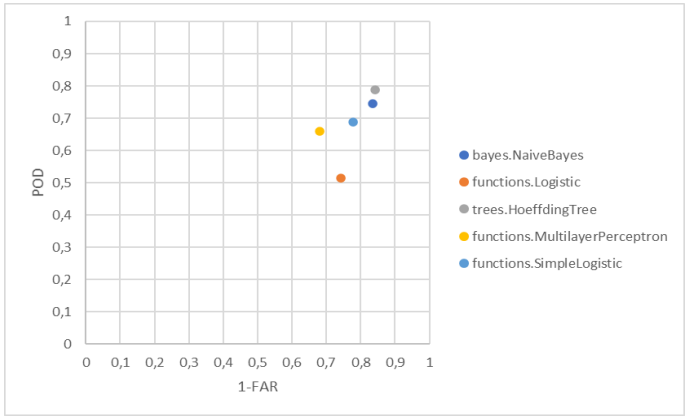


(e)

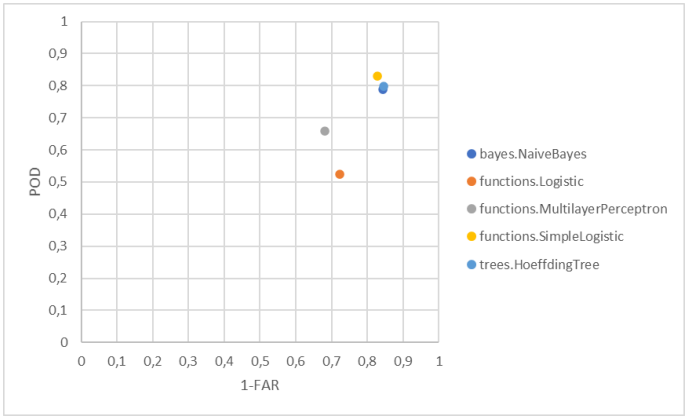
Figura 5.13 - Valores de POD versus 1-FAR de treinamento para cinco configurações de dados, para previsão às 12Z, considerando os cinco algoritmos utilizados na área II.



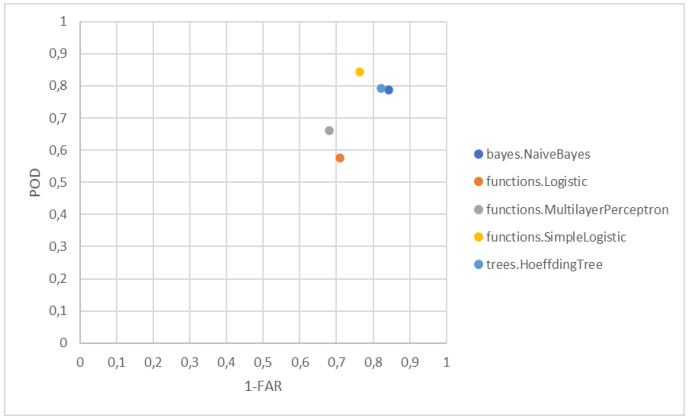
(a)



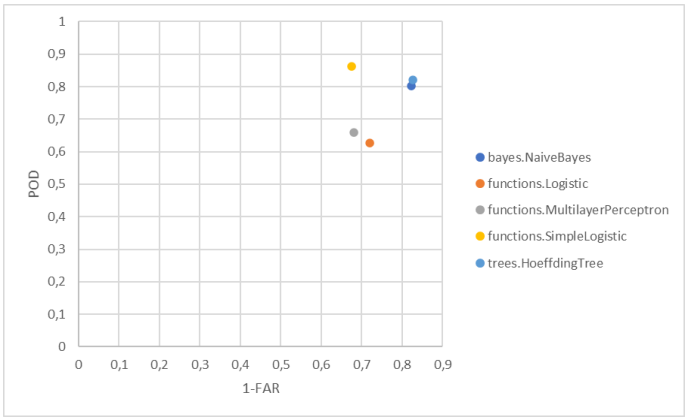
(b)



(c)



(d)



(e)

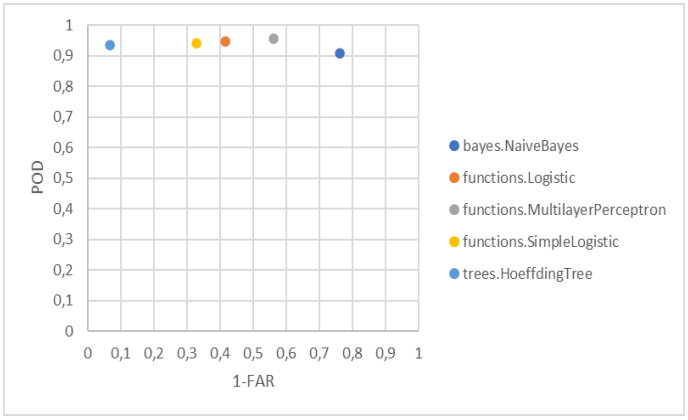
Figura 5.14 - Valores de POD versus 1-FAR de teste para cinco configurações de dados, para previsão às 12Z, considerando os cinco algoritmos utilizados na área II.

5.5.2 13Z

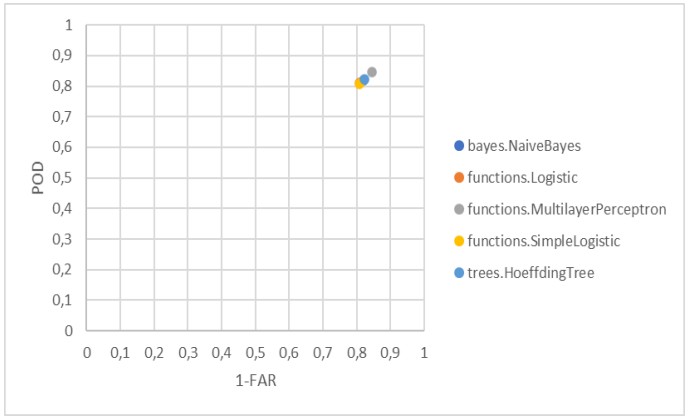
5.5.2.1 Área I

As Figuras 5.15 e 5.16 representam o treinamento e teste, nessa ordem, para a área I às 13Z. A Figura 5.15 com os dados originais do treinamento (1) tem comportamento muito semelhante ao das 12Z. Altos índices de POD e baixos valores de 1-FAR. Apenas o classificador *Naive Bayes* é considerado aceitável para boas predições. Com o balanceamento dos dados, o desempenho é acurado, sendo o *Multilayer Perceptron* (85% de POD e 1-FAR) configuração (2) (Figura 5.15(b)) o melhor classificador.

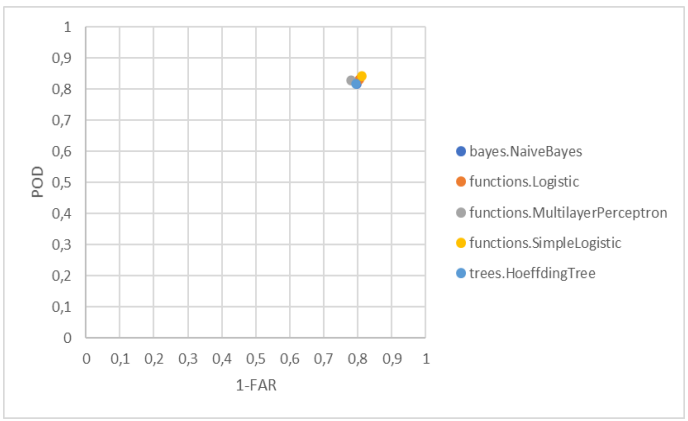
Para o teste (Figura 5.16), o classificador *Logistic* foi o que apresentou melhor desempenho com os dados balanceados na configuração (4) (Figura 5.16(d)).



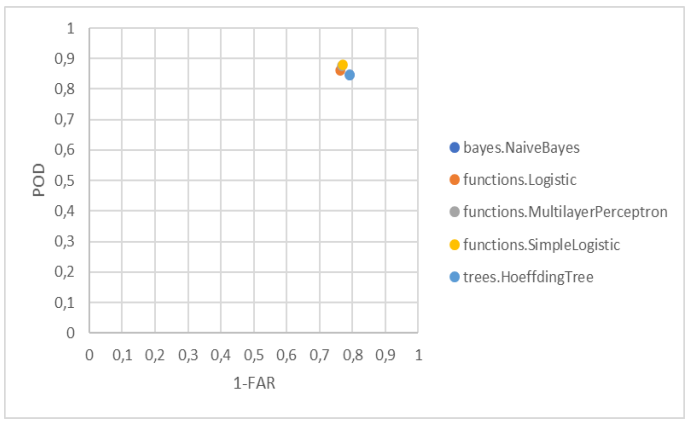
(a)



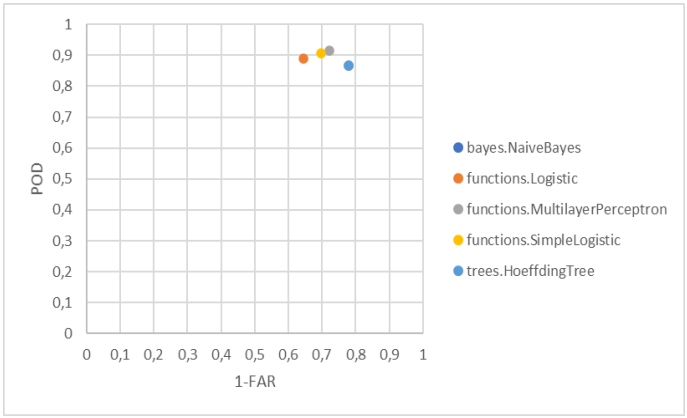
(b)



(c)

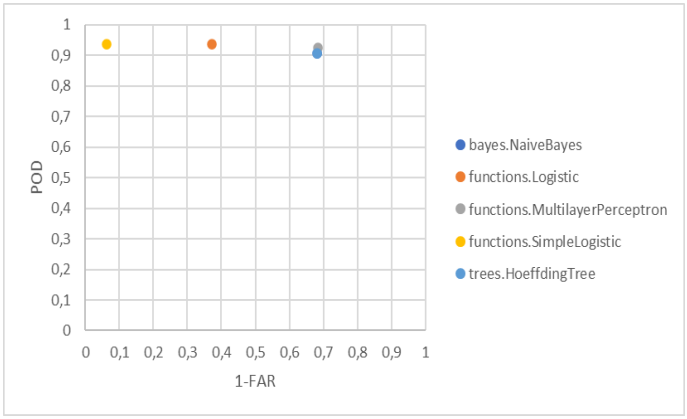


(d)

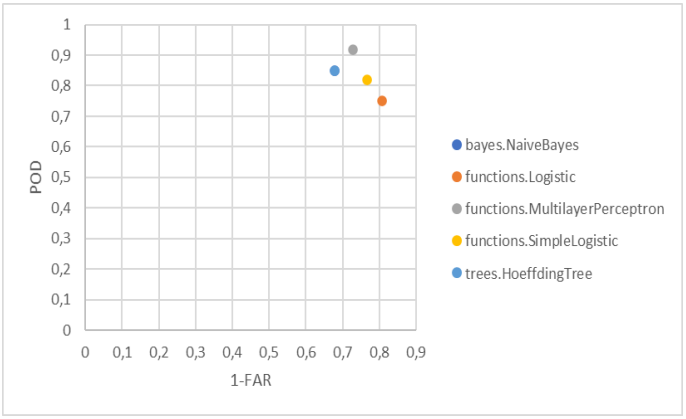


(e)

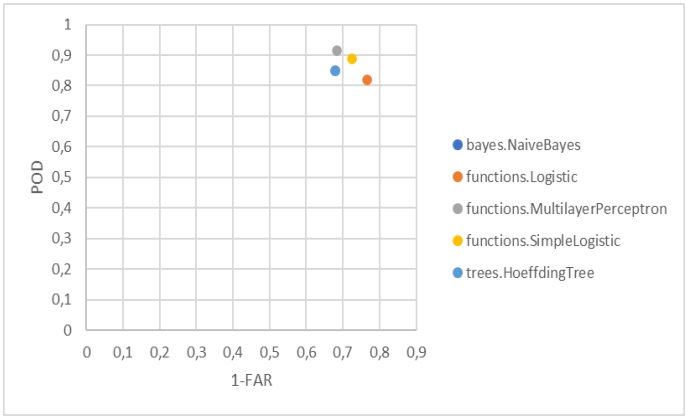
Figura 5.15 - Valores de POD versus 1-FAR de treinamento para cinco configurações de dados, para previsão às 13Z, considerando os cinco algoritmos utilizados na área I.



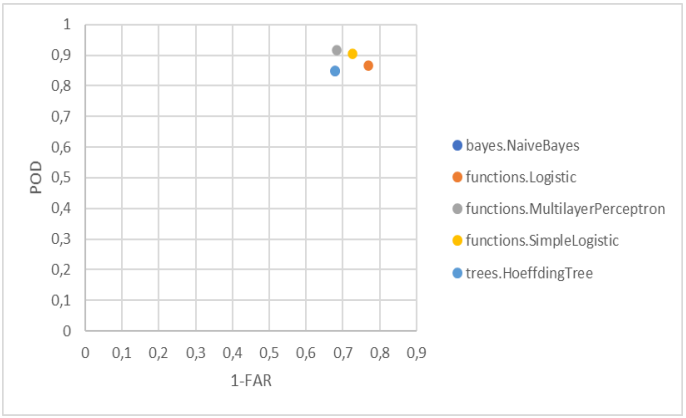
(a)



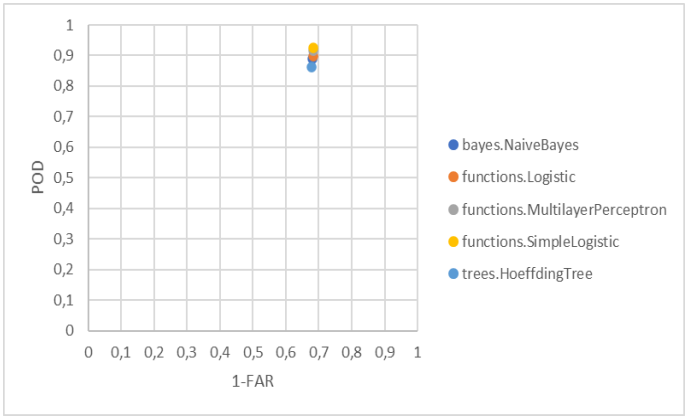
(b)



(c)



(d)



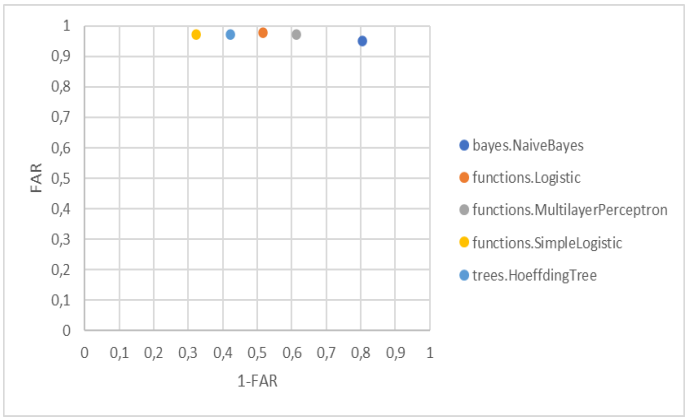
(e)

Figura 5.16 - Valores de POD versus 1-FAR de teste para cinco configurações de dados, para previsão às 13Z, considerando os cinco algoritmos utilizados na área I.

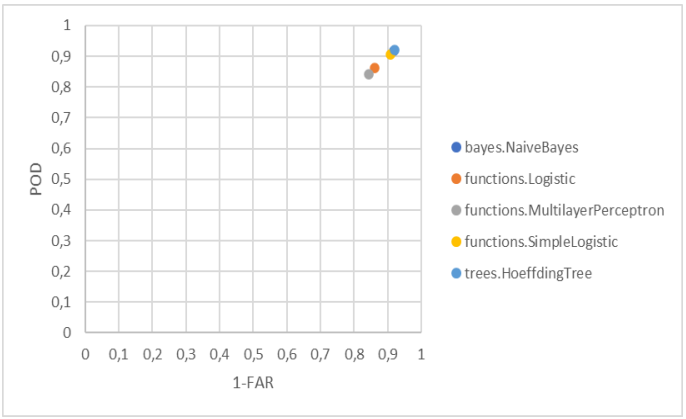
5.5.2.2 Área II

A área II é apresentada pelas Figuras 5.17 e 5.18 para o período das 13Z. De forma já mencionada para os casos anteriores de treinamentos, os dados originais (1) sempre apresentam valores elevados de POD, porém com alto índice de falso alarme. Com o balanceamento dos dados, esse falso alarme é corrigido e há uma melhora substancial. Todos os classificadores apresentam predições com altos valores de acerto, sendo o *Simple Logistic* (93% de POD e 92% de 1-FAR) para a configuração (4) (Figura 5.17(d)) o melhor classificador.

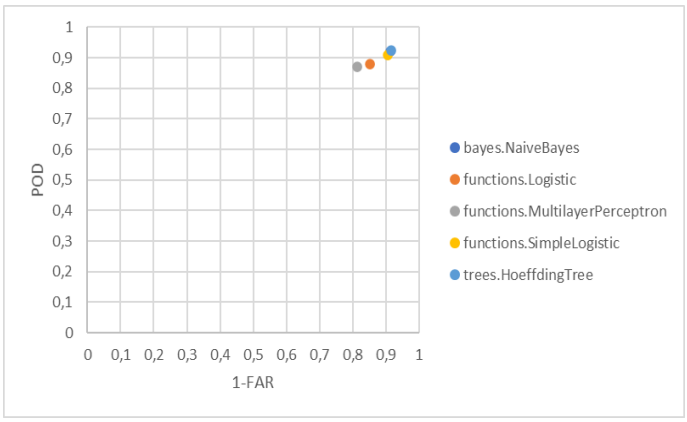
Para o teste da Figura 5.18 (a)-(e), o *Multilayer Perceptron* mostrou-se o único classificador com valores estatísticos aceitáveis para boas predições.



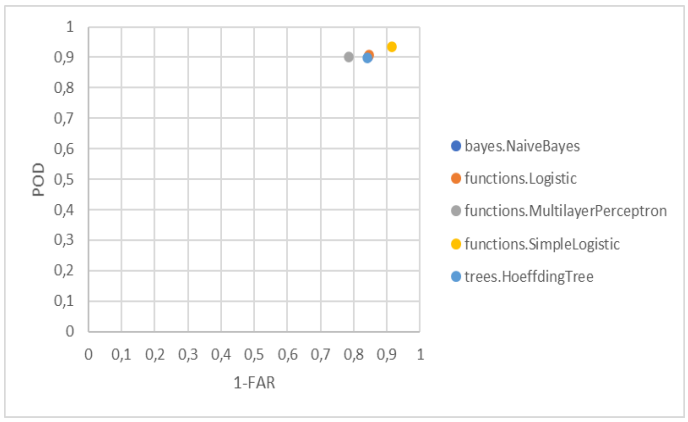
(a)



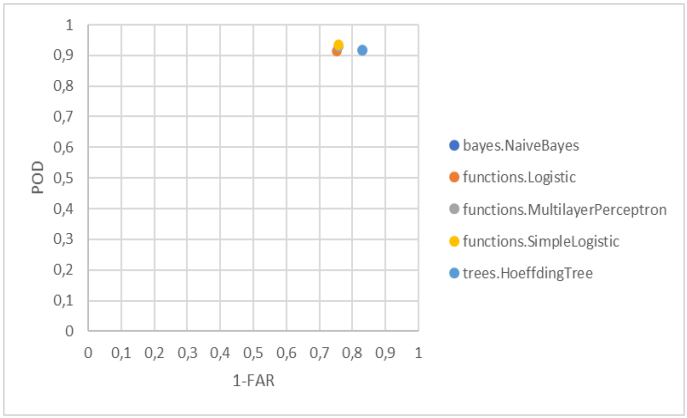
(b)



(c)

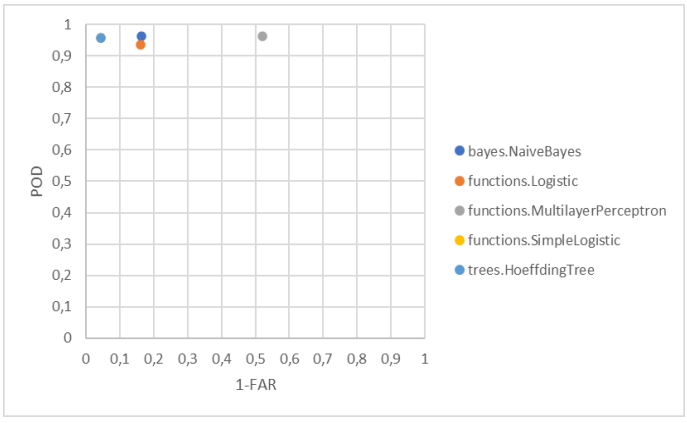


(d)

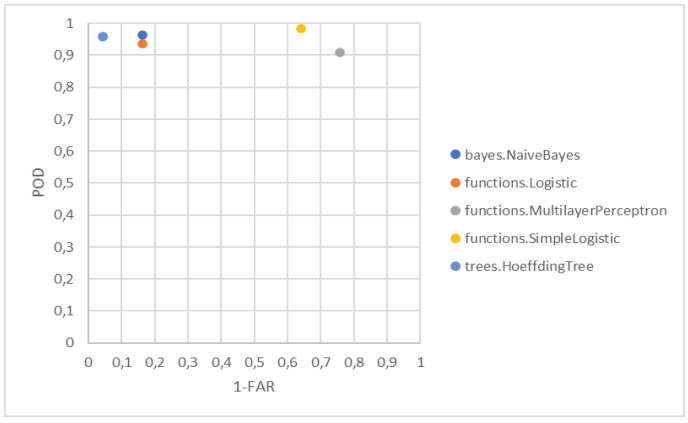


(e)

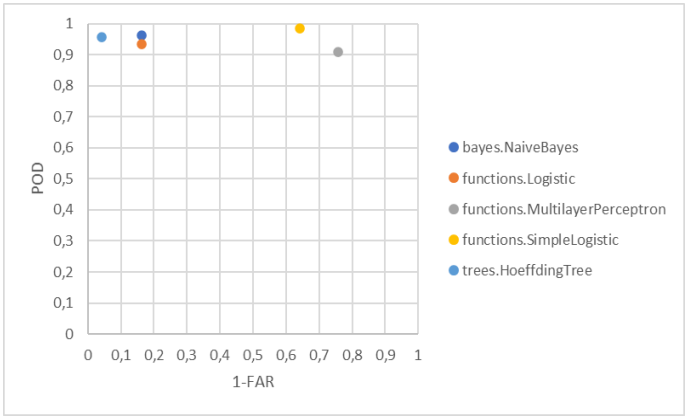
Figura 5.17 - Valores de POD versus 1-FAR de treinamento para cinco configurações de dados, para previsão às 13Z, considerando os cinco algoritmos utilizados na área II.



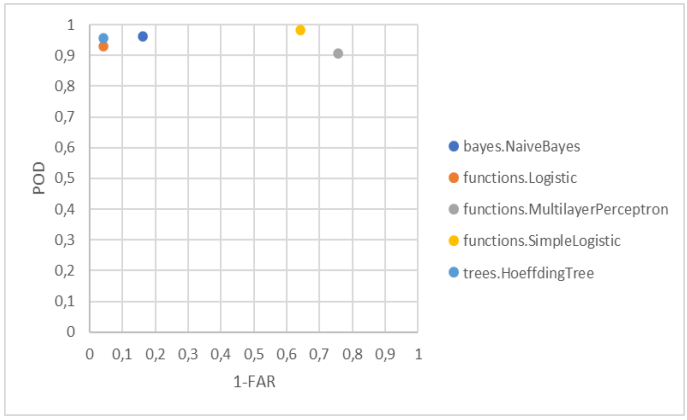
(a)



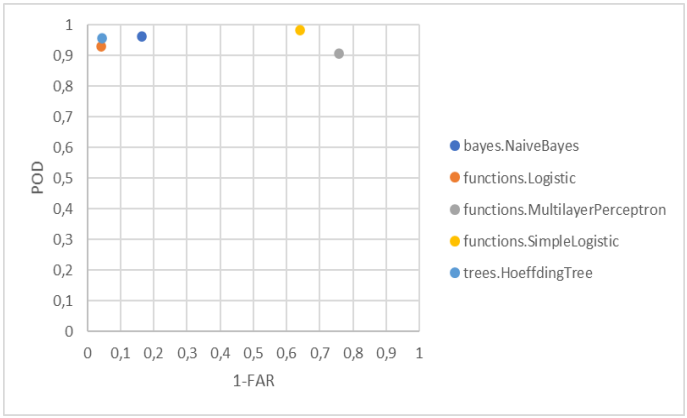
(b)



(c)



(d)



(e)

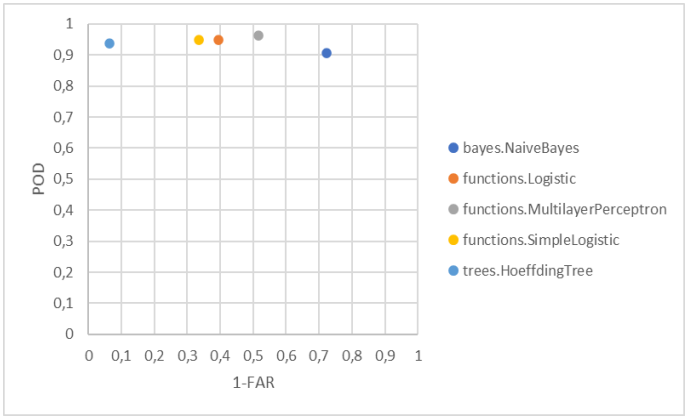
Figura 5.18 - Valores de POD versus 1-FAR de teste para cinco configurações de dados, para previsão às 13Z, considerando os cinco algoritmos utilizados na área II.

5.2.3 14Z

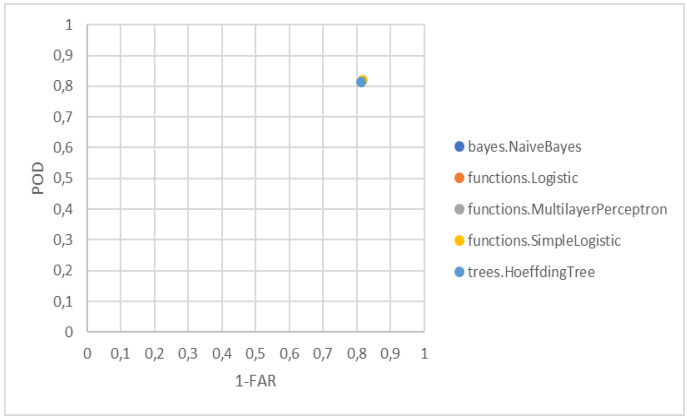
5.2.3.1 Área I

O estudo para a área I no horário das 14Z é representado pelas Figuras 5.19 e 5.20. Para o treinamento (Figura 5.19), no caso dos dados balanceados na configuração (2) (Figura 5.19(b)), todos os classificadores mostraram-se eficientes, superando o valor esperado.

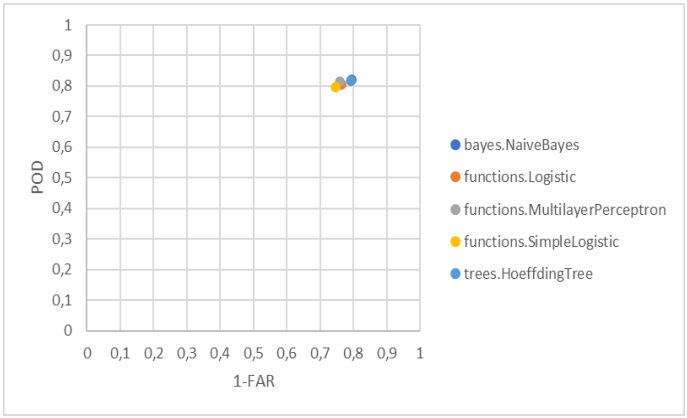
Em relação ao teste (Figura 5.20), o algoritmo de melhor performance foi o *Logistic* (86% de POD e 81% de 1-FAR) para os dados balanceados na configuração (3) (Figura 5.20(c)).



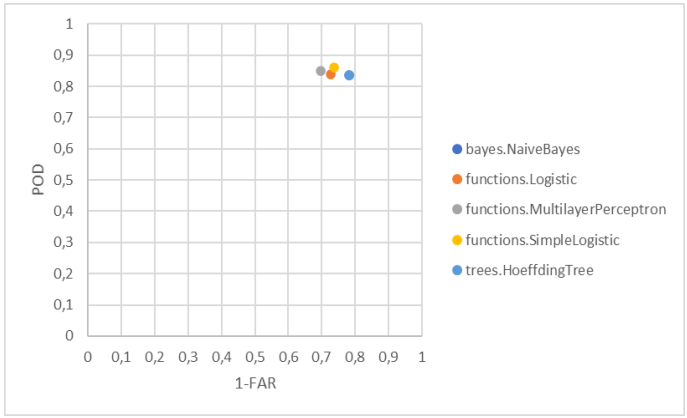
(a)



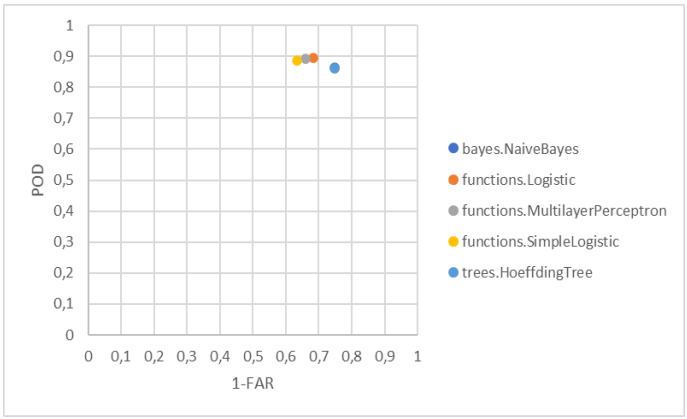
(b)



(c)

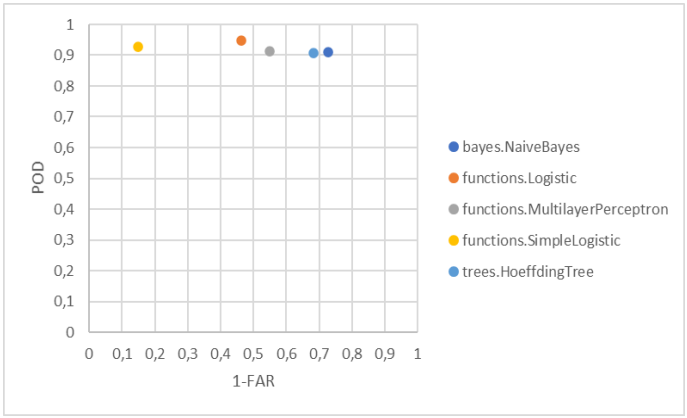


(d)

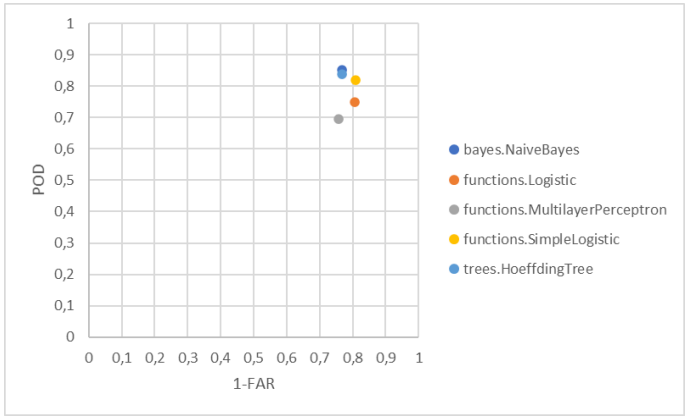


(e)

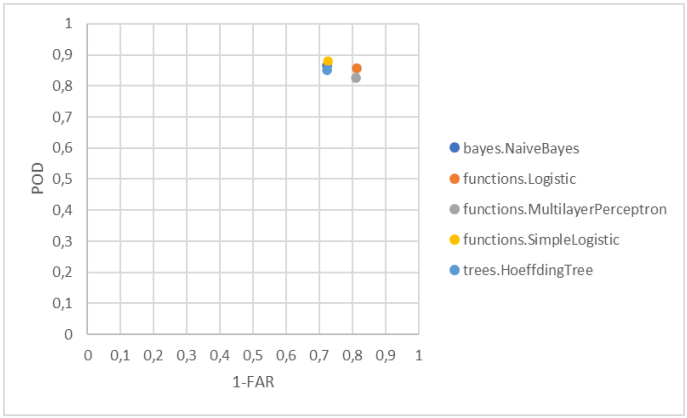
Figura 5.19 - Valores de POD versus 1-FAR de treinamento para cinco configurações de dados, para previsão às 14Z, considerando os cinco algoritmos utilizados na área I.



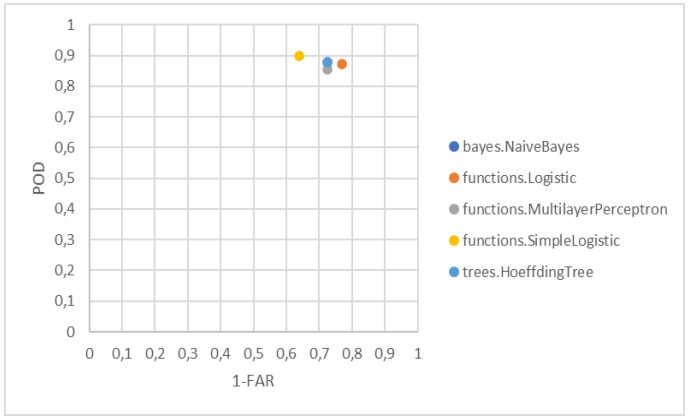
(a)



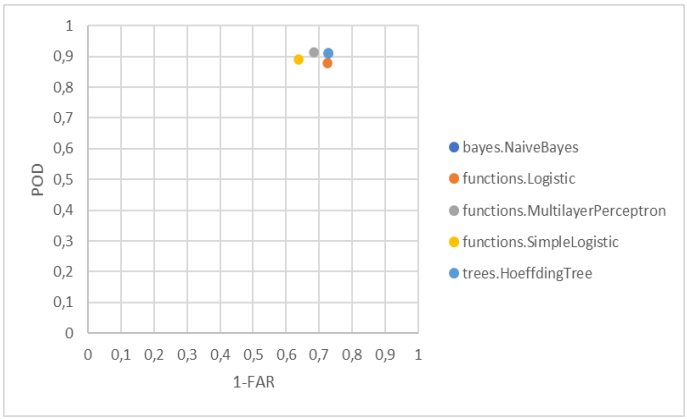
(b)



(c)



(d)



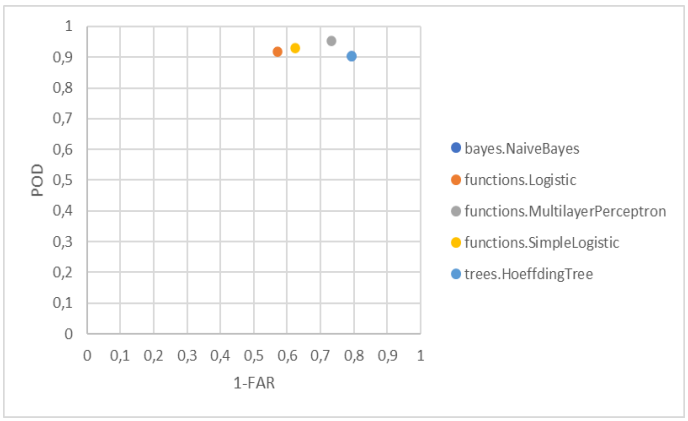
(e)

Figura 5.20 - Valores de POD versus 1-FAR de teste para cinco configurações de dados, para previsão às 14Z, considerando os cinco algoritmos utilizados na área I.

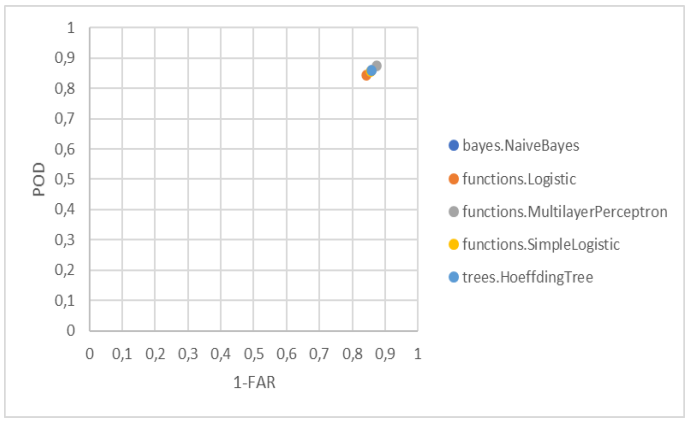
5.2.3.2 Área II

As Figuras 5.21 e 5.22 mostram os resultados para a área II às 14Z. Diferente dos demais casos, o treinamento para os dados originais (1) apresentam-se de maneira mais satisfatória, e que os classificadores se mostram menos dispersos (Figura 5.21(a)). O melhor resultado foi obtido pelo *Multilayer Perceptron* (92% POD e 84% 1-FAR) para os dados balanceados em (4) (Figura 5.21(d)).

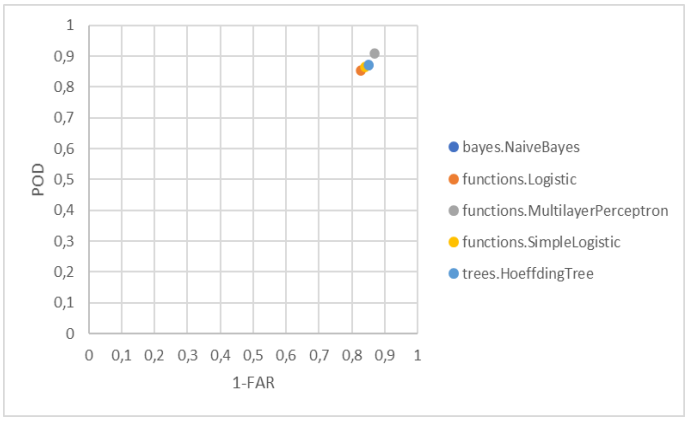
Para o teste, o *Naive Bayes* apresentou os melhores resultados, tanto para os dados originais (1) quanto para os balanceados na configuração (5) (Figura 5.22 (a) e (e)).



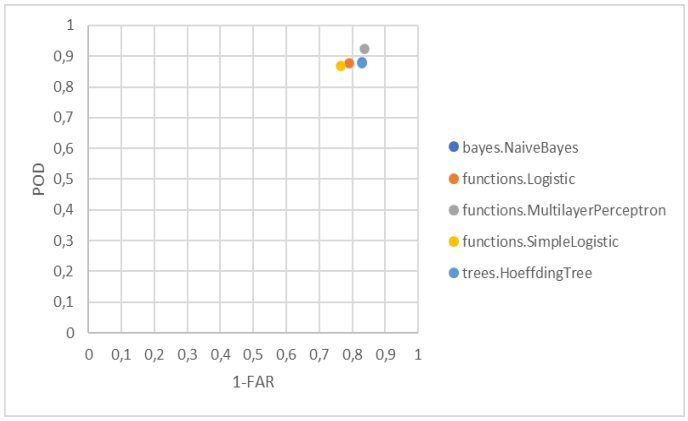
(a)



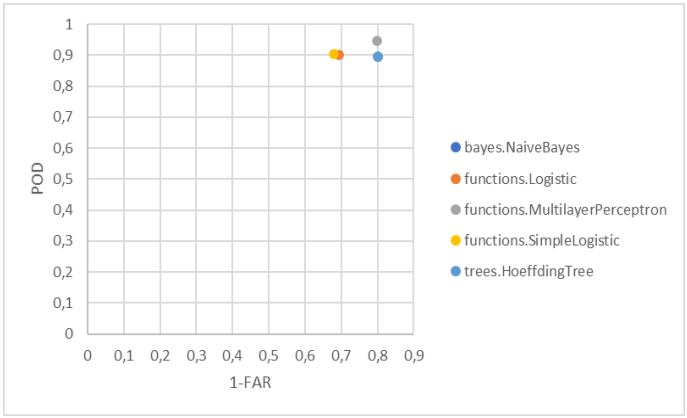
(b)



(c)

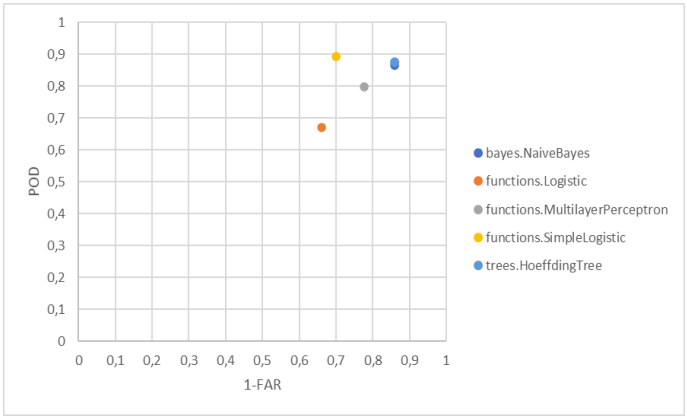


(d)

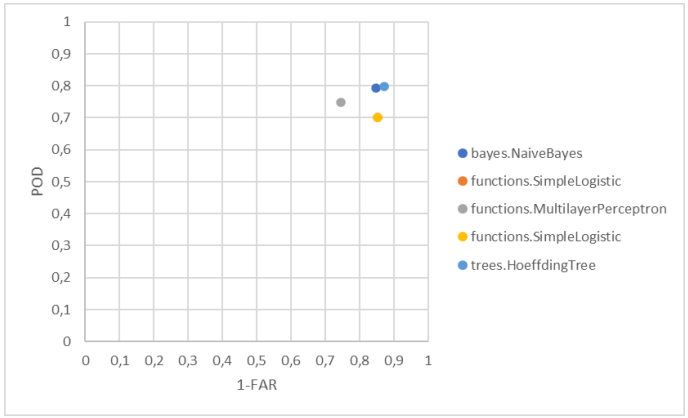


(e)

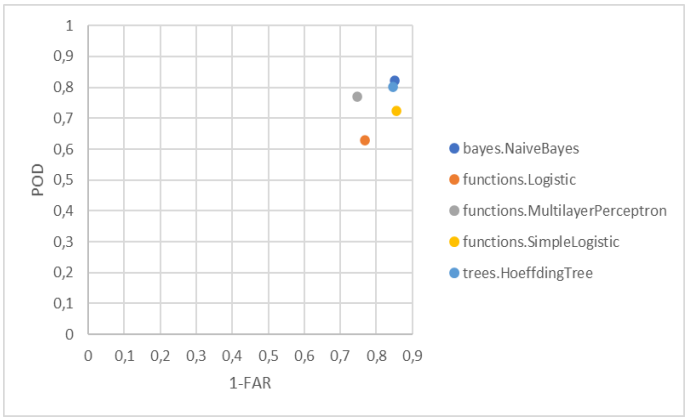
Figura 5.21 - Valores de POD versus 1-FAR de treinamento para cinco configurações de dados, para previsão às 14Z, considerando os cinco algoritmos utilizados na área II.



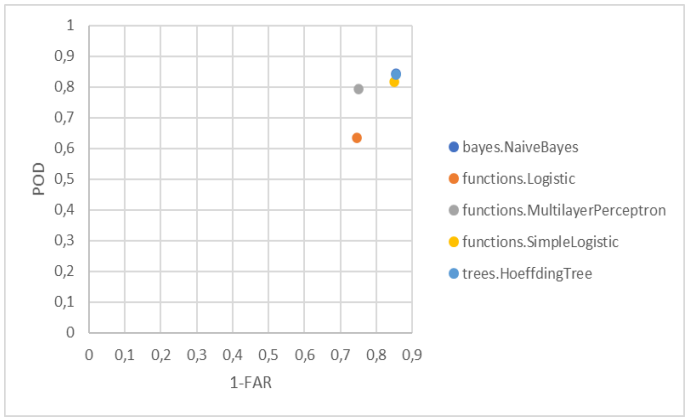
(a)



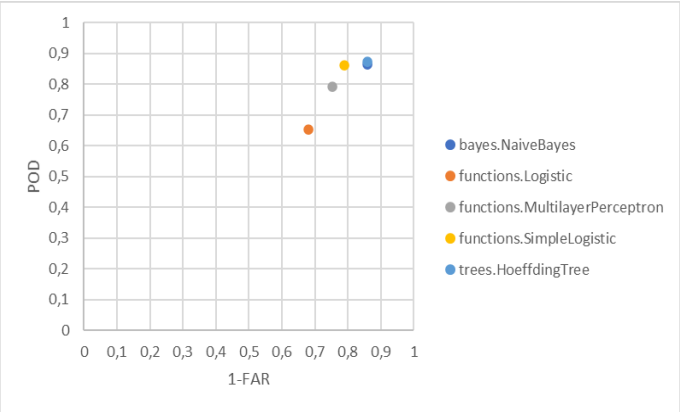
(b)



(c)



(d)



(e)

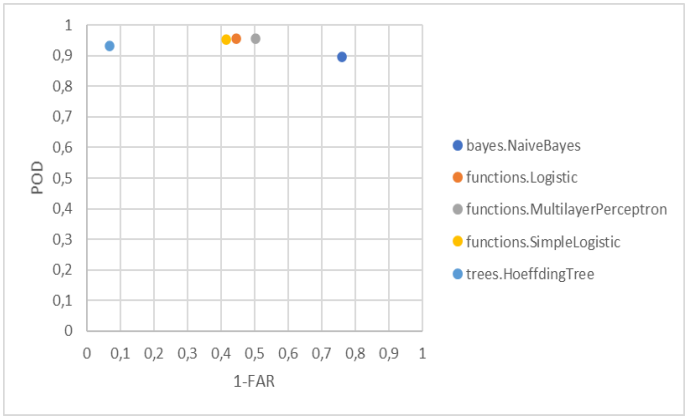
Figura 5.22 - Valores de POD versus 1-FAR de teste para cinco configurações de dados, para previsão às 14Z, considerando os cinco algoritmos utilizados na área II.

5.5.4 15Z

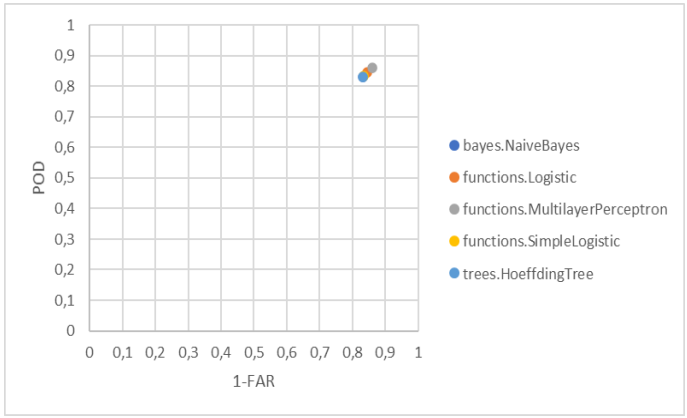
5.5.4.1 Área I

As Figuras 5.23 e 5.24 apresentam os gráficos de POD *versus* 1-FAR para a área I às 15Z. Para o treinamento (Figura 5.23) com os dados originais (1), novamente o *Naive Bayes* foi o único que se aproximou do valor esperado. Com os dados balanceados, todos os classificadores são aperfeiçoados e o melhor resultado é fornecido pelo *Logistic* (85% de POD e 1-FAR) configuração (2) (Figura 5.23(b)).

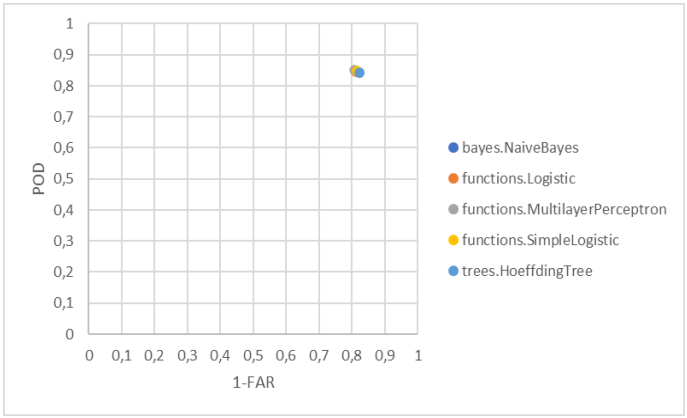
Em relação ao teste o *HoeffdingTree* para dados balanceados em (5) (Figura 5.24(e)), mostrou-se o mais eficiente.



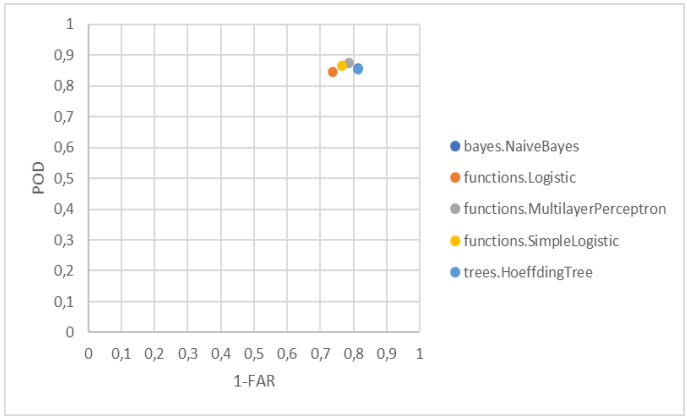
(a)



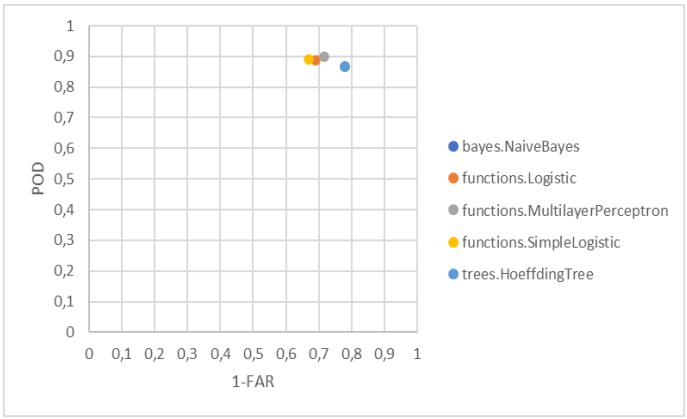
(b)



(c)

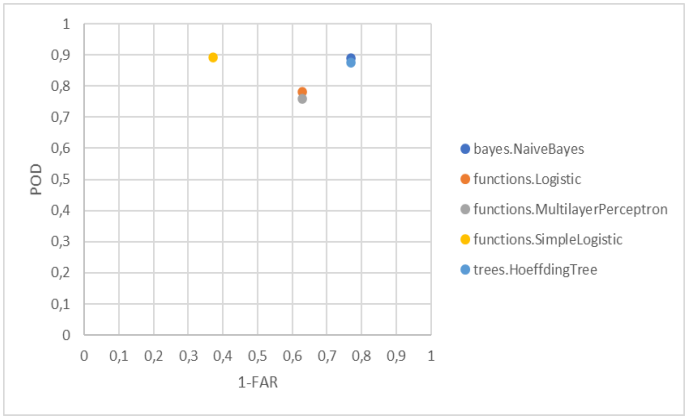


(d)

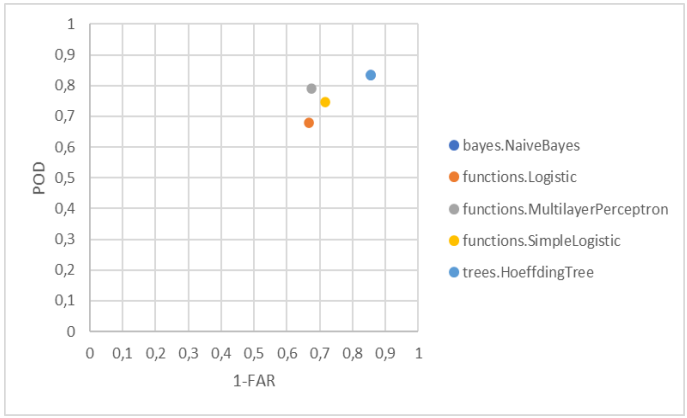


(e)

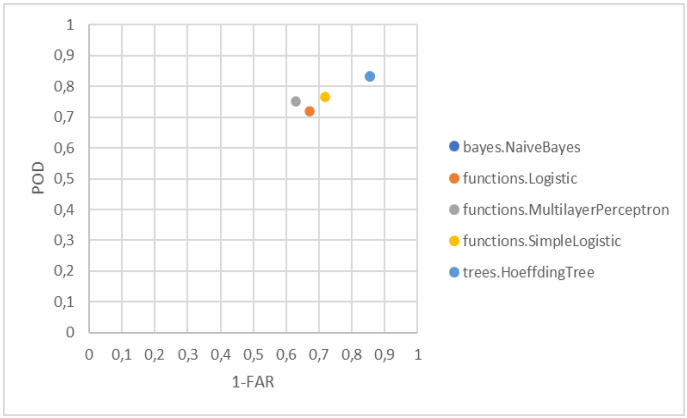
Figura 5.23 - Valores de POD versus 1-FAR de treinamento para cinco configurações de dados, para previsão às 15Z, considerando os cinco algoritmos utilizados na área I.



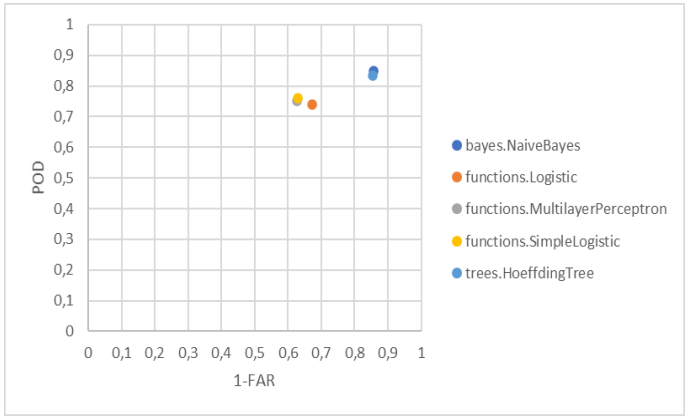
(a)



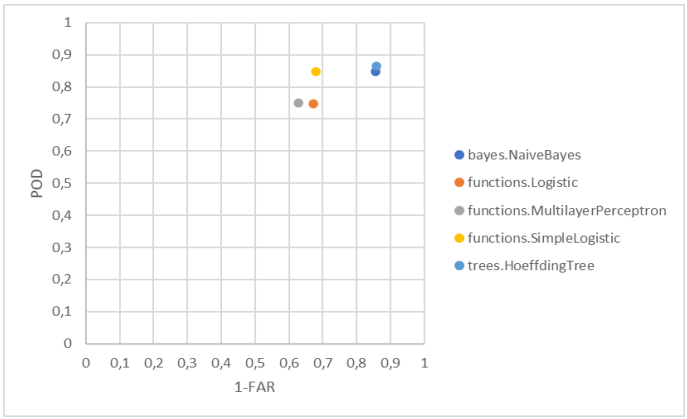
(b)



(c)



(d)



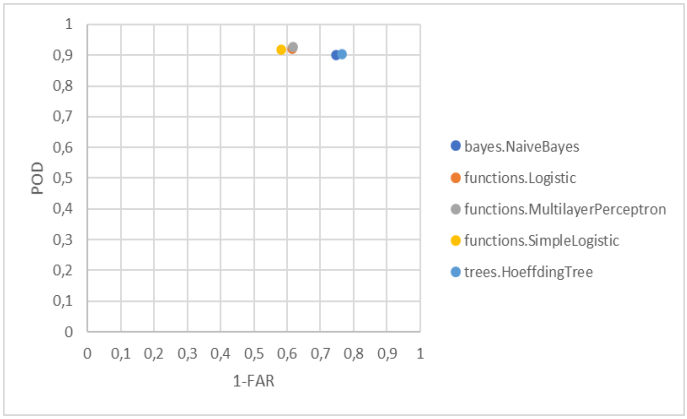
(e)

Figura 5.24 - Valores de POD versus 1-FAR de teste para cinco configurações de dados, para previsão às 15Z, considerando os cinco algoritmos utilizados na área I.

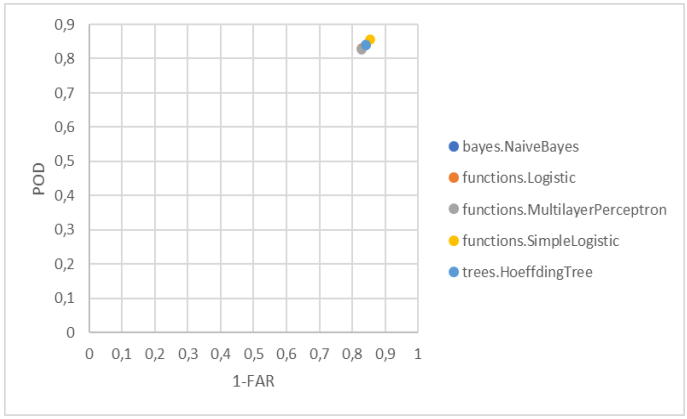
5.5.4.2 Área II

Nas Figuras 5.25 e 5.26 são fornecidas as estatísticas POD *versus* 1-FAR para a área II às 15Z. Como já mencionado, o treinamento (Figura 5.25) com os dados balanceados apresentam melhores resultados. Nesse caso, o classificador *Simple Logistic* com os valores equilibrados em (2) para a ocorrência e não ocorrência das descargas atmosféricas foi o que obteve melhor desempenho (86% de POD e 1-FAR, Figura 5.25(b)).

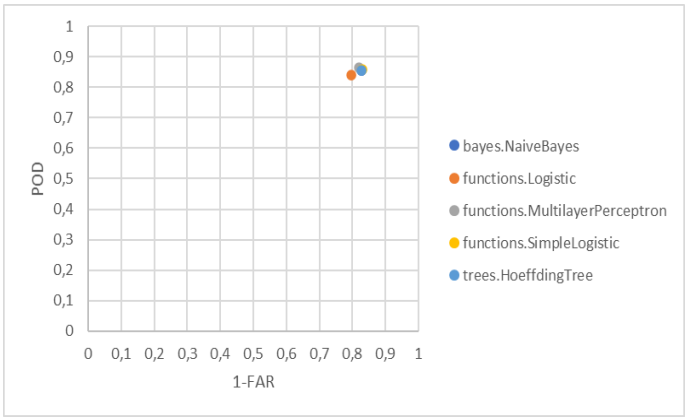
Em relação ao teste (Figura 5.26), tanto os dados originais (1) quanto os balanceados em (5) (Figura 5.26 (a) e (d)) apresentaram valores iguais (90% de POD e 85% de 1-FAR) para o classificador *Simple Logistic.*



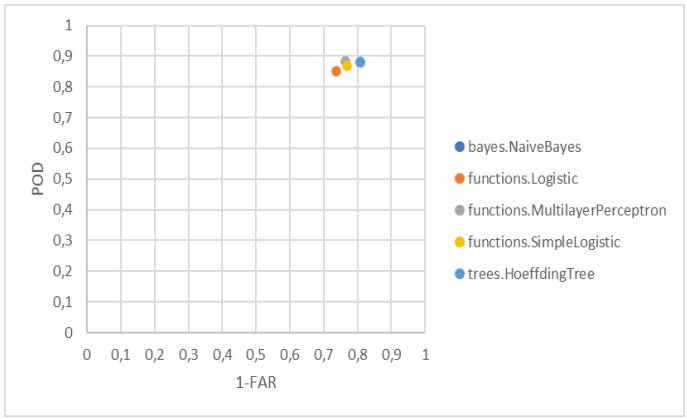
(a)



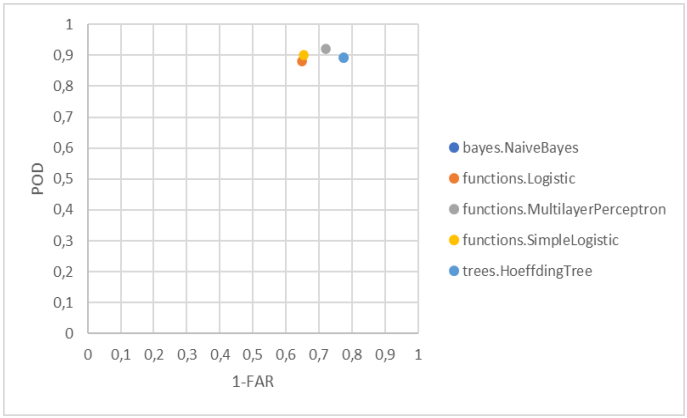
(b)



(c)

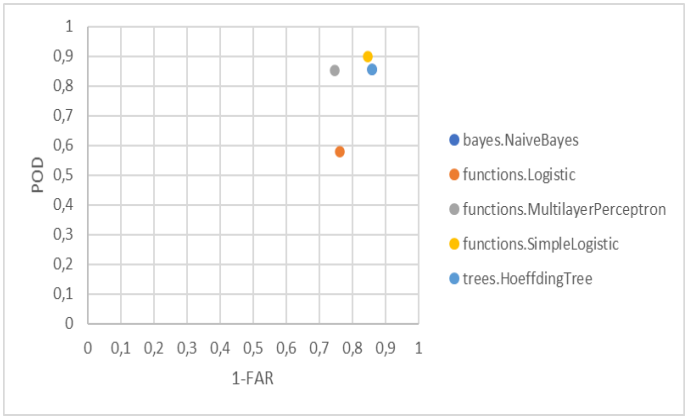


(d)

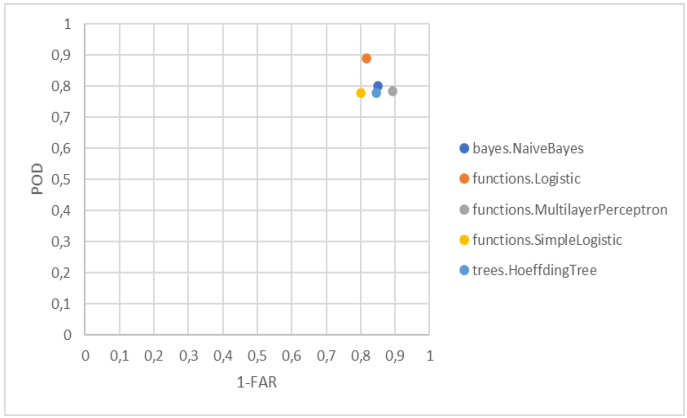


(e)

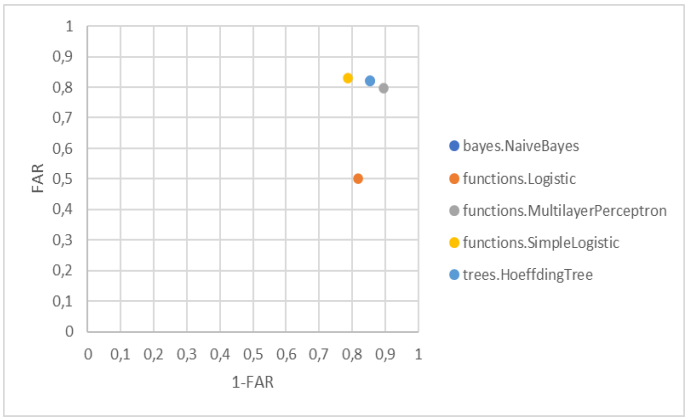
Figura 5.25 - Valores de POD versus 1-FAR de treinamento para cinco configurações de dados, para previsão às 15Z, considerando os cinco algoritmos utilizados na área II.



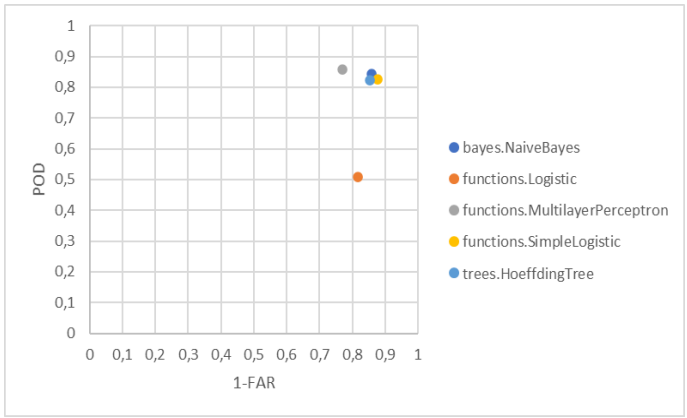
(a)



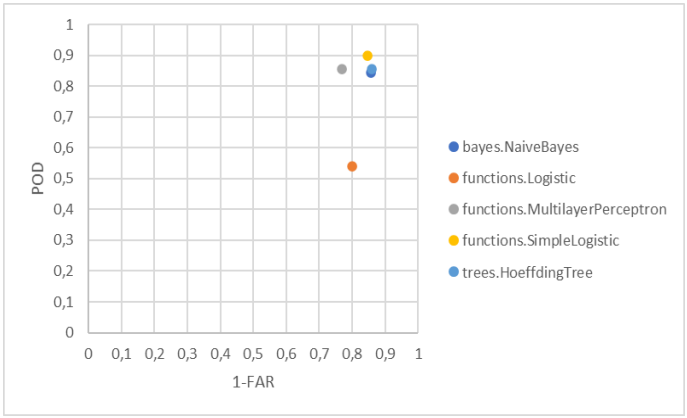
(b)



(c)



(d)



(e)

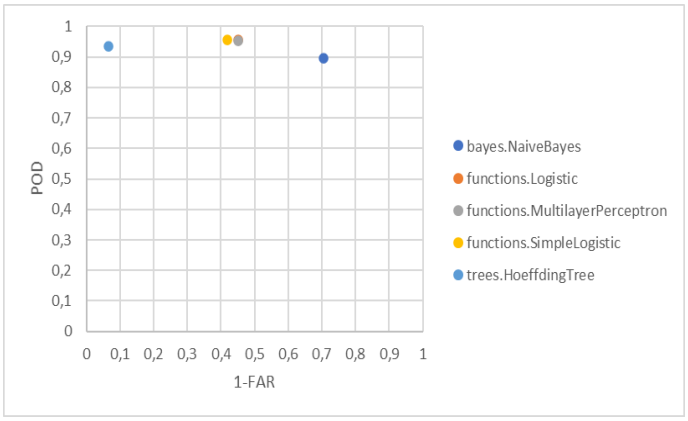
Figura 5.26 - Valores de POD versus 1-FAR de teste para cinco configurações de dados, para previsão às 15Z, considerando os cinco algoritmos utilizados na área II.

5.5.5 16Z

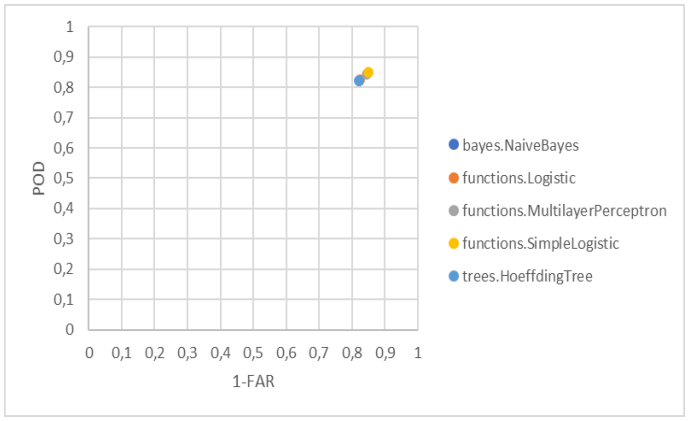
5.5.5.1 Área I

As Figuras 5.27 e 5.28 apresentam os gráficos relativos à área 16Z. O resultado do treinamento (Figura 5.27) com os dados originais para este horário mostrou-se insuficiente para todos os classificadores. Para os dados balanceados em (2) e (3) (Figura 5.27 (b) e (c)) todos os algoritmos obtiveram bons desempenhos.

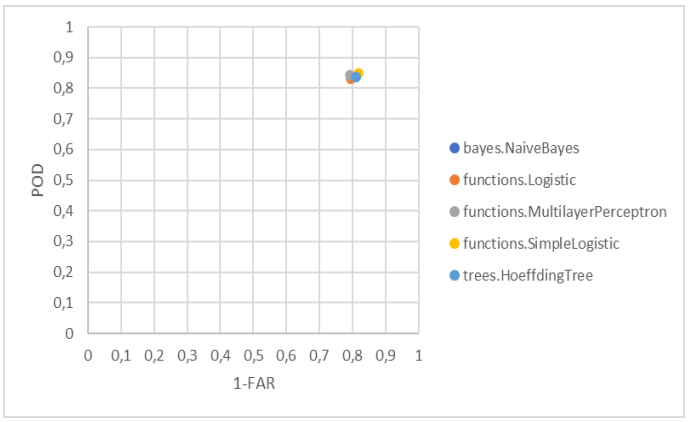
No teste da Figura 5.28 (a)-(e), nenhum classificador apresentou valores de 1-FAR satisfatórios.



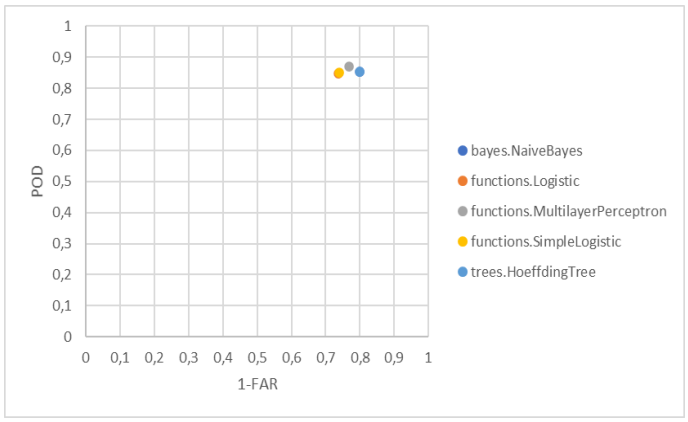
(a)



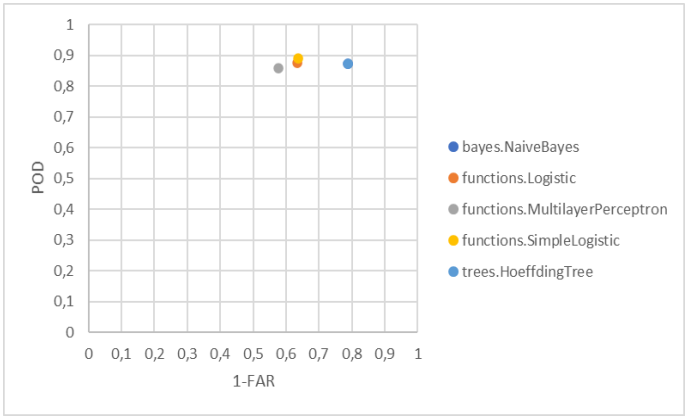
(b)



(c)

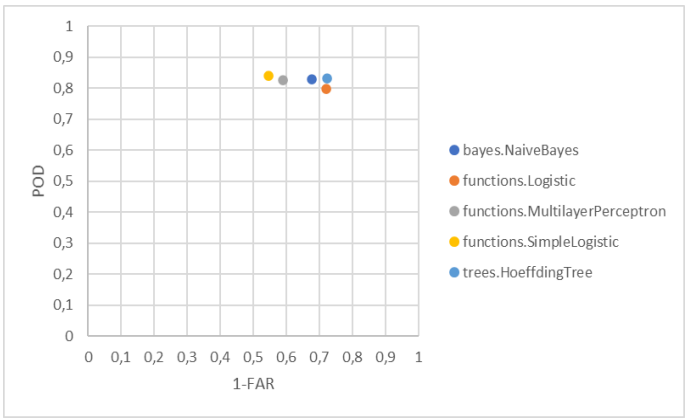


(d)

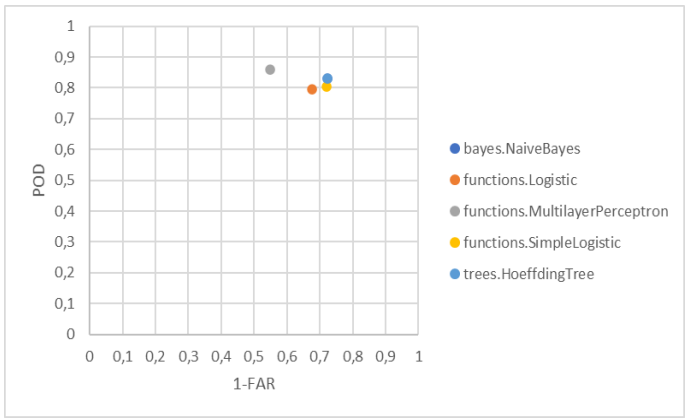


(e)

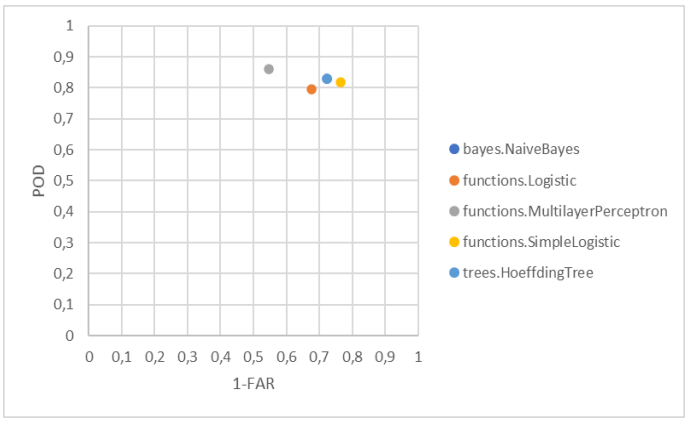
Figura 5.27 - Valores de POD versus 1-FAR de treinamento para cinco configurações de dados, para previsão às 16Z, considerando os cinco algoritmos utilizados na área I.



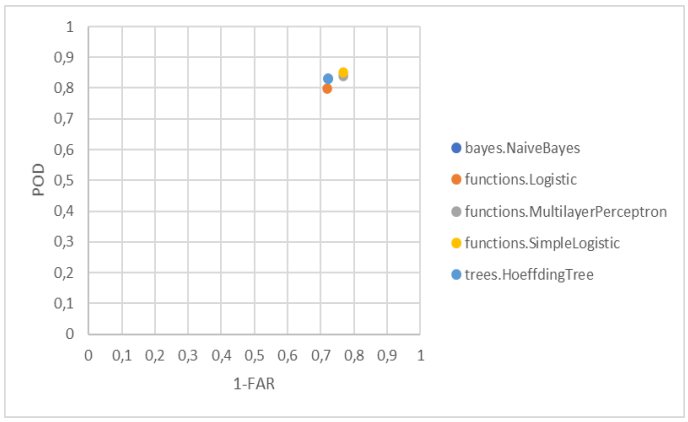
(a)



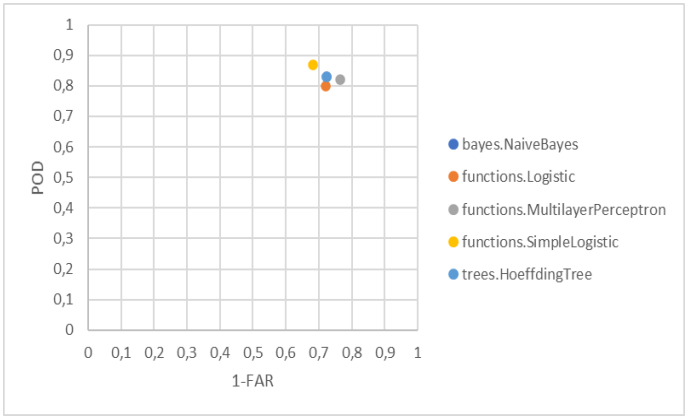
(b)



(c)



(d)



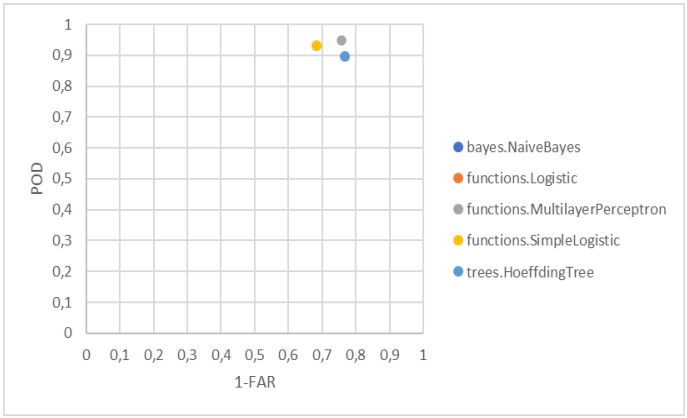
(e)

Figura 5.28- Valores de POD versus 1-FAR de teste para cinco configurações de dados, para previsão às 16Z, considerando os cinco algoritmos utilizados na área I.

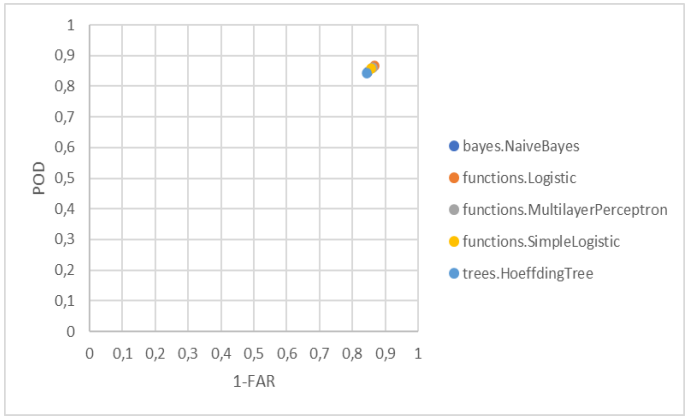
5.5.5.2 Área II

As Figuras de 5.29 e 5.30 apresentam os gráficos estatísticos para a área II às 16Z. Nesse caso os resultados das predições dos classificadores são superiores ao da área I. Todos os classificadores apresentam boas performances a partir dos dados originais (1), sendo o *Multilayer Perceptron* para dados balanceados na configuração (4) (Figura 5.29 (d)) o que obteve melhor desempenho.

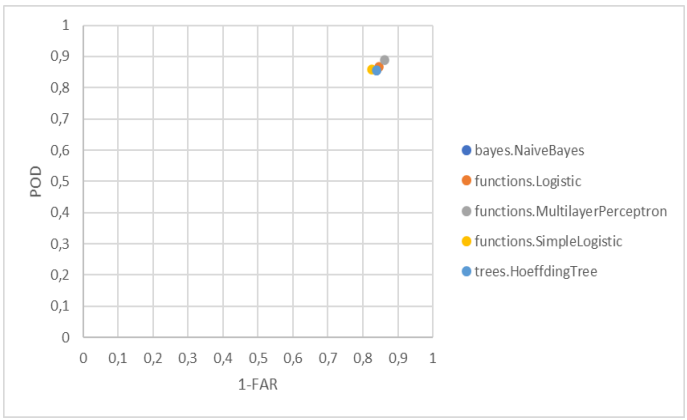
Em sequência para o teste da Figura 5.30, o *Simple Logistic* (84% POD e 80% de 1-FAR) balanceado em (5) (Figura 5.30 (d)) mostrou-se ser o classificador mais assertivo.



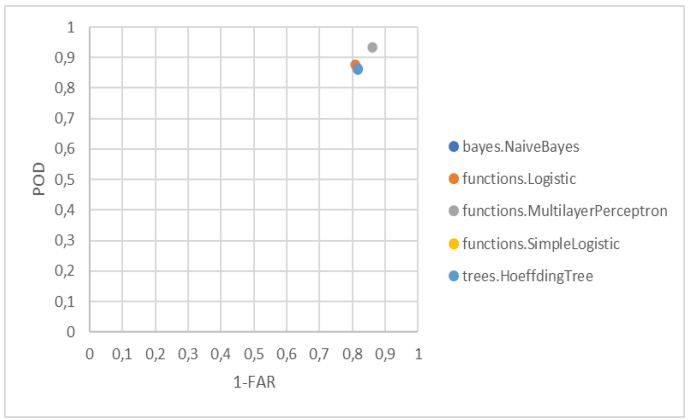
(a)



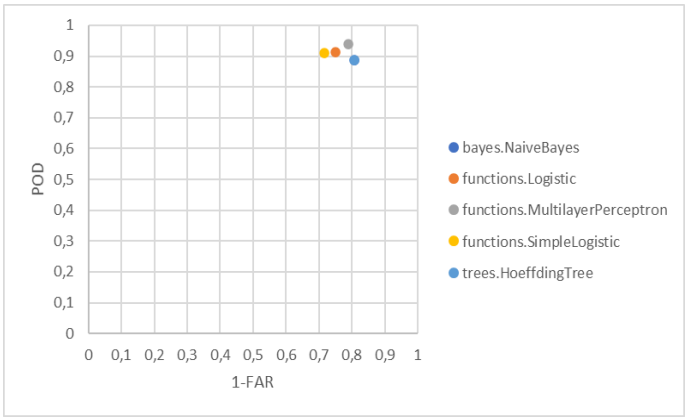
(b)



(c)

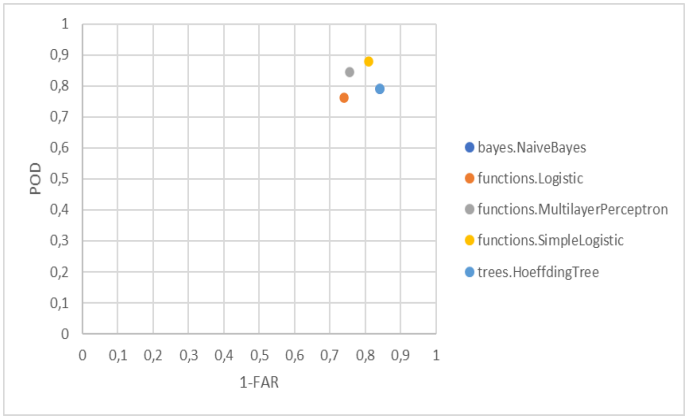


(d)

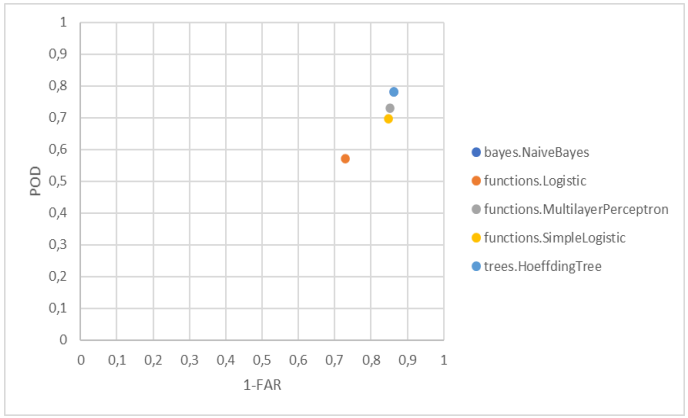


(e)

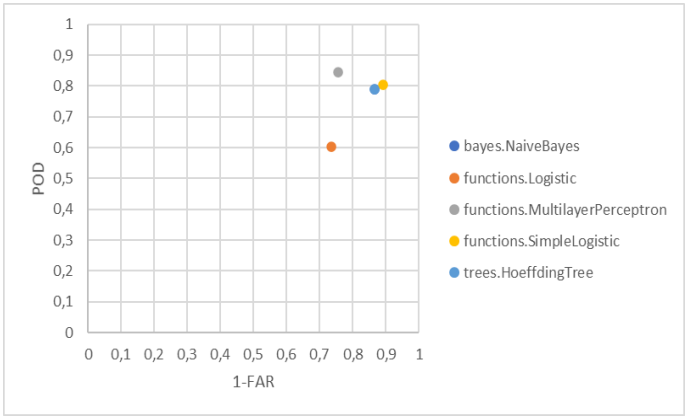
Figura 5.29 - Valores de POD versus 1-FAR de treinamento para cinco configurações de dados, para previsão às 16Z, considerando os cinco algoritmos utilizados na área II.



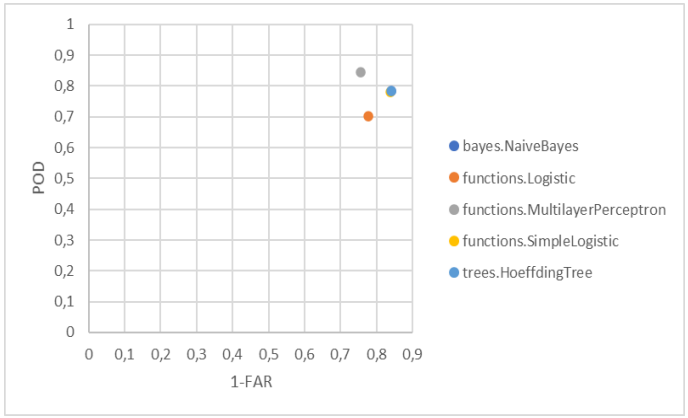
(a)



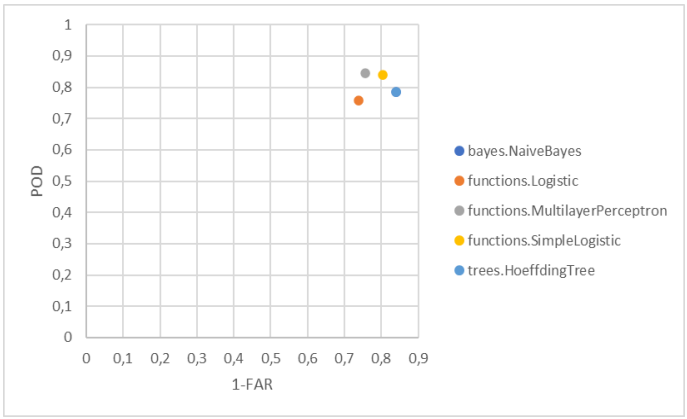
(b)



(c)



(d)



(e)

Figura 5.30 - Valores de POD versus 1-FAR de teste para cinco configurações de dados, para previsão às 16Z, considerando os cinco algoritmos utilizados na área II.

5.3 AVALIAÇÃO

Nesta sessão, será apresentado os valores estatísticos de cada classificador (via correlação cruzada) para a ocorrência ou não do evento.

As Tabelas 5.1 à 5.10 discorrem sobre os valores estatísticos dos melhores modelos de cada horário, onde “S”, “N” e “MD” significam, nos acrônimos das estatísticas de desempenho, “sim”, “não”, média, respectivamente.

De acordo com a Tabela 4.5 de Viera e Garrett (2005) todos os classificadores apontam no mínimo concordância substancial, sendo o horário das 13Z para a área II o maior valor de KAPPA encontrado (85%), indicando concordância perfeita.

Os horários das 13Z e 15Z (Tabelas 5.4 e 5.7), conforme valores de KAPPA (85% e 72%) e *F-Measure* (MD) (94% e 85%), indicam que os algoritmos *Simple Logistic e Multilayer Percepetron* são os de melhor desempenho para realização da previsão de 8h (área II) e 6h (área I) de EMC e não EMC, respectivamente.

Tabela 5.1 - Estatísticas dos algoritmos preditivos utilizados para a área I com a configuração (3) para às 12Z.

BALANCEADO (60% "SIM"/ 40% "NÃO")	12Z - ÁREA 1											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
Naive Bayes	0,85	0,76	0,76	0,85	0,85	1,01	0,77	0,98	0,82	0,80	0,81	0,61
Logistic	0,90	0,73	0,73	0,90	0,87	1,08	0,78	0,87	0,83	0,80	0,83	0,64
Multilayer Perceptron	0,96	0,70	0,70	0,96	0,89	1,16	0,79	0,76	0,86	0,80	0,85	0,69
Simple Logistic	0,91	0,70	0,70	0,91	0,86	1,11	0,76	0,84	0,82	0,78	0,82	0,62
Hoeffding Tree	0,85	0,76	0,76	0,85	0,85	1,01	0,76	0,98	0,82	0,80	0,82	0,61

Tabela 5.2 - Estatísticas dos algoritmos preditivos utilizados para a área II com a configuração (2) para às 12Z.

BALANCEADO (50% "SIM"/ 50% "NÃO")	12Z - ÁREA2											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
Naive Bayes	0,88	0,78	0,78	0,88	0,84	1,09	0,82	0,91	0,83	0,84	0,83	0,66
Logistic	0,87	0,77	0,77	0,87	0,83	1,11	0,81	0,89	0,82	0,82	0,82	0,64
Multilayer Perceptron	0,93	0,71	0,71	0,93	0,83	1,22	0,79	0,78	0,82	0,82	0,81	0,63
Simple Logistic	0,87	0,78	0,78	0,87	0,83	1,09	0,82	0,91	0,83	0,83	0,83	0,65
Hoeffding Tree	0,88	0,78	0,78	0,88	0,84	1,09	0,82	0,91	0,83	0,83	0,83	0,66

Tabela 5.3 - Estatísticas dos algoritmos preditivos utilizados para a área I com a configuração (2) para às 13Z.

BALANCEADO (50% "SIM"/ 50% "NÃO")	13Z - ÁREA1											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
Naive Bayes	0,86	0,78	0,78	0,86	0,83	1,08	0,81	0,88	0,82	0,82	0,82	0,64
Logistic	0,87	0,75	0,75	0,87	0,82	1,12	0,80	0,88	0,81	0,81	0,81	0,62
Multilayer Perceptron	0,91	0,78	0,78	0,91	0,85	1,13	0,83	0,87	0,85	0,85	0,84	0,69
Simple Logistic	0,87	0,75	0,75	0,87	0,82	1,13	0,80	0,88	0,81	0,81	0,81	0,62
Hoeffding Tree	0,86	0,78	0,78	0,86	0,83	1,08	0,81	0,92	0,82	0,82	0,82	0,64

Tabela 5.4 - Estatísticas dos algoritmos preditivos utilizados para a área II com a configuração (4) para às 13Z.

BALANCEADO (70% "SIM"/ 30% "NÃO")	13Z - ÁREA2											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
Naive Bayes	0,95	0,80	0,80	0,95	0,93	1,03	0,83	0,93	0,90	0,84	0,90	0,76
Logistic	0,95	0,80	0,80	0,95	0,94	1,04	0,84	0,91	0,91	0,85	0,91	0,77
Multilayer Perceptron	0,99	0,70	0,70	0,99	0,93	1,11	0,81	0,73	0,90	0,79	0,90	0,74
Simple Logistic	0,95	0,90	0,90	0,95	0,95	0,99	0,89	1,02	0,94	0,92	0,94	0,85
Hoeffding Tree	0,94	0,80	0,80	0,94	0,93	1,02	0,82	0,94	0,90	0,84	0,90	0,75

Tabela 5.5 - Estatísticas dos algoritmos preditivos utilizados para a área I com a configuração (2) para às 14Z.

BALANCEADO (50% "SIM"/ 50% "NÃO")	14Z - ÁREA1											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
Naive Bayes	0,85	0,77	0,77	0,85	0,82	1,08	0,81	0,92	0,81	0,81	0,81	0,63
Logistic	0,86	0,77	0,77	0,86	0,82	1,08	0,81	0,92	0,82	0,82	0,82	0,63
Multilayer Perceptron	0,89	0,74	0,74	0,89	0,83	1,15	0,80	0,85	0,82	0,82	0,82	0,63
Simple Logistic	0,86	0,77	0,77	0,86	0,83	1,09	0,81	0,91	0,82	0,82	0,82	0,64
Hoeffding Tree	0,86	0,77	0,77	0,86	0,82	1,08	0,81	0,92	0,81	0,81	0,81	0,63

Tabela 5.6 - Estatísticas dos algoritmos preditivos utilizados para a área II com a configuração (3) para às 14Z.

BALANCEADO (60% "SIM"/ 40% "NÃO")	14Z - ÁREA2											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
Naive Bayes	0,91	0,81	0,81	0,91	0,90	1,06	0,84	0,94	0,87	0,85	0,87	0,73
Logistic	0,91	0,77	0,77	0,91	0,88	1,06	0,81	0,90	0,86	0,83	0,85	0,69
Multilayer Perceptron	0,99	0,79	0,79	0,99	0,93	1,13	0,87	0,81	0,91	0,87	0,91	0,80
Simple Logistic	0,91	0,79	0,79	0,91	0,89	1,05	0,82	0,92	0,86	0,84	0,86	0,71
Hoeffding Tree	0,91	0,81	0,81	0,91	0,90	1,04	0,84	0,94	0,87	0,85	0,87	0,73

Tabela 5.7 - Estatísticas dos algoritmos preditivos utilizados para a área I com a configuração (2) para às 15Z.

BALANCEADO (50% "SIM"/ 50% "NÃO")	15Z - ÁREA1											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
Naive Bayes	0,88	0,78	0,78	0,88	0,84	1,10	0,82	0,90	0,83	0,83	0,83	0,66
Logistic	0,85	0,84	0,84	0,85	0,85	1,00	0,85	1,00	0,85	0,85	0,85	0,69
Multilayer Perceptron	0,94	0,78	0,78	0,94	0,87	1,16	0,85	0,84	0,86	0,86	0,86	0,72
Simple Logistic	0,86	0,81	0,81	0,86	0,84	1,04	0,83	0,96	0,83	0,83	0,83	0,67
Hoeffding Tree	0,88	0,78	0,78	0,88	0,84	1,10	0,82	0,90	0,83	0,83	0,83	0,66

Tabela 5.8 - Estatísticas dos algoritmos preditivos utilizados para a área II com a configuração (2) para às 15Z.

BALANCEADO (50% "SIM"/ 50% "NÃO")	15Z - ÁREA2											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
Naive Bayes	0,91	0,78	0,78	0,91	0,85	1,13	0,83	0,87	0,84	0,84	0,84	0,68
Logistic	0,88	0,78	0,78	0,88	0,84	1,11	0,82	0,89	0,83	0,83	0,83	0,66
Multilayer Perceptron	0,92	0,74	0,74	0,92	0,84	1,19	0,81	0,81	0,83	0,83	0,83	0,66
Simple Logistic	0,89	0,82	0,82	0,89	0,86	1,08	0,85	0,92	0,86	0,86	0,85	0,71
Hoeffding Tree	0,91	0,78	0,78	0,91	0,85	1,13	0,83	0,87	0,84	0,84	0,84	0,68

Tabela 5.9 - Estatísticas dos algoritmos preditivos utilizados para a área I com a configuração (2) para às 16Z.

BALANCEADO (50% "SIM"/ 50% "NÃO")	16Z - ÁREA1											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
Naive Bayes	0,88	0,76	0,76	0,88	0,83	1,12	0,81	0,88	0,82	0,82	0,82	0,64
Logistic	0,86	0,79	0,79	0,86	0,83	1,10	0,82	0,94	0,83	0,83	0,83	0,65
Multilayer Perceptron	0,93	0,76	0,76	0,93	0,86	1,17	0,83	0,83	0,84	0,84	0,84	0,69
Simple Logistic	0,87	0,83	0,83	0,87	0,85	1,05	0,85	0,95	0,85	0,85	0,85	0,70
Hoeffding Tree	0,88	0,76	0,76	0,88	0,83	1,12	0,81	0,88	0,82	0,82	0,82	0,64

Tabela 5.10 - Estatísticas dos algoritmos preditivos utilizados para a área II com a configuração (4) para às 16Z.

BALANCEADO (70% "SIM"/ 30% "NÃO")	16Z - ÁREA2											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
Naive Bayes	0,90	0,78	0,78	0,90	0,90	0,99	0,77	1,00	0,86	0,82	0,86	0,67
Logistic	0,92	0,76	0,76	0,92	0,91	1,02	0,79	0,94	0,88	0,81	0,87	0,70
Multilayer Perceptron	0,99	0,80	0,80	0,99	0,95	1,07	0,88	0,83	0,93	0,86	0,93	0,83
Simple Logistic	0,95	0,72	0,72	0,95	0,92	1,09	0,78	0,83	0,88	0,79	0,88	0,70
Hoeffding Tree	0,90	0,78	0,78	0,90	0,90	0,99	0,78	1,02	0,87	0,82	0,87	0,68

5.3.1 Severidade

Como descrito anteriormente, a severidade dos EMC foi avaliada para os modelos ótimos (modelos de melhor desempenho para cada área) às 15Z (área I) e 13Z (área II). Portanto esses horários são os que apresentam maior disponibilidade de energia e atividade convectiva.

As Figuras 5.31 e 5.32 apresentam os gráficos POD versus 1-FAR dos algoritmos que obtiveram os melhores desempenhos. Esses algoritmos denominados J48 e Randon Forest foram treinados e testados via correlação cruzada triplicando os eventos com valores de DA maior que 130 raios; os algoritmos com o valor “50”, são os dados balanceados na configuração (2).

A Figura 5.31, que mostra o gráfico de POD versus 1-FAR relativo à área I, revela que os dados originais (1) apresentam um alto valor de detecção dos eventos, porém com um alto índice de falso alarme. Com o balanceamento dos dados (2), o poder de detecção é reduzido assim como o falso alarme, ficando em torno de 78% de POD e 1-FAR.

Já na Figura 5.32, os dois modelos correspondentes à área II apresentam altos valores de POD e 1-FAR (em torno de 90%).

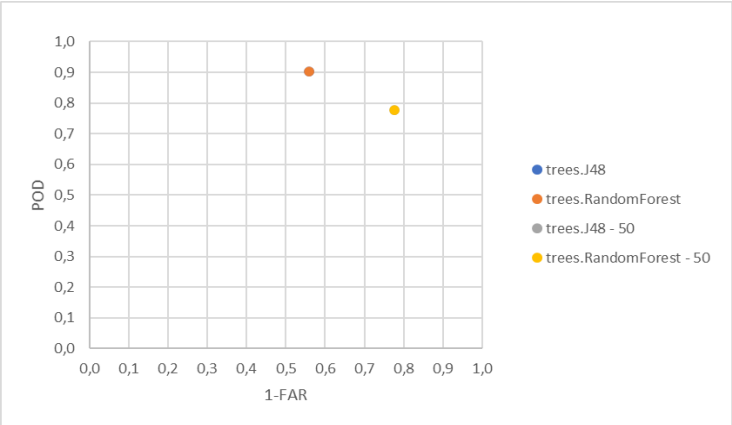


Figura 5.31 - Gráfico POD *versus* 1-FAR da previsão da severidade para os cinco algoritmos selecionados com dados de *input* das 15Z para a área I.

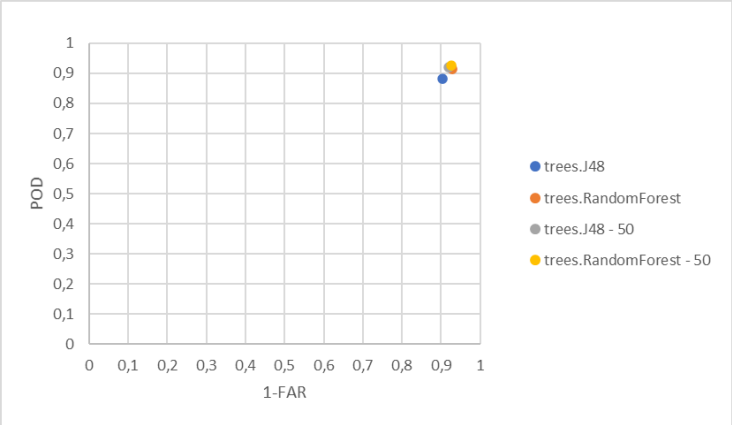


Figura 5.32 - Gráfico POD *versus* 1-FAR da previsão da severidade para os cinco algoritmos selecionados com dados de *input* das 13Z.

As Tabelas 5.11 à 5.14 apresentam as estatísticas utilizadas para avaliar o desempenho desses modelos. As Tabelas 5.11 e 5.12 relacionadas à área I mostram valores estatísticos inferiores à área II, isso se evidencia mais fortemente na Tabela 5.11. As Tabelas 5.13 e 5.14 que correspondem à área II, indicam valores expressivos que implicam que os modelos elaboram boas predições.

Analisando a severidade, identifica-se que o modelo é mais sensível à predição de eventos significativos, ou seja, quanto maior e mais intenso o EMC melhor a capacidade do modelo em detectá-lo.

Tabela 5.11 – Estatísticas de desempenho, para correlação cruzada – configuração (1), dos algoritmos de previsão de severidade dos EMC com dados de *input* das 15Z.

DADOS ORIGINAIS	**15Z - ÁREA1**											
	POD (S)	**1-FAR(S)**	**POD(N)**	**1-FAR(N)**	**F-MEASURE(S)**	**BIAS(S)**	**F-MEASURE(N)**	**BIAS(N)**	**POD (MD)**	**1-FAR (MD)**	**F-MEASURE(MD)**	**KAPPA**
J48	0,99	0,47	0,47	0,99	0,94	1,11	0,63	0,49	0,90	0,56	0,89	0,58
RandomForest	0,99	0,47	0,47	0,99	0,95	1,11	0,63	0,48	0,90	0,56	0,89	0,58

Tabela 5.12 – Estatísticas de desempenho, para correlação cruzada – configuração (2), dos algoritmos de previsão de severidade dos EMC com dados de *input* das 15Z.

BALANCEADO (50% "SIM"/ 50% "NÃO")	15Z - ÁREA1											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
J48	0,89	0,67	0,67	0,89	0,80	1,22	0,75	0,78	0,78	0,78	0,77	0,55
RandomForest	0,89	0,67	0,67	0,89	0,80	1,22	0,75	0,78	0,78	0,78	0,77	0,55

Tabela 5.13 - Estatísticas de desempenho, para correlação cruzada – configuração (1), dos algoritmos de previsão de severidade dos EMC com dados de *input* das 13Z.

DADOS ORIGINAIS	13Z - ÁREA2											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
J48	0,92	0,87	0,87	0,92	0,81	1,26	0,92	0,90	0,88	0,90	0,89	0,73
RandomForest	0,94	0,91	0,91	0,94	0,86	1,19	0,94	0,93	0,92	0,93	0,92	0,80

Tabela 5.14 - Estatísticas de desempenho, para correlação cruzada – configuração (2), dos algoritmos de previsão de severidade dos EMC com dados de *input* das 13Z.

BALANCEADO (50% "SIM"/ 50% "NÃO")	13Z - ÁREA2											
	POD (S)	1-FAR(S)	POD(N)	1-FAR(N)	F-MEASURE(S)	BIAS(S)	F-MEASURE(N)	BIAS(N)	POD (MD)	1-FAR (MD)	F-MEASURE(MD)	KAPPA
J48	0,98	0,86	0,86	0,98	0,92	1,12	0,91	0,88	0,92	0,92	0,92	0,84
RandomForest	0,98	0,87	0,87	0,98	0,93	1,11	0,92	0,89	0,93	0,93	0,93	0,85

5.3.2 Estudo de caso

Nas Tabelas 5.15 e 5.16 são, respectivamente, apresentados para as áreas I e II os registros diários dos valores dos índices termodinâmicos das 15Z (área I) e 13Z (área II) (colunas 2 à 6), presença de DA (coluna 7), previsão dos modelos preditivos de EMC e não-EMC (coluna 8) no mês de abril de 2019. Ressalta-se que os dias 8, 14, 15, 23 e 24 os dados dos índices de instabilidade do GOES-R não foram disponibilizados.

Objetivando realizar o *hand cast* de eventos convectivos severos (aqui classificados como os dias com DA maior ou igual a 130 raios) e não-eventos (dias com DA inferior à 130 raios), foram executados os modelos para ambas as áreas com índices de instabilidades das 13Z, conforme Tabelas 5.14 e 5.15. Os resultados revelam que os modelos obtiveram expressivos e similares *hand casts* e sendo capazes de identificarem, com 8h de antecedência, corretamente 100 % dos dias não-EMC e 96% dos dias de EMC. Os erros das previsões observados correspondem apenas a um dia de EMC para cada área, ou seja, dia 13, na área I, e dia 28, na área II, conforme coluna 8 das Tabelas 5.15 e 5.16, respectivamente. Visando avaliar as falhas dos modelos, analisaram-se os valores dos índices de instabilidade e percebe-se que estes são compatíveis aos que ocorrem em dias

típicos de EMC, com a exceção do valor do índice CAPE, que é igual a zero e LI positivo (atípico para dias EMC conforme dados das Tabelas). Intuitivamente, uma razão plausível, é que não houve o devido controle de qualidade nos dados e, assim, o cálculo do valor do CAPE e LI foram obtidos com dados corrompidos resultando em *hand casts* errôneas para os dias específicos.

Tabela 5.15 - Valores dos índices termodinâmicos e descargas atmosféricas para a área I no mês de abril de 2019.

DATA	CAPE	KI	LI	SI	TT	RAIOS	PREVISÃO
01/04	4,54	26,47	0,52	4,73	40,66	0	SIM
02/04	0,00	13,23	4,68	8,35	35,49	0	SIM
03/04	0,00	20,81	2,00	6,95	37,89	0	SIM
04/04	0,00	22,28	3,75	6,00	37,21	0	SIM
05/04	44,78	35,28	-0,99	0,25	44,45	985	SIM
06/04	65,06	35,61	-1,06	-0,30	43,66	29271	SIM
07/04	159,55	35,32	-1,94	-0,02	43,94	3655	SIM
08/04	SEM INFORMAÇÃO						
09/04	4,98	32,04	0,01	0,66	43,16	9	SIM
10/04	1,62	25,81	1,43	3,22	41,29	3	SIM
11/04	0,01	23,10	2,63	5,06	38,61	0	SIM
12/04	0,00	18,55	3,16	6,27	37,06	0	SIM
13/04	0,00	24,24	0,54	3,56	41,09	175	NÃO
14/04	SEM INFORMAÇÃO						
15/04	SEM INFORMAÇÃO						
16/04	92,83	31,97	-0,88	0,98	43,85	26244	SIM
17/04	247,71	25,40	-0,33	1,54	43,63	5450	SIM
18/04	2,24	1,49	4,38	8,83	36,48	0	SIM
19/04	2,59	-9,03	5,32	10,94	35,96	0	SIM
20/04	0,00	3,89	7,44	10,90	31,99	0	SIM
21/04	0,00	23,85	3,77	3,85	40,04	88	SIM
22/04	5,25	32,04	0,41	1,39	43,59	0	SIM
23/04	SEM INFORMAÇÃO						
24/04	SEM INFORMAÇÃO						
25/04	20,84	27,45	0,66	4,16	39,51	4	SIM
26/04	78,03	29,23	-0,60	2,31	42,23	849	SIM
27/04	5,32	31,71	-0,03	1,23	42,92	0	SIM
28/04	7,45	29,49	-0,05	1,98	43,30	6286	SIM
29/04	0,00	18,86	2,42	5,70	38,34	4	SIM
30/04	0,00	3,28	3,53	8,37	37,46	0	SIM

Tabela 5.16 - Valores dos índices termodinâmicos e descargas atmosféricas para a área II no mês de abril de 2019.

DATA	CAPE	KI	LI	SI	TT	RAIOS	PREVISÃO
01/04	16,30	25,71	0,37	5,60	40,80	0	SIM
02/04	0,00	12,21	3,61	9,02	35,42	0	SIM
03/04	0,27	15,31	0,66	7,59	38,97	0	SIM
04/04	0,00	15,44	2,31	8,75	33,70	0	SIM
05/04	21,07	31,21	-0,77	2,46	42,41	2620	SIM
06/04	16,70	35,89	-0,75	-0,02	43,42	345	SIM
07/04	282,99	26,86	-2,57	2,87	43,59	5735	SIM
08/04	SEM INFORMAÇÃO						
09/04	55,59	33,98	-0,66	0,91	42,94	1743	SIM
10/04	5,18	29,42	0,25	2,66	41,72	0	SIM
11/04	0,00	23,31	2,13	6,83	36,82	0	SIM
12/04	0,00	18,25	3,40	8,46	34,02	0	SIM
13/04	4,89	25,47	0,62	4,07	39,55	0	SIM
14/04	SEM INFORMAÇÃO						
15/04	SEM INFORMAÇÃO						
16/04	31,69	29,03	-0,29	2,58	41,99	28	SIM
17/04	410,52	27,65	-1,44	1,34	44,29	17407	SIM
18/04	23,58	9,39	2,58	7,67	36,42	11	SIM
19/04	53,33	0,81	2,35	7,71	39,74	0	SIM
20/04	0,00	-18,02	3,66	14,15	29,21	0	SIM
21/04	0,00	19,00	5,12	5,92	37,27	0	SIM
22/04	1,72	29,14	0,40	3,09	41,75	0	SIM
23/04	SEM INFORMAÇÃO						
24/04	SEM INFORMAÇÃO						
25/04	37,47	25,44	0,24	5,38	38,68	0	SIM
26/04	16,43	24,55	0,43	5,52	39,09	0	SIM
27/04	3,36	28,72	1,25	3,72	40,22	0	SIM
28/04	0,00	26,38	1,67	3,78	40,56	1205	NÃO
29/04	1,82	21,07	0,72	4,72	39,72	5	SIM
30/04	0,70	10,60	2,93	5,72	38,16	0	SIM

6. CONCLUSÕES

Neste trabalho foi desenvolvido um modelo de previsão, baseado em IA, de detecção e severidade de eventos meteorológicos convectivos, para a rota RJ – SP. Os principais resultados obtidos são resumidos a seguir:

- As descargas atmosféricas ocorrem em maior frequência no verão e são associadas preferencialmente às ilhas de calor, como a grande São Paulo, e à influência orográfica, como na região próximo à cidade de Resende;
- Foi identificado o período das 18-00h como o de maior ocorrência de atividade convectiva;
- Os modelos que têm início às 10h (área II) e 12h (área I) local são os de melhores desempenhos;
- O modelo desenvolvido demonstra capacidade de prever e classificar a severidade (com 8h de antecedência), com alta probabilidade de detecção, baixa taxa de falso alarme e pouco enviesado, os EMC e não-EMC;
- Os resultados de *hand casts* de 6h e 8h dos modelos revelam desempenhos satisfatórios para identificarem EMC e não-EMC para área estudada.

Como trabalhos futuros, sugere-se que a expansão da amostra de dados visando a consistência estatística dos resultados e recomenda-se o devido controle de qualidade dos dados para qualquer ação visando o teste de operacionalidade dos modelos aqui desenvolvidos.

7. REFERÊNCIAS

AFFONSO, E. T. F.; SILVA, A. M.; SILVA, M. P.; RODRIGUES, T. M. D.; MOITA, G. F. Uso redes neurais MultiLayer Perceptron (MLP) em sistema de bloqueio de websites baseado em conteúdo. Mecánica Computacional, v. XXIX, p. 9075-9090, 2010.

ALMEIDA, MANOEL VALDONEL. Aplicação de técnicas de redes neurais artificiais na previsão de curtíssimo prazo da visibilidade e teto para o aeroporto de guarulhos – SP. 2009. 206 f. Tese (Doutorado) - Curso de Engenharia Civil, Coppe, Ufrj, Rio de Janeiro, 2009.

ALMEIDA, VINÍCIUS ALBUQUERQUE; FRANÇA, GUTEMBERG BORGES; VELHO, HAROLDO FRAGA DE CAMPOS. Novel short-range forecasting system for meteorological convective events in Rio de Janeiro using remote sensing of atmospheric discharges. International Journal of Remote Sensing, 2020. DOI 10.1080/01431161.2020.1717669. Disponível em: https://www.researchgate.net/publication/340511327_Short-range_forecasting_system_for_meteorological_convective_events_in_Rio_de_Janeiro_using_remote_sensing_of_atmospheric_discharges.

BENDER, ANDRÉIA. Condições Atmosféricas Conducentes a Tempestades Severas e sua Relação com a Urbanização na RMSP. 2018. 129 f. Tese (Doutorado) - Curso de Ciências, Departamento de Ciências Atmosféricas, Universidade de São Paulo, São Paulo, 2018.

BISHOP, C.M., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford, p. 504.

BLUESTEIN, H. B.: Synoptic-Dynamic Meteorology in Midlatitudes. Volume II: Observations and Theory of Weather Systems. Oxford University Press, New York, EUA, 1993. 594 p.

BOUCKAERT, R.R.: Bayesian Network Classifiers in WEKA for Version 3-5-7. University of Waikato (2007).

BREIMAN, L. Machine Learning (2001) 45: 5. https://doi.org/10.1023/A:1010933404324.

BROOKS, H.E., LEE, J.W. e CRAVEN, J.P. 2003. The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. Atmospheric Research, 67, pp.73-94.

BROWN, B.G., Murphy, A.H., 1996. Verification of aircraft icing forecasts: the use of standard measures and meteorological covariates. Preprints, 13th Conf. Probability and Statistics in the Atmospheric Sciences, San Francisco, California, USA, Amer. Meteorol. Soc., pp. 251 – 252.

BROWNLEE, J. Master Machine Learning Algorithms. Discover How They Work and Implement Them from Scratch; Machine Learning Mastery: Vermont, VIC, Australia, 2016.

CENIPA. Aeródromos - Sumário Estatístico 2008-2017. Disponível em: file:///C:/Users/Windows%2010/Downloads/sumario_estatistico_aerodromos.pdf.

COHEN, J. A Coeficient of Agreement for Nominal Scales. Educational and Measurment. v. XX, n. 1, p. 37-46, 1960.

DINES, W. H. (1917). Meteorology and aviation. Monthly Weather Review, 45, 401.

DOC9750. ICAO: Global Air Navigation Plan 2016-2030. Montreal, 2016.

Equipe de Desenvolvimento do QGIS (2019). Sistema de Informações Geográficas do QGIS. Projeto Código Aberto Geospatial Foundation. http://qgis.osgeo.org.

FAUSETT, L., 1994. Fundamentals of Neural Networks. Architectures, Algorithms, and Applications. Prentice-Hall, Upper Saddle River, NJ, p. 462.

FRANÇA, G. B., M. V. ALMEIDA, and A. C. ROSSETE. 2016. “An Automated Nowcasting Model of Significant Instability Events in the Flight Terminal Area of Rio De Janeiro, Brazil.” Atmospheric Measurement Techniques 9: 2335–2344. doi:10.5194/amt-9-2335-2016.

FRANÇA, G. B., M. V. ALMEIDA, BONNET S. M., and ALBUQUERQUE NETO F. A.: Nowcasting model of low wind profile based on neural network using SODAR data at Guarulhos Airport, Brazil, International Journal of Remote Sensing, 39:8, 2506-2517, doi: 10.1080/01431161.2018.1425562, 2018.

FRANK, EIBE; WITTEN, IAN H. “Generating Accurate Rule Sets Without Global Optimization”. In: Proceedings of the Fifteenth international Conference on Machine Learning (July 24 - 27, 1998). J. W. Shavlik, Ed. Morgan Kaufmann Publishers, San Francisco, CA, 144-151.

GALWAY, J.G. 1956. The lifted index as a predictor of latent instability. Bulletin of the American Meteorological Society, 43, 528-529.

GEORGE H. JOHN, PAT LANGLEY, Estimating continuous distributions in Bayesian classifiers, Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, p.338-345, August 18-20, 1995, Montréal, Qué, Canada.

GEORGE, J. J. Weather Forecasting for Aeronautics. Academic Press, 673 pp., 1960.

GROSSMAN, K.S. 2010. Estudo da combinação de índices de instabilidade como ferramenta de auxílio na previsão do tempo. Departamento de Meteorologia (Instituto de Geociências). Universidade Federal do Rio de Janeiro, 68p.

GOES-R SERIES PRODUCT DEFINITION AND USERS’ GUIDE. Disponível em: <https://www.goes-r.gov/products/docs/PUG-L2%2B-vol5.pdf>

GULTEPE, I.; SHARMAN, R.; WILLIAMS, P.D.; ZHOU, B.; ELLROD, G.; MINNIS, P.; TRIER, S.; GRIFFIN, S.; YUM, S.S.; GHARABAGHI, B.; *et al*. A review of high impact weather for aviation meteorology. Pure Appl. Geophys. 2019, 176, 1869–1921.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. E WITTEN, IH (2009). O software de mineração de dados WEKA: uma atualização. *Boletim de Explorações da ACM SIGKDD*, *11* (1), 10–18.

HALLAK, C. & FILHO, A.J.P., 2012, Análise do desempenho de índices de instabilidade atmosférica na previsão de fenômenos convectivos de mesoescala na região metropolitana de São Paulo entre 28 de Janeiro e 04 de fevereiro de 2004. Revista Brasileira de Meteorologia, v.27, n.2. pag 173-206.

HAYKIN, S. 2001. Redes neurais: princípios e prática. Porto Alegre: Bookman.

HAYKIN, S. (2002). Neural Networks. A Comprehensive Foundation, Macmillan, New York, NY.

HERMSDORFF, JULIANA. previsão de instabilidade atmosférica significativa usando árvore de decisão na região metropolitana do Rio de Janeiro. 2018. 109 f. Dissertação (Mestrado) - Curso de Meteorologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2018.

HOUZE, R. A: Cloud Dynamics. Academic Press, 573 pp, 1993.

HULTEN GEOFF, LAURIE SPENCER, and PEDRO DOMINGOS. Mining time-changing data streams. In Proc. 2001 ACM SIGKDD *Intl. Conf. On Knowedge Discovery and Data Mining*, pages 97-206, 2001.

ISAAC, G. A., BAILEY, M., COBER, S. G., DONALDSON, N., DRIEDGER, N., GLAZER, A., GULTEPE, I., HUDAK, D., KOROLEV, A., REID, J., RODRIGUEZ, P., STRAPP, J. W., and FABRY, F.: Airport Vicinity Icing and Snow Advisor, in: AIAA 44th Aerospace Sci. Meeting and Exhibit, Reno Nevada, AIAA-2006-1219, 2006.

ISAAC, G. A., BOUDALA, F., COBER, S. G, CRAWFORD, R., DONALDSON, N., GULTEPE, I., HANSEN, B., HECKMAN, I., HUANG, L., LING, A., REID, J., and FOURNIER, M.: Decision Making Regarding Aircraft De-Icing and In-Flight Icing

Using the Canadian Airport Nowcasting System (CAN-Now), in: SAE 2011 International Conference on Aircraft and Engine Icing and Ground Deicing, 13–17 June 2011, Paper Number: 2011-38- 0029; doi:10.4271/2011-38-0029, 2011.

ISAAC, G. A., BURROWS, W. R., COBER, S. G., CRAWFORD, R. W., DONALDSON, N., GULTEPE, I., HANSEN, B., HECKMAN, I., HUANG, L. X., LING, A., MAILHOT, J., MILBRANDT, J. A., REID, J., and FOURNIER, M.: The Canadian airport nowcasting system (CANNow), Meteorol. Appl., 21, 30–49, 2014.

JOHN, GH E P. LANGLEY (1995). Estimando distribuições contínuas em classificadores bayesianos. Em P. Besnard e S. Hanks (Eds.), *Procedimentos da Décima Primeira Conferência sobre Incerteza em Inteligência Artificial* (pp. 338-345). San Francisco, CA: Morgan Kaufmann.

JOURDAN, P. et al. Relações Espaciais e Temporais entre o Desligamento de uma Linha de Transmissão e a Detecção de Descargas Atmosféricas em casos de ruptura do cabo OPGW. XXIV SNPTEE – Seminário Nacional de Produção e Transmissão de Energia Elétrica. 22 a 25 de outubro de 2017. Curitiba, PR.

KALNAY, E.; e diversos autores. The NCEP/NCAR 40-year reanalysis project. Bull. Amer. Meteor. Soc., v. 77, p. 437-471, 1996.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: **International joint Conference on artificial intelligence**. [S.l.: s.n.], 1995. v. 14, p. 1137–1145.

KOTZ S., JOHNSON N. L. Encyclopedia of statistical sciences. New York: John Wiley& Sons. v.4, p.352-354, 1983.

KUMAR, Y.; SAHOO, G. Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA. Information Technology and Computer Science, v. 7, p. 43-49, 2012.

LIU, J. G.; MASON, P. J. Image Processing and GIS for Remote Sensing: Techniques and Applications. Wiley Blackwell, 2016.

MCCANN, D., 1992: A neural network short-term forecast of significant thunderstorms. Wea. Forecasting, 7, 525–534, https://doi.org/ 10.1175/1520-0434(1992)007,0525:ANNSTF.2.0.CO;2.

MILLER, R. C. Notes on analysis and severe storm forecasting procedures of the Air Force Global Weather Central. Tech. Report 200, Air Weather Service, United States Air Force, 190 pp., 1972.

MOURA, C.R.W.; ESCOBAR, G.C.J.; ANDRADE, K.M. 2013. Padrões de circulação em superfície e altitude associados a eventos de chuva intensa na Região Metropolitana do Rio de Janeiro. Revista Brasileira de Meteorologia, v.28, n.3, p. 267 – 280.

MOURÃO, C. E. F. & MENEZES, W. F., 2006. "Estudo do comportamento de indicadores de tempo severo em casos de tempestades sobre o Rio de Janeiro". Anais do XIV Congresso Brasileiro de Meteorologia. Florianópolis-SC.

MUELLER, C., SAXEN, T., ROBERTS, R., WILSON, J., BETANCOURT, T., DETTLING, S., OIEN, N., and YEE, J.: NCAR auto-nowcast system, Weather Forecast, 18, 545–561, 2003.

NASCIMENTO, E. L.: Previsão de tempestades severas utilizando-se parâmetros convectivos e modelos de mesoescala: uma estratégia operacional adotável no Brasil? Revista Brasileira de Meteorologia, v. 20, p. 121-140, 2005.

NUNES, LUCÍ HIDALGO. Urbanização e Desastres Naturais. São Paulo: Oficina de Textos, 2015.

PAULUCCI, T. B. 2017. Caracterização Espaço-Temporal de Descargas Atmosféricas e Tempestades Elétricas na Região Metropolitana do Rio de Janeiro entre 2001 e 2016. Curso de Meteorologia, Universidade Federal do Rio de Janeiro, Monografia, 73p.

PAULUCCI, T. B.; GUTEMBERG, F.; LIBONATI, R.; RAMOS, A. M.: Long-Term Spatial– Temporal Characterization of Cloud-to-Ground Lightning in the Metropolitan Region of Rio de Janeiro. Pure and Applied Geophysics. 10.1007/s00024-019- 02216-1, 2019.

Quantum GIS User Guide – Version 2.8.1. Disponível para download em: <http://www.qgis.org/en/site/>

QUINLAN, J. R. 1993. C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, California.

SARAIVA, Jaci (2005). O projeto Sivam, o Sipam e a sua contribuição para a segurança da navegação aérea na Amazônia legal. Boletim da Sociedade Brasileira de Meteorologia. Volume 29. 27-30.

SASAKI, Y. (2007). "The truth of the F-measure".

SHOWALTER, A. K.: A stability index for forecasting thunderstorms. Bulletin of the American Meteorological Society, v. 34, p. 250-252, 1947.

SUMNER M., FRANK E., HALL M. (2005) Acelerando a Indução de Árvore Modelo Logística. Em: Jorge AM, Torgo L., Brazdil P., Camacho R., Gama J. (eds) Descoberta do conhecimento em bancos de dados: PKDD 2005. PKDD 2005. Notas de aula em ciência da computação, vol 3721. Springer, Berlim, Heidelberg.

Technical Regulations Basic Documents No. 2 Volume II – Meteorological Service for International Air Navigation.
Disponível em: https://library.wmo.int/index.php?lvl=notice_display&id=5790#.XZQSjEZKiUn.
Acessado em: 02 de outubro de 2019.

UNIVERSITY OF WAIKATO. WEKA 3 – Machine Learning Software in Java. Disponível no site da University of Waikato (2010). URL: http://www.cs.waikato.ac.nz/ml/WEKA.

VIERA, A. J.; MD; GARRETT, J. M. Understanding Interobserver Agreement: The KAPPA Statistic. Family Medicine Journal, v. 37, n. 5, p. 360-363, 2005.

WILK, K. E. and GRAY, K. C.: Processing and analysis techniques used with the NSSL weather radar system, in: Preprints, 14th Conf. on Radar Meteorology, Tucson, AZ, 520 American Meteorological Society, 369–374, 1970.

WILKS, S. D. – Statistical Methods in the Atmospheric Sciences. 2ª ed., New York, USA, Academic Press, 2006.

WILSON, J. W.: Movement and predictability of radar echoes, Tech. Memo.: ERTM-NSSl-28, National Severe Storms Laboratory, Springfield, VA, USA, 30 pp., 1966.

ZHANG, H. The Optimality of Naive Bayes. American Association for Artificial Intelligence, 2004.