UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

NICKSSON CKAYO ARRAIS DE FREITAS

UMA ABORDAGEM DE MINERAÇÃO DE DADOS PARA ESTIMATIVA DA VELOCIDADE DO VENTO

MOSSORÓ - RN

2018

NICKSSON CKAYO ARRAIS DE FREITAS

UMA ABORDAGEM DE MINERAÇÃO DE DADOS PARA ESTIMATIVA DA VELOCIDADE DO VENTO

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação - associação ampla entre a Universidade do Estado do Rio Grande do Norte e a Universidade Federal Rural do Semi-Árido, para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Marcelino Pereira dos Santos Silva
Coorientadora: Profª. Drª. Meiry Sayuri Sakamoto

MOSSORÓ - RN

2018

Catalogação da Publicação na Fonte.
Universidade do Estado do Rio Grande do Norte.

F866a Freitas, Nicksson Ckayo Arrais de
UMA ABORDAGEM DE MINERAÇÃO DE DADOS PARA ESTIMATIVA DA VELOCIDADE DO VENTO. / Nicksson Ckayo Arrais de Freitas. - Mossoró, Rio Grande do Norte, Brasil., 2018.
92p.

Orientador(a): Prof. Dr. Marcelino Pereira dos Santos Silva.
Coorientador(a): Profa. Dra. Meiry Sayuri Sakamoto.
Dissertação (Mestrado em Programa de Pós-Graduação em Ciência da Computação). Universidade do Estado do Rio Grande do Norte.

1. Recursos Renováveis. 2. Energia Eólica. 3. Mineração de Dados. 4. Bancos de Dados. 5. Inteligência Artificial.. I. Silva, Marcelino Pereira dos Santos. II. Universidade do Estado do Rio Grande do Norte. III. Título.

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC´s) foi desenvolvido pela Diretoria de Informatização (DINF), sob orientação dos bibliotecários do SIB-UERN, para ser adaptado às necessidades da comunidade acadêmica UERN.

NICKSSON CKAYO ARRAIS DE FREITAS

Uma Abordagem de Mineração de Dados para Estimativa da Velocidade do Vento.

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do título de Mestre em Ciência da Computação.

APROVADA EM: 08 / 02 / 2018.

Prof. Dr. Marcelino Pereira dos Santos Silva
Orientador e Presidente

Profa. Dra. Carla Katarina de Monteiro Marques
Universidade do Estado do Rio Grande do Norte

Prof. Dr. Jerffeson Teixeira de Souza
Universidade Estadual do Ceará

A minha família e meus amigos

AGRADECIMENTOS

Agradeço primeiramente à Deus por me dar forças nesta caminhada.

A minha família agradeço pelo apoio diário, suporte, prontidão e paciência que tiveram comigo. Em especial, agradeço a minha mãe, Nivea Rivânia Arrais de Freitas, por nunca desistir de mim, sempre confiar no meu potencial e me incentivar nos momentos que mais necessitei. Agradeço ao meu pai, Carlos Feitosa de Freitas pelo suporte e conselhos que levarei por toda a vida. Agradeço a minha amiga, namorada e noiva, Larissa Fernandes de Oliveira, pelo suporte, compreensão, colaboração, paciência e prontidão. Aos meus avós, Antônio Noronha de Freitas e Maria de Fátima Feitosa por me ajudarem nos momentos que muito necessitei.

Sou muito grato ao amigo, professor e orientador Prof. Marcelino Pereira dos Santos pela paciência, confiança, conselhos, orientação, acompanhamento e todos os demais elementos que contribuíram para minha evolução como pessoa e cientista. Desde há algum tempo, tem compartilhado seus conhecimentos e sua experiência com muita serenidade e profissionalismo. Também deixo meus agradecimentos para amiga e coorientadora Meiry Sayuri Sakamoto pela parceria, confiança, ajuda, prontidão e paciência que foram fundamentais para o desenvolvimento de toda a pesquisa científica.

Agradeço ao amigo Átila Negreiros Maia, por sua colaboração no primeiro ano de projeto e pela amizade de longa data, “tamo junto!”. Além disso, deixo meus agradecimento para os meus colegas de turma do mestrado em ciência da computação (UERN/UFERSA), por contribuírem cada um de sua forma durante as atividades do programa.

A todos os professores que, de alguma forma, contribuíram para o meu crescimento com seus ensinamentos. No departamento de Computação da UERN, deixo minha homenagem para os professores Marcelino Pereira dos Santos e Antônio Oliveira Filho, e as professoras Carla Katarina de Monteiro Marques e Cicilia Raquel Maia Leite.

Obrigado a Fundação Cearense de Meteorologia e Recursos Hídricos (FUNCEME), a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e as empresas eólicas que contribuíram com subsídios e dados durante a execução do projeto.

Muito obrigado a todos que não tiveram nomes mencionados, mas contribuíram de uma forma ou outra em minha formação acadêmica.

“A intermitência do sonho
é que nos permite suportar
os dias de trabalho”.
(Pablo Neruda)

RESUMO

Recursos renováveis são as alternativas mais promissoras para geração de energia, considerando que o uso de combustíveis fósseis tem causado fortes impactos no ecossistema terrestre e no clima. Como um recurso para produção de eletricidade, indústrias eólicas têm levado vantagem em relação às outras fontes e, consequentemente, a capacidade de geração dessas indústrias vem crescendo no mundo inteiro. No entanto, previsões de energia são elementos cruciais para os operadores de sistemas elétricos, pois permitem-os tomarem melhores decisões relacionadas ao mercado elétrico e às suas atividades operacionais. Vale salientar-se que a saída de potência dos parques eólicos depende da natureza estocástica do vento, um recurso natural, intermitente, incerto e incontrolável. De fato, estimativas consistentes da velocidade do vento podem evitar prejuízos, garantir a oferta segura e sustentável de eletricidade, facilitar a regulamentação de sistemas eólicos e aumentar a produtividade operacional nas indústrias através de uma tomada de decisão mais confiável. Todavia, a previsão de vento é um problema complexo e desafiador devido à falta de ferramentas apropriadas e aos eventos que influenciam as suas condições como rotação da terra, efeitos físicos e fatores climáticos. Para propor soluções neste contexto, ainda devemos considerar que dados meteorológicos têm acumulado enormes volumes de informação nos bancos de dados espaciais, o que demanda a investigação de meios relevantes para extração de informação estratégica. A tecnologia de mineração de dados constitui-se em solução para extrair, de forma semiautomática e inteligente, conhecimento relevante de enormes conjuntos de dados. Este trabalho apresenta uma nova abordagem de mineração de dados para previsão da velocidade do vento que tem baixo custo, contempla relevantes algoritmos de inteligência artificial e fornece recursos eficientes para tratamento de bancos de dados. No geral, a abordagem tem se mostrado promissora, flexível e bem fundamentada nos dois estudos de casos realizados no Brasil. Redes neurais, máquina de vetores de suporte, árvore de decisão e k-vizinho mais próximos são métodos envolvidos na construção de diversos modelos de previsão da velocidade do vento.

Palavras-chave: Recursos Renováveis, Energia Eólica, Velocidade do Vento, Mineração de Dados, Bancos de Dados, Inteligência Artificial.

ABSTRACT

Renewable sources are the most promising alternatives for power generation, whereas the use of fossil fuels has caused strong impacts on terrestrial ecosystems and the climate. Wind industries, as a power source, have advantages over other sources, as a consequence, wind energy generation capacity had a tremendous growth worldwide. However, energy forecasts are crucial elements for electrical system operators, because they can make better decisions on the electrical market and support operational activities. It is worth emphasizing that the output of energy from wind farms depends on the stochastic nature of the wind, which is a natural, intermittent, uncertain and difficult-to-control resource. In fact, wind speed prediction may avoid economic losses, ensure the safe and sustainable supply of electricity, facilitate regulation of wind systems, and increase the operational efficiency of industries through a more reliable decision making. Wind speed prediction is a complex and challenging problem due to the lack of appropriate tools and the events that influence wind conditions like earth moving, physical effects, and climatic factors. For proposing solutions in this context, we must consider that weather data have accumulated huge volumes of information in spatial databases, demanding the investigation of relevant means for knowledge extraction. Data mining arises as a solution to extract relevant knowledge intelligently and semi-automatically from huge datasets. This paper presents a new and low-cost data mining approach for wind speed forecasting, which incorporates relevant artificial intelligence algorithms and provides effective treatment of datasets. The approach has proven to be flexible, promising, and well-founded in two case studies carried out in Brazil. Neural networks, support vector machines, decision trees, and k-nearest neighbors are methods involved in building the diverse models for wind speed estimation.

Keywords: Renewable Sources, Wind Energy, Wind Speed, Prediction, Databases, Data Mining, Artificial Intelligence.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CO	Camadas Ocultas
CPTEC	Centro de Previsão de Tempo e Estudos Climáticos
DV	Direção do Vento.
DV_n	Direção do Vento Nominal.
EIA	*U.S. Energy Information Administration's*
EPE	Empresa de Pesquisa Energética
FUNCEME	Fundação Cearense de Meteorologia e Recursos Hídricos
GWEC	*Global Wind Energy Council*
HD	Hora do Dia
IA	Inteligência Artificial
IDC	*International Data Corporation*
INMET	Instituto Nacional de Meteorologia
INPE	Instituto Nacional de Pesquisas Espaciais
KDD	*Knowledge Discovery in Databases*
KNN	*K-Nearest Neighbors*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MLP	*Multilayer Perceptron*
MME	Ministério de Minas e Energia
MRC	*Meteorological Resource Center*
MSE	*Mean Square Error*
MS	Mês
MS_n	Mês Nominal
NOAA	*National Oceanic and Atmospheric Administration*

NN	*Neural Networks*
NNR	*Neural Network Recurrent*
NS	Número da Semana
PA	Pressão atmosférica
RL	Regressão Linear
RMSE	*Root Mean Square Error*
SINDA	Sistema Integrado de Dados Ambientais
SMAPE	*Symmetric Mean Absolute Percentage Error*
SMO	*Sequential Minimal Optimization*
SONDA	Sistema de Organização Nacional de Dados Ambientais
SVM	*Support Vector Machine*
TP	Temperatura do Ar
UR	Umidade Relativa
USGS	*United States Geological Survey*
VV	Velocidade do Vento
WEKA	*Waikato Environment for Knowledge Analysis*

LISTA DE SÍMBOLOS

CO_2	Dióxido de Carbono
°C	Graus Celsius
Kg/m^3	Quilograma por Metro Cúbico
mb	Milibar
m^2	Metros Quadrados
m/s	Metros por Segundo
m/s^2	Metros por Segundo ao Quadrado
m	Metros
mw	Megawatts
s	Segundos

SUMÁRIO

1 INTRODUÇÃO

A demanda por energia vem crescendo de forma acelerada em todo o mundo devido ao aumento populacional e à evolução industrial. A geração de energia através da queima de combustíveis fósseis tornou-se insustentável e nociva, pois libera dióxido de carbono na atmosfera em grandes quantidades, promovendo um forte desequilíbrio nos mais diversos ecossistemas terrestres e prejudicando a saúde humana.

Diante deste cenário, a maioria dos países criaram uma perspectiva global para expandir suas matrizes nacionais por meio de recursos renováveis, visando combater a poluição do meio ambiente, o aumento no preço dos combustíveis fósseis e uma possível escassez desses.

A energia eólica vem se desenvolvendo rapidamente no mundo inteiro devido às suas vantagens sobre as outras fontes renováveis na geração em larga escala. De fato, a evolução nos equipamentos, a possibilidade de produção elétrica vinte quatro horas por dia, os subsídios governamentais e a redução nos custo de instalação são fatores que motivam os altos investimentos nas indústrias eólicas.

No Brasil, a expansão eólica tem promovido benefícios sociais, econômicos e ambientais. Em particular, a instalação de novas indústrias tem proporcionado o aumento na oferta de empregos (principalmente no Nordeste), a redução da emissão de gases do efeito estufa, a redução da dependência majoritária de uma única fonte renovável (hidrelétrica) e o crescimento econômico através de parcerias internas e externas.

No entanto, a inconsistência, a imprevisibilidade e a insegurança na oferta de energia são os problemas persistentes que têm freado o maior progresso desses sistemas. A produção eólica depende diretamente da força dos ventos, um recurso natural e inesgotável, porém intermitente, imprevisível e incontrolável.

As estimativas da velocidade do vento são requisitos para o funcionamento eficiente dos sistemas eólicos, pois permitem aos operadores gerenciar a oferta segura de energia, tomar decisões confiáveis no comércio, determinar manutenções e aumentar a eficácia das turbinas eólicas. Atualmente, a falta de ferramentas para previsão da velocidade do vento tem dificultado as operações e a regulamentação dos sistemas eólicos. Além do mais, prever as condições de ventos é um problema complexo, considerando que o movimento do ar é originado pela diferença de pressão entre regiões e influenciado por fatores físicos e climáticos, o que demanda recursos tecnológicos apropriados.

Para propor soluções neste contexto, devemos considerar que dados meteorológicos têm sido acumulados em enormes volumes nos bancos de dados espaciais, uma vez que são captados a todo momento através de instrumentos como satélites, barômetros, anemômetros, radares e veículos aéreos não tribulados. Embora existam profissionais especializados para analisar

manualmente cada modelo de dado (imagem, texto e planilhas), a detecção de padrões a partir de uma coleção de dados tão ampla e diversificada supera a capacidade racional humana. Nesse contexto, informações relevantes mantêm-se ocultas nos bancos de dados, demandando recursos poderosos que facilitem a extração de informação pelo homem.

As mudanças climáticas causando danos ao ambiente, o aumento na demanda elétrica, o amplo potencial a ser explorado no Brasil, além da carência de recursos apropriados para auxiliar os operadores de energia nas previsões da velocidade do vento a partir dos gigantescos volumes de dados são fatores que motivam a investigação de métodos, técnicas, ferramentas e algoritmos capazes de extrair conhecimento estratégico de bancos de dados maciços.

A tecnologia de mineração de dados fornece recursos relevantes para tratamento eficiente de repositórios, assim como algoritmos semiautomáticos de inteligência artificial capazes de extrair informação de enormes conjuntos de dados. O processo de mineração de dados é um domínio orientado à aplicação que tem promovido soluções interessantes relacionadas à previsão em diversas áreas, como medicina, física, biologia, comércio e indústria. Diante do exposto, essa tecnologia tem potencial para lidar com os problemas de previsão da velocidade do vento e pode fornecer contribuições significativas ao setor eólico.

Na literatura, algumas abordagens estão sendo executadas para previsão da velocidade do vento, porém muitas dessas ignoram aspectos de processamento e transformação nos dados, bem como recursos que tem potencial para melhorar os resultados de performance e precisão dos modelos preditivos. Nesta proposta, uma abordagem de mineração de dados que contempla etapas bem definidas para previsão da velocidade do vento é discutida. Na abordagem, enormes volumes de dados são manuseados de forma eficiente, inteligente e estratégica. Em seguida, algoritmos robustos e semi-automáticos são usados para construção de diversos modelos de previsão.

1.1 OBJETIVO

Este trabalho tem como principal objetivo apresentar uma nova abordagem estratégica de mineração de dados, que surgiu a partir de fundamentos de mineração de dados e limitações detectadas nas abordagens tradicionais da literatura, para construir modelos de previsão da velocidade do vento visando o avanço quantitativo e qualitativo na operação das indústrias eólicas.

1.2 OBJETIVOS ESPECÍFICOS

- Efetuar levantamento de estudos nacionais e internacionais referentes ao tema em questão;
- Estabelecer parcerias;
- Detectar aspectos prioritários ou limitações no setor eólico;
- Levantar bancos de dados espaciais disponíveis;
- Extrair, processar, transformar e minerar os bancos de dados relevantes;
- Propor soluções de baixo custo para os problemas prioritários das indústrias eólicas;
- Validar as soluções propostas;
- Determinar algoritmos relevantes para extração de conhecimento;
- Propor ferramentas e recursos computacionais para auxiliar os operadores de energia;
- Submeter trabalhos a periódicos ou conferências nacionais e internacionais.

1.3 ESTRUTURA DO DOCUMENTO

Os capítulos deste trabalho estão organizados da seguinte forma: o capítulo 2 apresenta os conceitos relacionados à meteorologia, à energia eólica e à mineração de dados, assim como uma revisão da literatura e trabalhos relacionados. O capítulo 3 contém a fundamentação da abordagem proposta, juntamente com a ferramenta utilizada para auxiliar na construção dos modelos de previsão. O capítulo 4 descreve os estudos de casos realizados para exemplificar e avaliar a abordagem proposta. Por fim, o capítulo 5 traz as conclusões finais e as sugestões de trabalhos futuros.

2 REFERENCIAL TEÓRICO

Nesta seção, conteúdos que embasam a proposta deste trabalho são apresentados, assim como trabalhos relacionados.

2.1 INTRODUÇÃO À METEOROLOGIA

A meteorologia (do grego meteoros que significa “elevado no ar”) é a ciência que estuda os fenômenos da atmosfera terrestre, precisamente as condições de tempo e clima. O tempo se refere ao estado momentâneo da atmosfera, enquanto que o clima é a integração das condições de tempo para um período mais extenso (REBOITA et al., 2012). Fazendo uma analogia, estamos falando sobre o tempo ao dizer que a previsão será de chuva à tarde em tal cidade. Ao mencionar que o inverno será chuvoso (ou seco) nos referimos ao clima.

A atmosfera terrestre é formada por cinco camadas de gases que envolvem a Terra, conforme apresentado na Figura 1. Estas camadas foram divididas principalmente com base na temperatura e suas evidências foram determinadas por estudos com balões meteorológicos, ondas de rádio, sistemas de foguetes e satélites (BARRY; CHORLEY, 2013). A troposfera é a região mais próxima da Terra que contém cerca de 75% da massa molecular (ou gasosa) da atmosfera. Inclusive, é a camada onde os fenômenos meteorológicos ocorrem mais acentuadamente, como chuvas, tempestades, relâmpagos, furacões e neve.

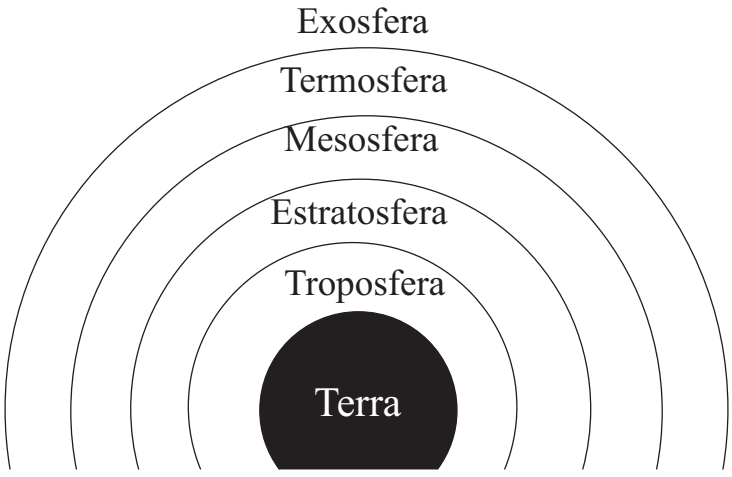

Figura 1 – Atmosfera terrestre

Fonte: Autoria Própria

Há cerca de 400 milhões de anos, a atmosfera vem modificando sua forma e composição. Consequentemente, examinar e compreender as mudanças que ocorrem no mundo, em geral,

é difícil. As condições de tempo, por exemplo, são descritas através da observação de vários elementos meteorológicos, tais como temperatura, umidade e pressão do ar, velocidade e direção do vento, tipos e quantidades de precipitação e de nuvem (AHRENS; HENSON, 2016).

Profissionais como meteorologistas e físicos usam princípios científicos e teorias para detectar e compreender um possível padrão comportamental nos eventos meteorológicos. Em seguida, eles disseminam explicações e informações relevantes para sociedade através dos veículos de comunicação, como internet, rádio, televisão e jornais.

Serviços meteorológicos incluem previsões meteorológicas, avisos públicos, consultas de informação e produtos para proteção e segurança. Tais serviços auxiliam diversos setores: na agricultura, facilitam o monitoramento das safras; no tráfego aéreo, possibilitam verificar as condições de voo; no comércio, auxiliam no processo de tomada de decisão. Além disso, eles são cruciais para detectar fenômenos destrutivos como tempestades, tornados e furacões, gerenciar recursos hídricos, monitorar áreas de incêndios florestais, analisar mudanças espaciais e atividades vulcânicas (TEXEIRA, 2016).

A meteorologia é relevante e seus estudos têm impacto direto na qualidade de vida das pessoas. Embora o seu avanço seja notório devido à evolução da tecnologia, uma série de problemas persistem. Nos Estados Unidos, por exemplo, ocorrem a cada ano cerca de 10.000 temporais, 5.000 inundações, 1.300 furacões, secas generalizadas, incêndios florestais e eventos climáticos, responsáveis por cerca de 90% de todos os desastres registrados e causam uma média de 65.020 mortes e 15 bilhões de dólares em danos por ano (NOAA, 2017). No Brasil, também sofremos com secas generalizadas, queimadas, enchentes e ciclones que provocam danos sociais, econômicos e ambientais.

Em decorrência desses problemas, a maioria dos países têm investido em estudos relacionados às ciências espaciais e atmosféricas. Estas iniciativas originaram diversas instituições especializadas com o objetivo de supervisionar a superfície terrestre continuamente, monitorando quando, como e onde os possíveis eventos meteorológicos poderão ocorrer. No Brasil, a Fundação Cearense de Meteorologia e Recursos Hídricos (FUNCEME), o Centro de Previsão de Tempo e Estudos Climáticos (CPTEC) do Instituto Nacional de Pesquisas Espaciais (INPE) e o Instituto Nacional de Meteorologia (INMET) são as instituições que mais se destacam na atuação em meteorologia. Em especial, a FUNCEME localizada no Ceará e fundada há mais de quatro décadas, tem por missão estudar meteorologia, recursos hídricos e ambientais, colaborando para o desenvolvimento sustentável do Nordeste, mais precisamente, do Ceará.

Embora a maioria dessas instituições dediquem-se diariamente a problemas espaciais, o apoio científico de pesquisadores é necessário e pode contribuir significativamente no entendimento de um problema ou determinada situação.

Atualmente, enormes volumes de dados estão disponíveis para a comunidade científica

na Web. No Brasil, dados espaciais podem ser encontrados no catálogo de imagem do INPE[1], no Sistema Integrado de Dados Ambientais (SINDA)[2], no Sistema de Organização Nacional de Dados Ambientais (SONDA)[3] e na base de dados no INMET[4]. Dados internacionais de satélites são oferecidos pelas duas plataformas do Levantamento Geológico dos Estados Unidos, do inglês *United States Geological Survey* (USGS)[5] [6].

A variedade de aplicações, a ampla disponibilidade de dados, a complexidade envolvida no monitoramento da superfície terrestre e o efeito direto na economia e infraestrutura da sociedade são fatores que têm motivado os estudos espaciais relacionados à meteorologia.

2.2 ENERGIA EÓLICA

Mudanças climáticas (ou aquecimento global) causam sérios impactos no planeta e são responsáveis por inúmeros problemas, como tempestades violentas, inundações, ciclones, chuva ácida e secas prolongadas. Estudos comprovam que as ações do homem são as principais responsáveis pelo aumento frenético na temperatura da Terra. Sobretudo, a queima de combustíveis fósseis na geração de energia libera na atmosfera gases do efeito estufa, principalmente CO_2.

O consumo de energia continua crescendo rapidamente devido ao aumento populacional, à evolução tecnológica industrial e à urbanização. Segundo a Administração de Informação de Energia dos Estados Unidos (EIA) (do inglês *U.S. Energy Information Administration*), o consumo mundial de energia aumentará 28% entre 2015 e 2040, ou seja, mais de um quarto da energia utilizada no mundo. Em relação aos recursos, a energia nuclear e o carvão terão uma projeção constante, enquanto que petróleo e gás natural terão um leve crescimento, conforme ilustrado na Figura 2.

Embora os combustíveis fósseis apresentem um pequeno aumento, as suas reservas estão diminuindo e provavelmente irão acabar com o passar do tempo. Enquanto isso, os custos da energia proveniente desses recursos tendem a subir cada vez mais. Neste cenário, as fontes renováveis ganharam força globalmente para combater as mudanças climáticas e o aumento no preço dos combustíveis poluentes.

Pensando nas futuras gerações, a maioria dos países vêm se esforçando para alcançar o desenvolvimento sustentável (ou social, econômico e ambiental) através da geração de energia

[1] http://www.dgi.inpe.br/CDSR/
[2] http://sinda.crn2.inpe.br
[3] http://sonda.ccst.inpe.br
[4] http://www.inmet.gov.br/
[5] https://earthexplorer.usgs.gov
[6] http://glovis.usgs.gov/

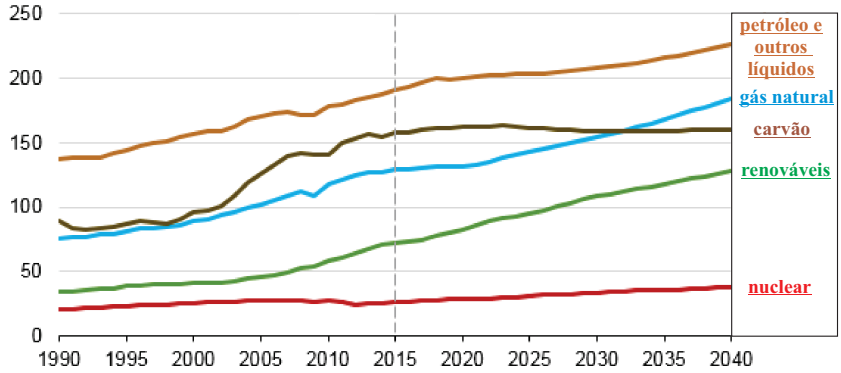

Figura 2 – Consumo de energia mundial entre 1990 e 2040

Fonte: Adaptada de (EIA, 2017)

provenientes de recursos naturais, tais como sol, vento, água e biomassa. Para a sociedade, uma diversificação e ampliação na geração de energia poderá fornecer tarifas mais baratas, energia limpa de melhor qualidade e também acabará com os problemas proporcionados pela dependência majoritária de uma única fonte de energia.

No Brasil, em 2001, ocorreu uma escassez energética devido à falta de planejamento governamental e à ampla dependência de uma única fonte de energia. Uma crise se alastrou causando danos irreparáveis, como redução do crescimento econômico, aumento no desemprego, aumento do déficit da balança comercial, perda de arrecadação de tributos, efeito inflacionário, além de incômodo com a privação de energia (TOLMASQUIM, 2000). Naquela época, aproximadamente 90% da energia produzida no Brasil tinha origem nas hidrelétricas, ou seja, era totalmente dependente das chuvas para manter os reservatórios de água em um nível adequado.

Diante do exposto, todos os recursos renováveis são relevantes para produção de energia, considerando que as matrizes energéticas são complementares e uma fonte pode suprir a deficiência da outra. De certa forma, os países tendem a uma particularidade, normalmente determinada por suas características climáticas. Felizmente, em razão da sua biodiversidade, o Brasil dispõe de potencial eólico, solar, hidrelétrico e de biomassa para produção de energia elétrica (PACHECO, 2006).

No entanto, a energia eólica vem apresentado muitas vantagens em relação às outras fontes renováveis: (1) a sua geração é dependente de um recurso inesgotável, o vento; (2) ao contrário dos sistemas solares, as indústrias eólicas produzem energia dia e noite; (3) instalação de novos parques eólicos significa aumento na oferta de emprego; (4) a geração eólica não produz resíduos poluentes ao meio ambiente (GAO et al., 2016); (5) indústrias eólicas têm vantagens econômicas quando comparado a outros sistemas para geração de energia em larga escala (COLAK; SAGIROGLU; YESILBUDAK, 2012); por último, (6) a energia do vento tem um alto custo-benefício social e ambiental, além de um ciclo de construção curto, baixa manutenção e flexibilidade para investimento (ZUO; LIU, 2012).

O Brasil, em particular, teve um crescimento significativo na produção de energia eólica. De acordo com o *Global Wind Energy Council* (GWEC), uma organização internacional especializada em energia eólica, o país ocupou o nono lugar no ranking mundial de capacidade de geração eólica acumulada e assumiu a quinta posição no mundo em potencial eólico instalado, alcançando 10.740 megawatts (mw) com um crescimento de 2.014 mw em relação ao ano de 2016 (GWEC, 2017). O ranking completo é apresentado na Tabela 1 para capacidade eólica acumulada e na Tabela 2 para potencial eólico instalado. Em ambos, a China lidera com boa vantagem em relação ao segundo colocado Estados Unidos.

Tabela 1 – Ranking mundial da capacidade de geração eólica acumulada

País	**mw**	**Percentual %**
China	168,690	34,7
Estados Unidos	82,184	16,9
Alemanha	50,018	10,3
Índia	28,700	5,9
Espanha	23,074	4,7
Reino Unido	14,543	3,0
França	12,066	2,5
Canadá	11,900	2,4
Brasil	10,740	2,2
Itália	9,257	1,9
Resto do Mundo	75,577	15,5
Total TOP 10	**411,172**	**84**
Total no Mundo	**486,749**	**100**

Fonte: Adaptada de (GWEC, 2017)

Tabela 2 – Ranking mundial do potencial eólico instalado

País	**mw**	**Percentual %**
China	23,328	42,7
Estados Unidos	8,203	15,0
Alemanha	5,443	10,0
Índia	3,612	6,6
Brasil	2,014	3,7
França	1,561	2,9
Turquia	1,387	2,5
Holanda	887	1,6
Reino Unido	736	1,3
Canadá	702	1,3
Resto do Mundo	6,727	12,3
Total TOP 10	**47,873**	**88**
Total no Mundo	**54,600**	**100**

Fonte: Adaptada de (GWEC, 2017)

Segundo o boletim do Ministério de Minas e Energia (MME), a indústria eólica foi a que mais cresceu no Brasil dentre as companhias de energias renováveis em 2016, com mais de

400 usinas e 5200 turbinas em operação. Dentre os estados, o Rio Grande do Norte com 34,7% apresentou a maior proporção de geração, seguido do Ceará com 18,8%. No fator de capacidade instalada, o Piauí teve o maior indicador com cerca de 48,4% (MME, 2017).

Para compreensão do funcionamento dos sistemas eólicos, introduz-se alguns conceitos chaves de seus elementos principais. Os aerogeradores (ou turbinas eólicas) são os instrumentos responsáveis pela geração de energia que funcionam de modo semelhante aos moinhos de vento, uma ferramenta bastante utilizada por homens do campo para bombear água e macerar minerais. Em síntese, a energia cinética do ar em movimento se torna energia mecânica pela força de rotação do rotor. Por conseguinte, o gerador elétrico que está ligado ao rotor, seja diretamente ou por intermédio de uma caixa de engrenagem, transforma a energia mecânica em eletricidade (UCZAI, 2012).

As turbinas modernas de grande porte são instaladas no topo de uma torre com certa altura sobre a superfície, compostas por um rotor horizontal, uma hélice com três pás e um anemômetro (ou sensor de vento) para medir a intensidade da velocidade dos ventos (normalmente a cada 10 minutos), além de outros itens apresentados na Figura 3.

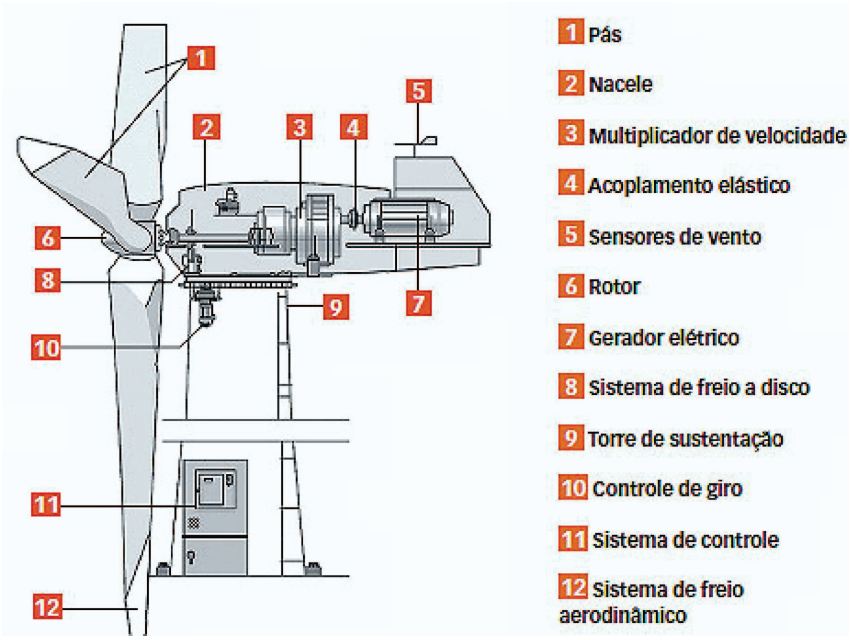

Figura 3 – Conjuntos de itens que compõe um modelo de aerogerador moderno

Fonte: Adaptada de (STAVISS, 2011)

Para que não haja nenhum dano às suas estruturas e alcançar uma produção elétrica economicamente viável, os aerogeradores possuem restrições técnicas que definem um limite superior (ou *cut-out*) e inferior (ou *cut-int*) para seu funcionamento. De acordo com Breeze (2016), normalmente as turbinas apenas começam a produzir eletricidade com ventos próximos a 3 metros por segundo (m/s); da mesma forma, fortes rajadas em torno de 25 m/s cessam a produção. Todavia, os limites de operação podem variam de acordo com as características de cada aerogerador e fabricante.

A maioria dos parques eólicos no mundo são instalados em terra (ou *onshore*), porém na Europa algumas usinas têm sido implantadas no mar (ou *offshore*). Apesar de apresentarem maiores custos de instalação e manutenção, é uma alternativa interessante devido à falta de locais apropriados em terra e ao bom aproveitamento elétrico proporcionado pela maior força dos ventos nos oceanos (TOLMASQUIM, 2016).

Um sistema elétrico robusto demanda informações relevantes para oferta segura de eletricidade. O funcionamento estável dos sistemas eólicos pode ser alcançado pelo balanço entre a estimativa de geração elétrica e a previsão de consumo elétrico, como em (MAÇAIRA; SOUZA; OLIVEIRA, 2016).

As previsões de energia são indispensáveis para o planejamento eficiente de operações nas indústrias, pois permite tomar melhores decisões sobre a manutenção de sistemas, a configuração de turbinas, o gerenciamento da energia e do comércio (COLAK; SAGIROGLU; YESILBUDAK, 2012). De acordo com o relatório do Laboratório Nacional de Energia Renovável dos Estados Unidos (do inglês *National Renewable Energy Laboratory* - NREL), operadores de energia podem ser penalizados, em alguns países, se uma produção for menor do que o valor estimado por eles. Inclusive, as empresas podem não receber qualquer pagamento pela energia gerada acima de uma estimativa (NREL, 2010).

De fato, o grande problema das indústrias eólicas é a inconsistência nas previsões de energia, o que tem dificultado a regulamentação desses sistemas. Na literatura, há duas abordagens para previsão de energia: a previsão direta da saída elétrica de uma turbina, como em (CATALÃO; POUSINHO; MENDES, 2009; CATALÃO; POUSINHO; MENDES, 2011), e a previsão indireta na qual uma estimativa de velocidade do vento é feita, em seguida, ela é convertida em eletricidade.

De acordo com Zhu e Genton (2012), a previsão indireta é mais relevante por duas razões: (i) parques eólicos vizinhos com diferentes modelos de turbinas podem compartilhar a mesma velocidade do vento, ou seja, em vez de realizar previsões de energia de forma separada em cada turbina, uma única curva da previsão de geração é determinada; (ii) a previsão da velocidade do vento, em geral, é mais precisa do que a previsão direta de energia eólica devido à maior correlação espacial do vento.

As previsões indiretas de energia podem ser obtidas através da equação 2.1, na qual a saída de energia P é dada em watts (W), C é o fator dependente do modelo da turbina (ou coeficiente de potência), S é a área rotor em metros quadrados (m^2), A é a densidade do ar em quilograma por metro cúbico (Kg/m^3), e VV a velocidade do vento em m/s (BURTON et al., 2001; EMEIS, 2013). Portanto, prever o comportamento do vento é primordial para as indústrias.

$$P = \frac{1}{2} \cdot C \cdot S \cdot A \cdot VV^3 \tag{2.1}$$

O vento é um recurso natural, intermitente, incerto e de difícil controle que afeta

diretamente a saída das turbinas eólicas. Sua origem é determinada através do gradiente de pressão entre regiões, ocasionado pelo aquecimento desigual das superfícies terrestre pelo sol. Por isso, a energia eólica é considerada um recurso secundário da energia solar. Além das diferenças de pressão, o vento é influenciado por mecanismos complexos, tais como a rotação da Terra, os efeitos físicos de montanhas, os eventuais obstáculos e a rugosidade dos terrenos (TOLMASQUIM, 2016). Consequentemente, prever a velocidade do vento é uma tarefa muito difícil.

Em geral, a influência de obstáculos e da rugosidade diminui em função da altura acima do solo, sendo observadas velocidades maiores proporcionalmente à altura. Por tal motivo, aerogeradores são instalados em lugares abertos e nas maiores alturas possíveis.

Além de ser importante para os sistemas eólicos, compreender a predominância do vento é útil para o planejamento urbano, pois ajuda a decidir onde serão construídos centros industriais, aeroportos, fábricas e lixões (AHRENS; HENSON, 2016).

No entanto, as previsões da velocidade do vento ainda são ineficientes em vários países e as melhorias nas metodologias e abordagens atuais são necessárias para atingir melhores resultados. Atualmente, estimar velocidade do vento pode ser considerada uma das questões de pesquisa mais relevantes e desafiadoras no mundo. Ainda mais, dados meteorológicos utilizados nas previsões são captados a todo momento e têm acumulado enormes volumes de informação nos bancos de dados espaciais, o que demanda teorias e ferramentas apropriadas para uma análise eficiente a partir desses conjuntos. Previsões consistentes evitam prejuízos econômicos e aumentam a eficiência operacional das indústrias através de uma tomada de decisão mais confiável.

2.3 MINERAÇÃO DE DADOS

O armazenamento constante de dados digitais tem provocado um crescimento desenfreado nos bancos de dados de instituições, indústrias e corporações. O aumento na quantidade e variedade de dados está relacionado a diversos fatores, como versatilidade da internet, redução no custo de dispositivos para armazenamento, evolução nas ferramentas de coleta de dados, popularidade de sistemas embarcados, crescimento do trabalho online, dentre outros.

Segundo a *International Data Corporation* (IDC), o universo digital duplica a cada dois anos. Eram 4,4 trilhões de gigabytes de dados no planeta em 2013 que deverá crescer para 44 trilhões de gigabytes até 2020 (IDC, 2014). A variedade e o volume de dados são tão imensos que provocam um ocultamento de informações nos bancos de dados.

Embora existam profissionais especializados que são treinados para analisar manualmente cada modelo de dado (como imagem, texto e planilhas), a detecção de padrão a partir de uma coleção de dados tão ampla e diversificada supera a capacidade racional humana. Nesse contexto, recursos tecnológicos são exigidos para facilitar a extração de informações relevantes pelo homem.

Na literatura, diversos pesquisadores trabalharam com a ideia de que conhecimento pode ser automaticamente detectado, validado e usado de forma inteligente para inúmeras finalidades (WITTEN et al., 2017). Logo, tais necessidades originaram a Descoberta de Conhecimento em Bancos de Dados - *Knowledge Discovery in Databases* (KDD), que é um processo não trivial para identificar padrões em dados que sejam novos, válidos, potencialmente úteis e compreensíveis. Mineração de dados é uma das etapas desse processo onde algoritmos específicos são aplicados para detectar padrões relevantes em um banco de dados sistemático (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). No entanto, "processo de mineração de dados" é um termo popularizado que vem sendo utilizado por analistas de dados e estatísticos como um sinônimo de KDD.

O processo de mineração de dados (ou KDD) envolve uma sequência de etapas interativas e iterativas. Nessas etapas, o conhecimento de pelo menos um especialista é fundamental para analisar, interpretar, compreender e validar os dados do processo. Apresenta-se na Figura 4 a sequência de etapas do processo de KDD que são: seleção dos dados, pré-processamento, transformação nos dados, mineração de dados e avaliação (ou interpretação). Vale salientar que, dependendo do domínio da aplicação, as etapas de pré-processamento poderão anteceder a de seleção dos dados e de transformação nos dados, como em (HAN; KAMBER; PEI, 2012). Cada etapa será descrita a seguir.

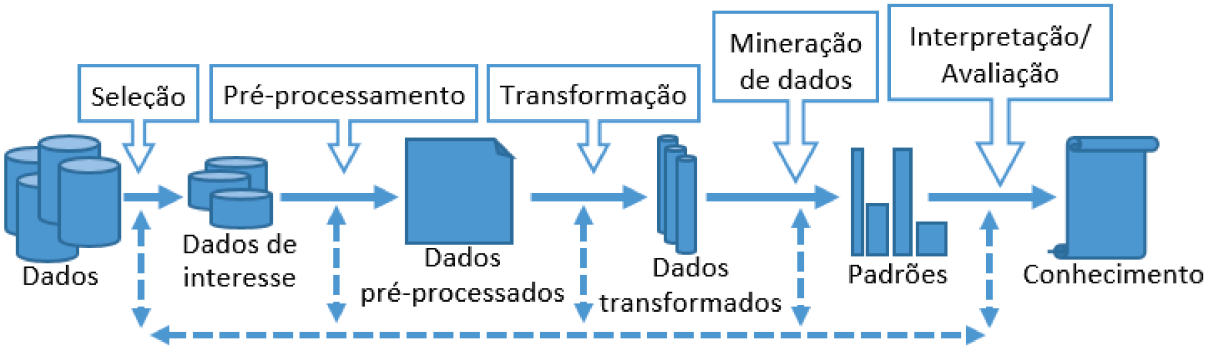

Figura 4 – Processo de mineração de dados

Fonte: Adaptada de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

Neste processo, tudo se inicia a partir de um conjunto de dados, que são elementos puros e quantificáveis (como fatos, números, imagens ou textos). Esses artefatos individualmente não oferecem qualquer embasamento para o entendimento da situação. Na etapa de seleção, define-se uma parte do conjunto de dados considerado relevante de acordo com o domínio do problema em

questão. Os dados, muitas vezes, são de diversas fontes e estão totalmente dispersos em vários repositórios. Portanto, o especialista define um subconjunto de dados para tratamento e análise.

Um repositório de dados, em geral, contém dados ruidosos, ausentes e inconsistentes por diversos motivos, como problemas na captação do sensor, erros humanos de digitação, deterioração, formato inadequado e falta de integração de dados. A qualidade dos dados pode ser avaliada pelo nível de exatidão, integridade, consistência, pontualidade, credibilidade e interoperabilidade (HAN; KAMBER; PEI, 2012).

O pré-processamento é uma etapa decisiva, no processo de KDD, para tratar a qualidade dos dados, podendo consumir aproximadamente 70% do tempo em relação às demais etapas. Algumas tarefas que podem ser realizadas no pré-processamento dos dados incluem: (i) ignorar o registro que tem o atributo incorreto, (ii) atribuir um valor manual, (iii) usar uma constante global, (iv) definir uma medida de tendência central (por exemplo, a média ou mediana), (v) usar um valor provável baseado na vizinhança (com técnicas de suavização) e outras técnicas encontradas em (GARCÍA; LUENGO; HERRERA, 2016).

Dentro de um banco de dados podemos considerar três tipos básicos de atributos: numéricos, mensuram valores inteiros ou reais; booleanos, que determinam duas possibilidades (verdadeiro ou falso); e nominais (ou discretos), que assumem um conjunto finito e determinado de possibilidades como, por exemplo, um dado de temperatura pode ser alta, média e baixa.

Vistos alguns dos tipos de atributos, podemos dizer que a etapa de transformação nos dados objetiva definir um novo conjunto de dados considerado "mais adequado" para ser utilizado na etapa de mineração de dados. Em outras palavras, os dados podem ser representados de uma nova forma para maximizar o desempenho na etapa seguinte; por exemplo, alguns métodos de mineração de dados apenas funcionam com dados nominais (como certos métodos de classificação). Nesse caso, se os dados forem numéricos, esses precisam ser transformados em dados nominais para o funcionamento correto dos algoritmos. Algumas das tarefas que podem ser realizadas na etapa de transformação incluem: discretização de dados, realizado por meio de técnicas de suavização; formatação dos dados, que consiste em representar os dados em uma outra forma, sem prejudicar a sua integridade; compressão de dados, que permite representar os dados de uma forma reduzida; normalização, que objetiva dimensionar a escala dos dados.

Com os dados transformados, podemos definir os algoritmos e ferramentas automáticas (ou semiautomáticas) e inteligentes para inspeção do banco de dados. Os métodos de mineração de dados têm dois objetivos principais: a descrição e a previsão (KANTARDZI, 2011). Os métodos de descrição buscam, em um conjunto de dados, padrões novos e informações não triviais disponíveis para revelar os relacionamentos entre dados, tais como os método para agrupamento (ou *clustering*), sumarização, modelagem de dependência, e detecção de desvio e mudanças. Por outro lado, os métodos de previsão produzem um modelo representativo (ex. um código executável), a partir de um conjunto de dados de treinamento, para previsão de novos dados, tais como classificação e regressão. Cada método tem seu objetivo e particularidade,

portanto cabe ao especialista detectar, analisar e definir a classe ou método de mineração de dados mais adequada para o objetivo projetado.

Os métodos de agrupamento (ou segmentação) visam separar um conjunto de dados N em K subconjuntos (n1, n2, n3...), conhecidos como *clusters*. Semelhanças e características em comum entre os dados são critérios usados para separação dos conjuntos. Um algoritmo de agrupamento bem conhecido é o K-Means, no entanto, muitos outros são discutidos em (ZAKI; MEIRA-JR, 2014).

Os métodos de sumarização propõem extrair de dados relacionamentos entre atributos. As regras de associação são métodos de sumarização muito aplicados que avaliam os atributos que ocorrem frequentemente. Uma aplicação comum no comércio trata da verificação de quais subconjuntos de produtos são comprados com mais frequência. Tal método pode detectar um padrão de relacionamento como, por exemplo, os clientes que compraram cervejas no final de semana também obtiveram fraldas de bebê (LAROSE; LAROSE, 2014).

Os métodos de modelagem de dependência, muitas vezes, são utilizados para derivar alguma estrutura causal entre os dados. Esses modelos podem ser probabilísticos ou determinísticos. Os métodos de estimativas de densidade e de causais explícitas se enquadram nessa categoria (FAYYAD, 1997).

Os métodos de detecção de desvios e mudanças identificam ocorrências tanto em sequências de informações como em séries temporais. As duas principais característica desses métodos são a sua capacidade de explicar a ordenação das observações e a busca por padrões com pouca incidência (FAYYAD, 1997). Uma aplicação relevante dessa categoria é apresentada em (GOLDSCHMIDT; PASSOS, 2005), na qual anomalias no consumo de energia elétrica de uma residência foram detectadas nos últimos 10 anos.

Além dos métodos de descrição, há os métodos de previsão. Os métodos de classificação são utilizados para previsão em dados categóricos (ou nominais), enquanto que os de regressão trabalham com a previsão de dados numéricos. Atualmente, muitos métodos para classificação são modificados para regressão e vice-versa. Podemos citar algoritmos de árvore de decisão, redes neurais e classificadores bayesianos que estão contidos em ambas categorias.

Assim que definimos a classe de método e selecionamos um algoritmo, o processo de mineração termina com a descoberta de padrões que precisam ser avaliados e interpretados por um especialista com algum grau de certeza para auxiliar as possíveis tomadas de decisão. O que torna um padrão interessante pode variar entre especialistas. Todavia, um padrão é relevante se compreendido facilmente pelos seres humanos, validado por meio de dados novos com algum grau de certeza, potencialmente útil (fornecer alguma vantagem ou utilidade), novo (previamente desconhecido) ou se validou uma hipótese que o usuário elaborou (HAN; KAMBER; PEI, 2012).

Percebe-se claramente que o processo de mineração é um domínio altamente orientado à aplicação. Com o passar dos anos, o campo incorporou muitas técnicas de outros domínios,

incluindo estatística, aprendizado de máquina, reconhecimento de padrões, sistemas de bancos de dados, *data warehouse*, recuperação de informação, visualização, algoritmos e computação de alto desempenho, conforme apresentados na Figura 5.

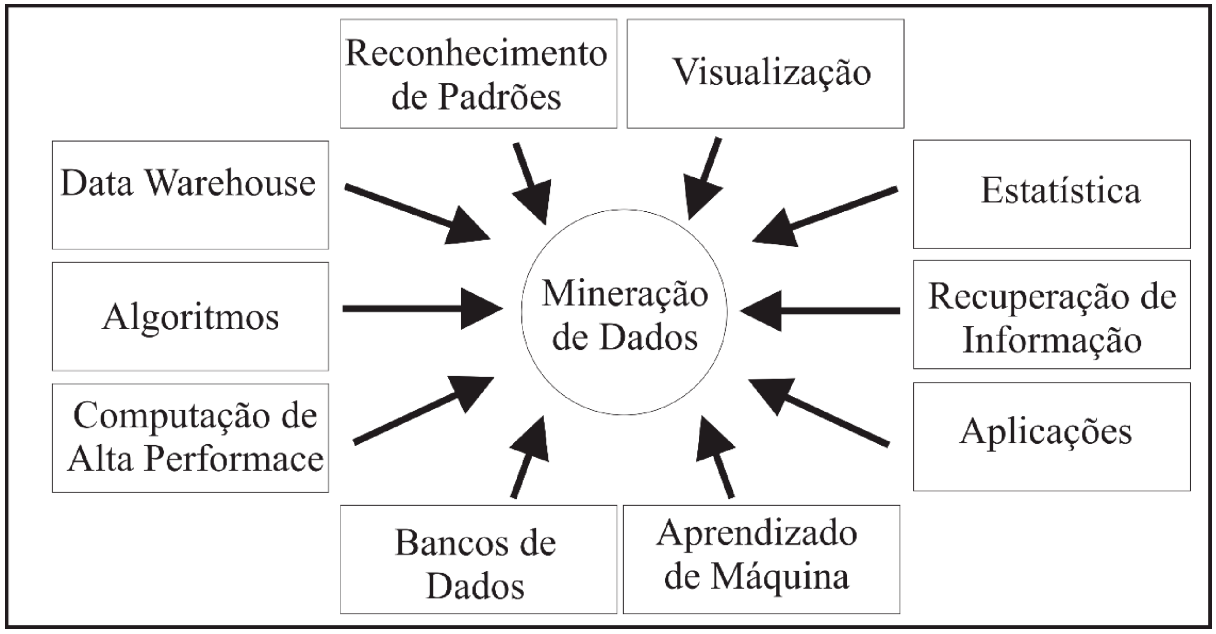

Figura 5 – Campos que envolvem a mineração de dados

Fonte: Adaptada de (HAN; KAMBER; PEI, 2012)

A comunidade de mineração de dados tem alcançado muitas soluções viáveis para diversas áreas e aplicações. Muitas aplicações sobre *business intelligence*, motor de pesquisa web e bioinformática são discutidas em (HAN; KAMBER; PEI, 2012). Aplicações envolvendo marketing e vendas, julgamento de decisão, imagens de satélites, mineração de conteúdo web, envolvendo previsões e diagnóstico médico são descritas em (WITTEN et al., 2017). Aplicações de negócios e nas ciências são mencionadas por (LUO, 2008).

2.4 REVISÃO DA LITERATURA: PREVISÃO DA VELOCIDADE DO VENTO

Até o momento, foi introduzido temas relacionados ao contexto desta proposta, como Meteorologia, Energia Eólica e Mineração de Dados. Nesta seção, é discutida uma visão geral do problema da previsão da velocidade do vento, assim como trabalhos relacionados.

Na literatura, o problema da previsão da velocidade do vento é dividido em quatro categorias de escala temporais; ainda que a divisão não seja tão clara e exata, podemos considerá-la da seguinte forma: (i) previsão de prazo longo (do inglês *long-term*) para uma semana até um ano ou mais à frente; (ii) previsão de prazo médio (do inglês *medium-term*) para quarenta e oito horas até uma semana; (iii) previsão de prazo curto (do inglês *short-term*) para um dia até

quarenta e oito horas; e (iv) previsão de prazo ultra curto (do inglês *very short-term*) para poucos segundos até um dia (FREITAS; SILVA; SAKAMOTO, 2018).

Cada categoria de previsão tem sua importância para os operadores de energia: estimativas de prazo longo são preparadas para dar suporte às decisões sobre o mercado elétrico e para otimizar os custos no planejamento de manutenções prolongadas, enquanto que as de prazo médio são usadas para tomada de decisão sobre ligamento ou desligamento de aerogeradores. Por outro lado, as estimativas de prazo curto são relevantes para planejamento do despache econômico de energia, ou seja, decisões relacionadas ao incremento (ou decremento) de carga (SOMAN et al., 2010). Por fim, estimativas de prazo ultra curto servem para configurar as turbinas e esclarecer informações do mercado elétrico (FAZELPOUR; TARASHKAR; ROSEN, 2016).

Nos últimos anos, muitas abordagens foram implementadas para previsão da velocidade do vento. Segundo Lei et al. (2009), quatro categorias de métodos podem ser consideradas: (1) físicos, (2) convencionais estatísticos, (3) de correlação espacial e (4) de inteligência artificial e novos modelos.

Métodos físicos fazem previsões considerando somente propriedades físicas, tais como características do terreno, obstáculos, temperatura e pressão do ar (LAZIĆ; PEJANOVIĆ; ŽIVKOVIĆ, 2010; EL-FOULY; EL-SAADANY; SALAMA, 2008). Essas abordagens normalmente demandam muitos recursos computacionais (ou supercomputadores) e não fornecem bons resultados nas previsões de prazo curto e ultra curto.

Métodos convencionais estatísticos são modelos matemáticos que estimam velocidade do vento a partir de uma análise estatística de séries temporais, tais como *autoregressive model* (MOHANDES; REHMAN; HALAWANI, 1998), *moving average model*, *autoregressive moving average model*, *autoregressive integrated moving average model* e *Kalman filter* (ERDEM; SHI, 2011). Em geral, essas abordagens fornecem bons resultados nas previsões de prazo curto.

Métodos de correlação espacial predizem a velocidade do vento explorando relacionamentos espaciais de estações eólicas vizinhas. Uma abordagem que combinou uma rede neural com correlação espacial apresentou bons resultados nas previsão de prazo curto (BARBOUNIS; THEOCHARIS, 2007; FINAMORE et al., 2016). No entanto, é muito difícil encontrar dados disponíveis em usinas vizinhas devido à forte competição no mercado de eletricidade.

Recentemente, vários trabalhos focaram nos métodos baseados na Inteligência Artificial (IA) e outros modelos, pois descrevem um relacionamento estatístico não linear e altamente complexo entre dados meteorológicos e velocidade do vento, tais como redes neurais (do inglês *Neural Networks* - NN)(VELO; LÓPEZ; MASEDA, 2014; MALIK; SAVITA, 2016; KAUR; KUMAR; SEGAL, 2016; SHAO; CUI; DENG, 2016), lógica fuzzy (DAMOUSIS et al., 2004; MOHANDES; REHMAN; RAHMAN, 2011), máquina de vetores de suporte (do inglês

Support Vector Machine - SVM) (HU et al., 2016; PINTO et al., 2015; LIU; KONG; LEE, 2014; LAHOUAR; SLAMA, 2014) e alguns métodos híbridos (CADENAS; RIVERA, 2010; JIANG; WANG; WANG, 2017; CHANG et al., 2017; GUO et al., 2011; WANG et al., 2015; WANG et al., 2014; LIU et al., 2014; BOUZGOU; BENOUDJIT, 2011). Essa categoria vem sendo a mais utilizada em razão dos resultados interessantes em todos os modelos de previsão.

2.4.1 TRABALHOS RELACIONADOS

Dentre os métodos da abordagens de IA e outros modelos, os modelos híbridos têm um ciclo de construção complexo, uma vez que combinam, de forma sequencial ou paralela, dois ou mais algoritmos para descrever o comportamento futuro da velocidade do vento. Essa abordagem traz bons resultados para uma região específica, considerando que os padrões de ventos numa região sejam bem compreendidos e implementados dentro do método. No entanto, como tais abordagens são construídas e adaptadas a um único conjunto de dados, há uma perca de acurácia quando são aplicadas em outros conjuntos devido às diferentes projeções dos sinais. Por tal motivo, desprezamos as abordagens híbridas neste trabalho.

Além dos modelos híbridos, as abordagens de IA mais utilizadas são os métodos de NN e SVM, embora há trabalhos que aplicaram os algoritmos K-Nearest Neighbors (KNN) e Regressão Linear (RL), conforme apresentados na Tabela 3. Em geral, esses modelos são construídos através de uma análise de variáveis meteorológicas, como temperatura do ar (TP), umidade relativa (UR), pressão atmosférica (PA), velocidade do vento (VV) e direção do vento (DV).

De acordo com Pinto et al. (2015), quinze diferentes implementações foram realizadas com SVM e NN para previsões da velocidade do vento a cada cinco minutos nos Estados Unidos. Como treinamento dos modelos foram considerados três anos de dados com os seguintes atributos: temperatura, velocidade e direção do vento. SVM teve a maior acurácia nos experimentos superando o melhor resultado da NN com valores de 0,7120 para MAE, 22,87% para MAPE e 21,17% para SMAPE (ver modelo n° 1 e 2 na Tabela 3).

Em (YESILBUDAK; SAGIROGLU; COLAK, 2013), vários modelos para previsão de ventos a cada 10 minutos foram implementados usando KNN a partir de um conjunto de dados que cobriu o mês de julho de 2010 em Poyracık, Turquia. Dentre os quatro modelos mais preciso (n° 3 até 6), o modelo n° 3 combinou temperatura, pressão do ar, umidade relativa e direção do vento, alcançando os melhores resultados tais como 0,7400 para MAE e 7,08% para MAPE.

Segundo Lahouar e Slama (2014), um algoritmo de SVM foi implementado para previsão horária, a partir de dados de velocidade e direção do vento (representado na forma nominal em 16 classes). O modelo mais relevante alcançou 0,8363 para MAE e 1,1800 para RMSE, quando

Tabela 3 – Abordagens baseadas em IA para previsão da velocidade do vento

n°	Intervalo de previsão	Métodos de IA	Atributos de Entrada	Região	MAE	RMSE	MSE	MAPE	SMAPE
1	5 minutos	SVM	TP, VV, DV	Estados Unidos	0,7120	-	-	22,87	21,17
2	5 minutos	NN	TP, VV, DV	Estados Unidos	0,8180	-	-	-	-
3	10 minutos	KNN	TP, PA, UR, DV	Turquia	0,7400	-	-	7,08	-
4	10 minutos	KNN	PA, UR, DV	Turquia	0,8210	-	-	7,71	-
5	10 minutos	KNN	TP, PA, UR	Turquia	0,8000	-	-	7,48	-
6	10 minutos	KNN	TP, PA, DV	Turquia	1,0130	-	-	9,52	-
7	horário	SVM	VV, DV	Tunísia	0,8363	1,1800	-	-	-
8	horário	NN	VV	Jilin, China	0,9305	1,2382	-	16,72	-
9	horário	NNR	VV	Jilin, China	0,9319	1,2435	-	14,95	-
10	horário	RL	VV	Jilin, China	0,9267	1,2359	-	15,93	-
11	horário	NN	VV	Gansu, China	0,9725	1,2685	-	26,49	-
12	horário	NNR	VV	Gansu, China	1.0037	1,2905	-	26,81	-
13	horário	RL	VV	Gansu, China	0.9735	1,2662	-	26,26	-
14	5 horas	NN	TP, PA, VV	Estados Unidos	0,8460	-	-	-	-
15	5 horas	NN	VV, DV	Estados Unidos	0,8430	-	-	-	-
16	5 horas	NN	VV	Estados Unidos	0,8180	-	-	-	-
17	10 horas	NN	VV	Estados Unidos	2,0660	-	-	-	-
18	diário	NN	VV	Estados Unidos	1,9770	-	-	-	-
19	diário	NN	TP, PA, VV	Itália	-	-	3,1500	-	-
20	diário	NN	TP, PA, VV	Itália	-	-	3,4500	-	-
21	diário	NN	TP, VV, DV	Estados Unidos	0,9789	1,2984	-	-	-
22	diário	NN	VV, DV	Espanha	1,5864	2,2126	-	-	-

Fonte: Autoria Própria

foi treinado com dados de 2009 e 2010, além de validado com dados de janeiro, fevereiro e março de 2011 (ver modelo n° 7). O manuscrito descreve que o SVM traz boa precisão quando comparado às redes neurais para previsões de prazo curto.

Para previsão horária nas regiões de Jilin e Gansu na China, uma rede neural *MultiLayer Perceptron* (MLP), uma rede neural recorrente (NNR) e um algoritmo de regressão linear foram implementados em (HU et al., 2016) (ver modelo n° 8 até 13). Os três modelos para cada região teve resultados relevantes numa validação usando dados horários cobrindo 4 meses. Na região de Jilin, a MLP teve uma maior acurácia nos resultados com 0,9725 para MAE, 1,2685 para RMSE e 21.88% para MAPE. Na região de Gansu, em contraposição, regressão linear foi um pouco superior aos dois modelos de redes neurais com 0,9267 para MAE, 1,2359 para RMSE e 15,93% para MAPE.

Em (RAMOS et al., 2011b), uma metodologia foi proposta para previsão da velocidade do vento. Cinco modelos foram construídos usando uma rede neural para diferentes previsões e combinações de atributos, conforme apresentados nos modelos n° 14 até 18. A previsão para cinco horas, usando apenas velocidade do vento como entrada, teve a maior precisão atingindo valor de 0,8180 para MAE.

De acordo com Finamore et al. (2015), uma rede neural foi construída para previsão diária de ventos em Campânia, Itália. Quatro anos de dados foram usados para treinamento do modelo, enquanto que dois meses foram destinados à validação. Atributos como pressão, temperatura e velocidade do vento foram utilizadas nas duas simulações em 2014: uma para

março como mostrada no modelo n° 19 e outra para junho, modelo n° 20. Na primeira simulação, o resultado foi melhor com cerca de 3,1500 para MSE.

Em (DARAEEPOUR; ECHEVERRI, 2014) foi elaborada uma rede neural para previsão diária em Kansas nos Estados unidos e em Galícia na Espanha. O modelo americano combinou temperatura, velocidade e direção do vento e teve melhores resultados com 0,9789 para MAE e 1,2984 para RMSE (ver modelo n° 21). O modelo espanhol (n° 22) teve como entrada apenas velocidade e direção do vento e atingiu 1,5864 para MAE e 2,2126 para RMSE.

De acordo com Zhao, Wang e Li (2011), novas metodologias ainda necessárias para melhorar os resultados de precisão, reduzir a incerteza e manter um tempo de computação aceitável nas previsões de energia eólica. No geral, a literatura tem buscado um único método que, se aplicado em qualquer situação, forneça os melhores resultados de precisão possíveis. No entanto, como os padrões de ventos são influenciados por muitos fatores e variam entre regiões, desenvolver um método global para previsão da velocidade do vento é muito difícil ou impossível. Além do mais, a comparação de resultados de modelos de previsão é complexa, pois depende de muitos fatores, como tempo de execução, configuração exata do método, volume de dados, parâmetros de entrada, precisão nos resultados, critério de validação considerado e características dos sinais.

A maioria dos trabalhos mencionados fizeram estudos usando alguns dos atributos meteorológicos para determinar o melhor método de previsão para uma dada situação sem considerar o tempo de execução e diversos aspectos relevantes. Em outras palavras, uma abordagem foi implementada para um conjunto de dados específico sem considerar aspectos que aumentam a qualidade do banco de dados e somente resultados de precisão foram comparados para determinar o método mais confiável para aquela situação.

Acredita-se que os resultados em todos os trabalhos mencionados poderiam ser melhores, em termos de performance e precisão, se uma abordagem eficiente, inteligente e estratégica fosse executada. Em razão disso, foi proposta uma abordagem de mineração de dados para previsão da velocidade do vento na qual diversos aspectos importantes são levados em consideração para construção de um modelo consistente, tais como uma análise eficiente para tratamento de enormes banco de dados, vários algoritmos semiautomáticos baseados na IA, mecanismos para validação dos resultados e critérios interessantes que se executados corretamente, podem fornecer melhorias satisfatórias nos resultados.

3 ABORDAGEM DE MINERAÇÃO DE DADOS

Considerando uma maior flexibilidade nas indústrias de eletricidade e visando facilitar o manuseio com os dados meteorológicos, propomos uma abordagem que tem início a partir de um banco de dados, como ilustrado na Figura 6. Nesta abordagem, os dados atravessam um tratamento estratégico e inteligente através de uma série de atividades nas etapas de processamento e transformação que visa aumentar a qualidade do banco de dados. Em seguida, um modelo de previsão é construído através da execução de um algoritmo de IA. O especialista avalia o modelo desenvolvido, se os resultados mostrarem-se relevantes, ele pode implementar o modelo e usá-lo em previsões futuras. Caso contrário, retorna-se a etapas anteriores com o objetivo de aumentar a qualidade dos resultados. O conjunto de etapas incorporadas nesta proposta de mineração de dados são descritas adiante.

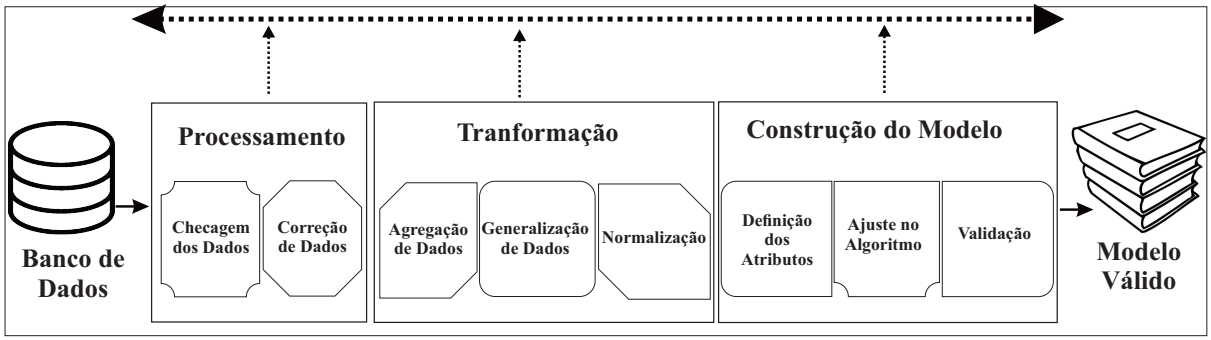

Figura 6 – Abordagem de mineração de dados para previsão da velocidade do vento

Fonte: Autoria Própria

Dados meteorológicos que são usados nas previsões de ventos têm diferentes comportamentos ao redor de toda a atmosfera. Desse modo, precisamos entender os valores e variações de cada atributo para uma região em particular e, então, podemos seguir até a etapa de processamento. Os atributos meteorológicos utilizado neste trabalho são descritos abaixo, seguindo as definições de Ahrens e Henson (2016).

- temperatura do ar – o grau de calor do ar dado em graus Celsius (°C);
- pressão do ar – a força do ar exercida sobre uma superfície, normalmente dada em milibar (mb);
- umidade relativa do ar – é a razão da média do vapor de água no ar pela média de vapor de água necessária para saturação, mensurada em porcentagem (%);
- velocidade do vento – é a quantificação do movimento do ar analisado por um observador estacionário numa fatia de tempo. Este movimento pode ser expresso como o número de metros percorridos em um segundo;

- direção do vento – consiste na direção da origem do vento a qual pode variar de 0 até 360 graus (°).

A temperatura do ar de uma dada região normalmente diminui em função da altura acima do solo devido ao resfriamento adiabático. Em geral, a temperatura tem uma variação um pouco lenta em relação à altitude. Normalmente, a cada 1000 metros de altura, há em média uma baixa de 6,5 °C de temperatura (AHRENS; HENSON, 2016). Por outro lado, a pressão atmosférica tem um perfil um pouco mais rápido: próximo ao nível do mar, a pressão pode variar em torno de 1000 mb, mas diminui 10 mb para cada 100 metros de altitude. Em síntese, quanto mais elevado um objeto está em relação ao solo, menores são os graus de pressão e de temperatura exercidos nele.

A umidade relativa pode mudar quando há alterações no conteúdo de vapor de água no ar ou na temperatura do ar (AHRENS; HENSON, 2016). Por outro lado, os ventos podem variam de 0 até 25 m/s, embora existam movimentações de ar acima de 25 m/s que não são consideradas ventos normais, uma vez que são furacões, tornados e tempestades violentas, de acordo com a escala de Francis Beaufort [1].

3.1 ATIVIDADES DE PROCESSAMENTO E TRANSFORMAÇÃO

Dada uma visão geral dos atributos meteorológicos, na etapa de processamento, podemos aumentar a qualidade do banco de dados, através da checagem e da correção dos dados. A checagem dos dados é uma atividade mutável que tem como objetivo detectar os atributos com valores nulos, raros e impossíveis. Recursos visuais como gráficos, implementações de código em uma linguagem de programação e consultas em SQL a partir de um banco de dados, dentre outras ferramentas, dão aos especialistas oportunidades de detectar, de forma eficiente e inteligente, anomalias em repositórios de dados de grandes volumes.

Na correção dos dados, o especialista toma suas decisões sobre os atributos, de acordo com a situação detectada na atividade anterior. Ele pode decidir ignorar o atributo suspeito, corrigir manualmente atributos, formatar casas decimais, definir uma constante global (media ou mediana), coletar informação necessária para modelar ou estimar ruído (e.g., regressão ou inferência), dentre outros (GARCÍA; LUENGO; HERRERA, 2016).

Com os dados processados, na etapa de transformação viabilizamos o aumento da performance de cada algoritmo. Uma agregação é realizada com o objetivo de reduzir o conjunto de dados e aumentar a performance na construção de modelos de previsão. A redução é definida de acordo com a necessidade do especialista: por exemplo, se for considerado a elaboração

[1] http://www.tempoagora.com.br/dia-a-dia/como-e-medida-velocidade-vento/

de um modelo de previsão diária de ventos a partir de dados horários, pode-se transformar os dados horários em diários, o que certamente reduziria significativamente o conjunto de dados e aumentaria a performance na execução dos algoritmos, uma vez que vinte quatro registros podem se unir para forma um único.

Algoritmos de IA, em geral, constroem um modelo preditivo de forma diferente e podem ter uma maior performance e precisão nos resultados, se executados em um conjunto de entrada adequado: por exemplo, um método pode construir um modelo para previsão mais acurado (ou mais rápido), se um atributo mês for representado na forma nominal "Janeiro", em vez da forma numérica *1*. Diante do exposto, o especialista pode definir diferentes formas para representar alguns dos atributos processados sem prejudicar a integridade deles e usá-los como entrada para os algoritmos na tentativa de conseguir melhores resultados.

Antes de iniciar a construção dos modelos, recomenda-se normalizar todos os atributos em uma escala de dados para facilitar e acelerar a execução dos algoritmos. A normalização lida com a mudança e padronização da dimensão de escalas dos dados, ou seja, atributos podem ser normalizados com casas decimais determinadas após a vírgula ou estratégias que definem uma única escala para todos os atributos (e.g., de 0 a 1). Em ambos os casos, a normalização pode aumentar a performance dos modelos, uma vez que permite ao algoritmo executar de forma mais simples desprezando cálculos extensos.

3.2 ATIVIDADES EXECUTADAS PARA CONSTRUÇÃO DOS MODELOS DE PREVISÃO

A partir dos dados transformados, podemos executar os algoritmos automáticos (ou semiautomáticos) e pertinentes para previsão da velocidade do vento. Os padrões de ventos variam por diversos fatores tornando difícil definir um único método para uma previsão. Por isso, vários algoritmos devem ser considerados no desenvolvimento de modelos de previsão. Desse modo, propomos três atividades fundamentais para alcançar resultados consistentes: análise do atributos de entrada, ajuste do algoritmo e validação.

Na análise de atributos propomos detectar um relevante conjunto de entrada para os algoritmos. O excesso de atributos irrelevantes na entrada do modelo prejudica os seus resultados de precisão e performance. Sendo assim, pode-se detectar nessa análise que um determinado algoritmo poderia executar em poucos segundos e alcançar melhores resultados de precisão se fosse combinado em sua entrada somente quatro atributos ao invés de dez. De fato, precisamos analisar e definir um relevante conjunto de entrada para cada algoritmo, de modo a alcançar um equilíbrio entre performance e acurácia nos resultados.

Para definir o melhor conjunto de dados usamos um coeficiente de correlação (R),

construído baseado no *Pearson correlation coefficient* que foi capaz de descobrir padrões escondidos de vendas no mercado em (CHEUNG; LI, 2012). *R* expressa um valor numérico entre -1 e 1 que significa o fator de correção linear existente entre duas variáveis, conforme a equação 3.1: valores próximo à -1 indicam uma alta correlação negativa; valores próximos à 1 indicam uma alta correlação positiva; por fim, valores próximos à 0 indicam uma baixa ou nenhuma correlação entre os dados (WITTEN et al., 2017). Segundo Taylor (1990), valores positivos ou negativos entre 0,36 e 0,67 revelam uma correlação moderada, entre -0,36 e 0,36 uma correlação insignificante, e entre 0,67 e 1 podem ser considerados de alta correlação.

$$R = \frac{\sum_{i=1}^{N}(x1 - x2)(y1 - y2)}{\sqrt{\sum_{i=1}^{N}(x1 - x2)^2(y1 - y2)^2}} \tag{3.1}$$

Para analisar o melhor conjunto de dados, determinamos uma estratégia exaustiva na qual todas as combinações possíveis de atributos são analisadas para determinar o conjunto mais adequado, ou seja, o que tem maior correlação nos experimentos. Por exemplo, se tivermos dois atributos A e B, realizaremos três testes: A, B e AB; se tivermos três atributos A, B e C serão sete testes: A, B, C, AB, AC, BC e ABC; e assim por diante. Existem outras abordagem que pode-se utilizar para definir o melhor conjunto. No entanto determinar o conjunto de forma exaustiva garante alcançar um resultado relevante para cada algoritmo, promovendo uma visão ampla das correlações entre atributos para cada método.

Assim que definimos um conjunto de dados, podemos ajustar os parâmetros de cada algoritmo e analisar os resultados. O ajuste dos algoritmos dependem de uma série de fatores, como volume de dados, projeções dos sinais, quantidade de atributos, dentre outros. Cabe ao especialista realizar novamente uma série de testes com a finalidade de detectar uma estratégia adequada de configuração nos parâmetros para cada um dos métodos.

3.3 MÉTRICAS DE VALIDAÇÃO DE MODELOS ESTATÍSTICOS

Após a execução dos algoritmos, um modelo de previsão é construído e deve ser validado por um especialista. Os resultados das previsão da velocidade do vento podem ser avaliados por muitas métricas estatísticas que determinam uma representação numérica para o erro. Na literatura, houve muita discussão sobre os critérios mais adequados para validar modelos de previsão: (WILLMOTT; MATSUURA, 2005; CHAI; DRAXLER, 2014; WANG; BOVIK, 2009; GOODWIN; LAWTON, 1999; TAYMAN; SWANSON, 1999). Contudo, ainda não existe uma única abordagem global para avaliar modelos de previsão, uma vez que cada métrica fornece uma diferente visão do erro para o especialista. Por isso, os trabalhos da literatura que foram

apresentados na Tabela 3 utilizaram duas ou mais métricas para julgar os modelos de previsão da velocidade do vento.

Nesta abordagem, apresentam-se as cinco métricas mais comuns e relevantes para validar os modelos de previsão e facilitar a comparação de resultados com a literatura, tais como *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE), *Mean Square Error* (MSE), *Mean Absolute Percentage Error* (MAPE) e *Symmetric Mean Absolute Percent Error* (SMAPE). Elas são *negatively-oriented scores*, o que significa que quanto mais baixo forem seus valores, melhores serão seus resultados. Além disso, são métricas complementares nas quais uma pode suprir a deficiência da outra.

Seja *R* o valor real da velocidade do vento, *P* o valor estimado e *N* o total de instâncias para o estado *i*. O MAE determina a média da magnitude dos erros absolutos sem considerar os seus sinais, ou seja, os diferentes erros têm pesos iguais, conforme apresentado na equação (3.2).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Pi - Ri| \tag{3.2}$$

O RMSE também estima a magnitude dos erros, mas tem proporção diferente pois considera a raiz quadrada na média dos valores, conforme apresentado na equação (3.3). Em alguns casos, RMSE pode fornecer um erro enganador (WILLMOTT; MATSUURA, 2005), mas costuma ser mais apropriado do que MAE para representar a performance do modelo quando se espera um erro distribuído como uma função gaussiana (CHAI; DRAXLER, 2014).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Pi - Ri)^2} \tag{3.3}$$

O MSE mensura a média dos erros ao quadrado, ou seja, os erros com valores maiores têm um peso maior do que os erros de menor valor, conforme ilustrado na equação (3.4).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Pi - Ri)^2 \tag{3.4}$$

Além dessas, existem métricas que expressam o erro em função da porcentagem dos dados. O MAPE é considerado uma métrica bem interpretável para os especialistas, pois expressa a média dos erros em função da quantidade de dados, de forma desbalanceada entre erros positivos e negativos. Todavia, MAPE pode fornecer um valor exagerado nos casos em que exitem muitos erros absolutos positivos, pois consideram na divisão somente a média dos valores preditivos, conforme apresentado na Equação 3.5.

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{Pi - Ri}{Pi}\right| \tag{3.5}$$

No entanto, SMAPE foi desenvolvido para lidar com as limitações do MAPE e calcular um erro de forma equilibrada, pois considera uma média dos erros positivos e negativos, conforme apresentado na equação (3.6) (GOODWIN; LAWTON, 1999; TAYMAN; SWANSON, 1999). Para MAE, RMSE e MSE a saída do erro é apresentada em (m/s²), enquanto que para MAPE e SMAPE é dada em porcentagem (%).

$$SMAPE = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{Pi - Ri}{\frac{Pi+Ri}{2}} \right| \tag{3.6}$$

Para comparação de modelos desenvolvidos a partir de uma mesma base de dados o MAE parece levar vantagem em relação ao RMSE e MSE, pois é a métrica mais utilizada na literatura que calcula a média dos erros absolutos de forma balanceada. No entanto, para comparação entre modelos que foram construídos a partir de bases de dados diferentes, MAE, RMSE e MSE torna-se inadequado pois a amplitude maior dos sinais tendem a aumentar a média dessas métricas. SMAPE pode ser mais adequado para comparação de modelos em conjuntos distintos pois fornece um erro interpretável e calculado de forma balanceada em função da base de dados.

A performance de cada modelo pode ser avaliada através do tempo de execução de cada algoritmo dado em segundos (s). Tendo em vista que os modelos são construídos a partir de enormes conjuntos de dados, um modelo pode se tornar inviável do ponto de vista operacional. Por isso, propomos fazer uma análise de tempo de execução que vem sendo ignorada na literatura.

No cenário atual de mineração de dados, ferramentas livres para extração de conhecimento tiveram uma rápida evolução e causaram uma competição intensiva com softwares comerciais. Atualmente, existem diversas ferramentas gratuitas no mercado que incorporam uma ampla variedade de recursos e algoritmos eficientes para extrair conhecimento de gigantescos repositórios. Ao invés de implementar, testar e validar algoritmos, definir um software robusto e popular que atravessou uma série de testes parece ser uma estratégia interessante para utilizar na etapa de construção dos modelos. Para as indústrias eólicas, um software livre pode ser uma ótima opção inicial de baixo custo para previsão de ventos, considerando que tenha bons algoritmos implementados.

3.4 WEKA

WEKA (do inglês *Waikato Environment for Knowledge Analysis*) é uma das ferramentas de mineração de dados mais flexíveis e robustas que traz uma performance sólida na maioria de seus recursos (AL-ODAN; AL-DARAISEH, 2015). Esse software foi desenvolvido em linguagem C, pela Universidade de Waikato, na Nova Zelândia em 1993. Mais tarde, em razão de alguns fatores, o software foi reescrito em linguagem Java seguindo o paradigma orientado a

objetos e distribuído sob licença GNU General Public License (HALL et al., 2009). Isto significa que o código do sistema pode ser baixado na plataforma da universidade [2] e adaptado para as necessidades de cada usuário.

A ferramenta dispõe de uma interface simples, de fácil manuseio e flexível para manusear bancos de dados, na qual o usuário pode carregar os dados de várias formas e selecionar facilmente métodos e algoritmos, conforme os itens descritos na Figura 7.

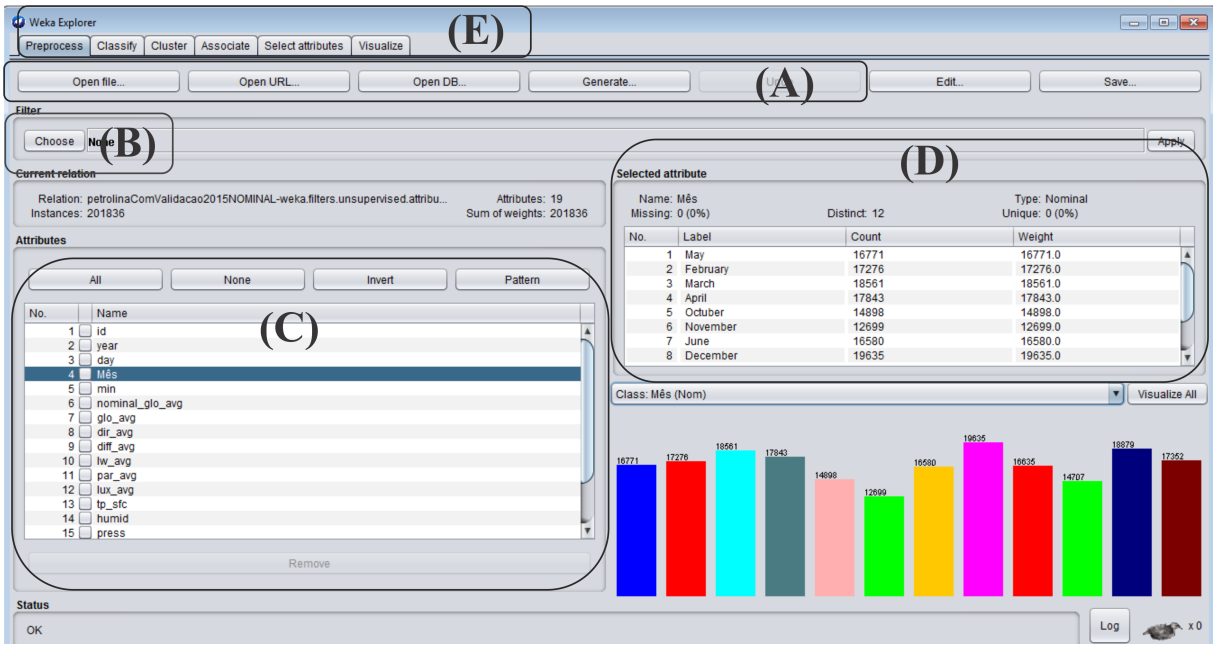

Figura 7 – Tela principal da ferramenta WEKA

Fonte: Autoria Própria

- Há um botão para criar um banco de dados e três configurações para entrada de dados no sistema conforme apresentadas em (A): na primeira, o usuário seleciona um arquivo que pode ser de diversos formatos, incluindo ARFF, CSV, JSON e C4.5; na segunda, uma URL da Web pode ser utilizada; por último, o usuário pode utilizar um banco de dados local e realizar uma conexão via JDBC;
- O usuário pode selecionar recursos de filtragem nos dados em (B), assim como mecanismos para processamento e transformação dos dados;
- Recursos de seleção de atributos para construção dos modelos podem ser selecionados via interface gráfica em (C);
- Informações sobre os dados são apresentadas em (D), tais como total de registros, tipo, valores máximo, médio, mínimo, distintos, iguais e de desvio padrão para cada atributo;

[2] <http://www.cs.waikato.ac.nz/ml/weka/>

- Outros recursos como métodos de classificação, *clustering*, regras de associação e visualização em dados podem ser acessados via botões em (E).

Na sua versão 3.9 (versão atual), a ferramenta incorporou diversos algoritmos relevantes de IA. Nesta proposta, selecionamos quatro abordagens que têm estratégias diferentes para previsão da velocidade do vento, tais como redes neurais, K-vizinhos mais próximos, máquina de vetores de suporte e árvore de decisão. Na prática, implementações dos três primeiros citados já foram citadas na literatura e revelaram bons resultados (ver Tabela 3). Em contraste, não foi encontrada em nossas pesquisas implementações utilizando algoritmos baseados na estrutura das árvores de decisão. Como são abordagens eficientes e ágeis para extrair conhecimento de gigantescos bancos de dados, foi decidido testar o M5p, um dos algoritmos de árvore de decisão que está implementado no WEKA.

3.4.1 Redes Neurais

Uma rede neural é um sistema paralelo distribuído, formado por unidades de processamento simples (ou nodos) que calculam determinadas funções matemáticas, normalmente não-lineares. O funcionamento dessa estrutura é inspirado em uma estrutura física concebida pela natureza: o cérebro humano (BRAGA; LUDERMIR; CARVALHO, 2000).

Dentre os modelos de redes neurais, o *Multilayer Perceptron* (MLP) é o mais utilizado e popular para previsão de sistemas complexos e não-lineares (CHANG; SHIN, 2006). MLP é uma rede *feed-forward* composta por inúmeras camadas de nodos interconectadas responsáveis por mapear, na forma unidirecional, um conjunto de entrada (*input layer*) em saídas apropriadas (*output layer*), passando por camadas ocultas (*hidden layers*) que modificam as funções de entrada, conforme apresentado na Figura 8.

Para treinamento da MLP, existe um método poderoso, computacionalmente eficiente e com boa capacidade de generalização conhecido como *backpropagation*. Tal abordagem supervisionada corrige os pesos de cada nodo partindo das camadas de saída até as de entrada para minimizar o erro calculado (BISHOP, 1995). O algoritmo de rede neural MLP está implementado no WEKA e seus parâmetros podem ser configurados via interface gráfica por um especialista.

3.4.2 Máquina de Vetores de Suporte

O algoritmo de Máquina de Vetores de Suporte (do inglês Support Vector Machine - SVM) é uma abordagem supervisionada relevante para tarefas de classificação e regressão em

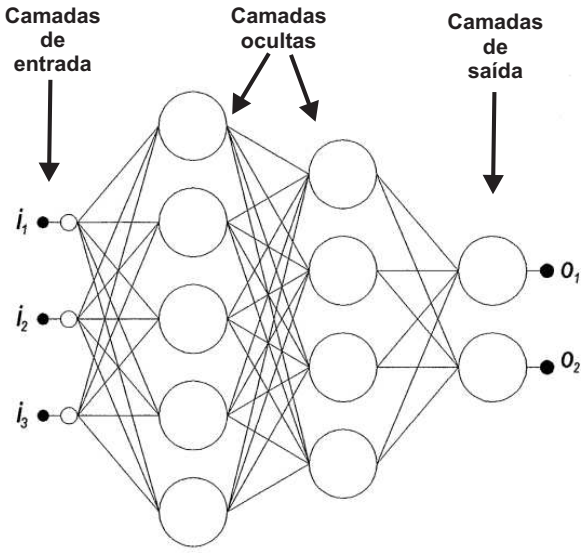

Figura 8 – Arquitetura de uma rede neural MLP

Fonte: Adaptada de (GARDNER; DORLING, 1998)

sistemas não-lineares. A ideia por trás do SVM consiste em desenhar um hiperplano para separar um conjunto de dados em duas classes. Para Cortes e Vapnik (1995), um hiperplano ótimo (do inglês *Optimal Hyperplane*) deve maximizar a margem ótima (do inglês *Optical margin*) entre os vetores das duas classe, ou seja, a distância entre o hiperplano e os elementos mais próximos até ele, conforme apresentado na Figura 9.

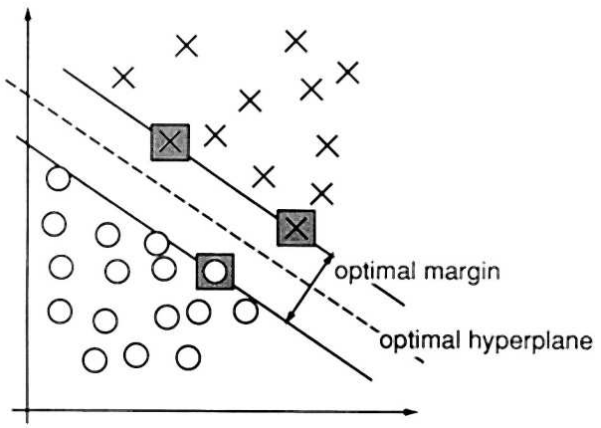

Figura 9 – Ideia do funcionamento do algoritmo de SVM

Fonte: (CORTES; VAPNIK, 1995)

No treinamento de um algoritmo de SVM, um problema de programação quadrática deve ser solucionado, do inglês *Quadratic Programming* (QP) (PLATT, 1998). *Sequential Minimal Optimization* (SMO) é um algoritmo robusto para treinamento de SVM, implementado no WEKA, conceitualmente simples, com excelente performance, fácil implementação, boa escalabilidade que passou por aprimoramentos em (SHEVADE et al., 2000) para resolver problemas difíceis. O problema de QP é dividido em subproblemas que são resolvidos rapidamente e analiticamente,

desprezando a necessidade de uma matriz extra para armazenamento, tornando-o o custo de memória para solução do problema linear ao conjunto de treinamento (PLATT, 1999). Por isso, SMO consegue ser mais rápido do que abordagens tradicionais de SVM em conjuntos de dados esparsos.

Para facilitar a simplicidade do cálculo e a capacidade de representar espaços abstratos, as técnicas de SVM implementam uma função *Kernel* que pode ser configurada no WEKA para alcançar sinais polinomiais, gaussianos e sigmoidais.

3.4.3 Árvore de Decisão

Uma árvore de decisão é uma estrutura hierárquica simples, intuitiva e de fácil compreensão que vem sendo utilizada para uma rápida descoberta de conhecimento exploratório em muitas áreas, tais como medicina, manufatura e produção, análise financeira, astronomia e biologia molecular (HAN; KAMBER; PEI, 2012). Sua estrutura é formada por nós, ramos e folhas conforme apresentado na Figura 10. Cada nó denota um teste para atributos, cada ramo uma saída do resultado desse teste e cada folha indica uma classe (ou uma distribuição de valor) (WITTEN et al., 2017). Os testes em atributos numéricos geralmente determinam se o valor de um atributo é maior ou menor do que uma constante pré-determinada. Em atributos nominais, eles verificam se a constante definida é igual ou diferente dos valores de uma classe. Desse modo, pode haver dois ou mais caminhos possíveis que envolvem diferentes decisões, constantes e atributos em toda a trajetória até as folhas.

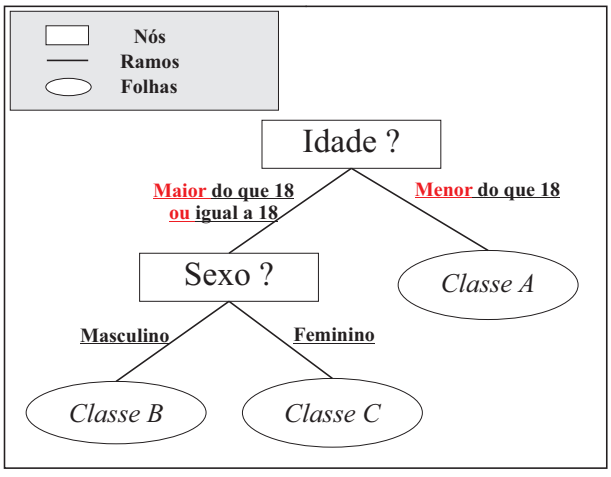

Figura 10 – Estrutura de uma árvore de decisão

Fonte: Autoria Própria

M5 é um algoritmo robusto, relevante e eficiente para extrair conhecimento de enormes bancos de dados que foi construído por Quilan baseado na ideia de divisão e conquista (do inglês *divide-and-conquer*). Isto significa que a instância de um problema é dividida em partes menores

e cada uma é resolvida separadamente; em seguida, as soluções menores se unem para produzir a solução do problema (QUINLAN, 1992). A saída do algoritmo é um modelo de árvore de decisão que maximiza a redução do erro esperado. No entanto, Quilan chama a abordagem de "*model tree*" pois o algoritmo tem expressões lineares nas folhas em vez de um valor numérico (como o algoritmo CART). Por isso, M5 consegue ser mais compreensível e acurado do que outras abordagens de árvore de decisão nas previsões de valores numéricos (QUINLAN, 1992).

A implementação do algoritmo M5 é chamada de M5p no WEKA devido aos aprimoramentos e alterações que o algoritmo recebeu na literatura (WANG; WITTEN, 1997). De acordo com os experimentos de Blomberg, Hemerich e Ruiz (2013), o algoritmo M5p alcançou melhores resultados do que outros algoritmos implementados no WEKA, tais como RepTree, KNN, SVM, RL e NN, quando foi analisado em vinte bancos de dados públicos distintos. Por isso, selecionamos o M5p para previsão da velocidade do vento.

3.4.4 Algoritmo KNN

O algoritmo do K-Vizinhos Mais Próximos (K-Nearest Neighbors - KNN) é um dos algoritmos mais simples, de fácil implementação que pode fornecer bons resultados, considerando que o algoritmo usa instâncias vizinhas para resolver de forma incremental tarefas supervisionadas de classificação e regressão. O KNN demanda dois elementos chaves antes de sua execução: (i) o valor de K que representa a quantidade de elementos vizinhos que serão analisados para determinar a instância alvo não rotulada; (ii) a métrica de cálculo de distância entre pontos, tais como distância euclidiana (mais comum), *Manhattan distance*, *Minkowski distance* e *Chebyshev distance* (OOI; NG; LIM, 2013; SINGH; YADAV; RANA, 2013).

O funcionamento do algoritmo é apresentado na Figura 9, na qual um elemento não rotulado será definido baseado na vizinhança formada por dois elementos rotulados. Assim que K é definido, a distância do elemento desconhecido até os K elementos é calculada; desse modo, tal elemento é direcionado a uma das classes especificadas. Em geral, na abordagem de regressão o elemento desconhecido recebe a média dos valores de sua vizinhança, enquanto que na classificação é considerada a votação majoritária de uma classe.

IBK é uma estratégia implementada no WEKA baseada em KNN, que traz a ideia de que um ponto P e seus vizinhos mais próximos pertencem a uma classe em um espaço n-dimensional (AHA; KIBLER; ALBERT, 1991). A saída do algoritmo IBK é uma descrição conceitual, ou seja, uma função que mapeia instâncias para categorias. O especialista tem a possibilidade de determinar o valor de K de forma manual ou automática no WEKA.

Embora cada método e algoritmo seja pré-configurado na ferramenta WEKA, para obter maior desempenho e precisão nos resultados, o especialista deve compreender e ajustar cada

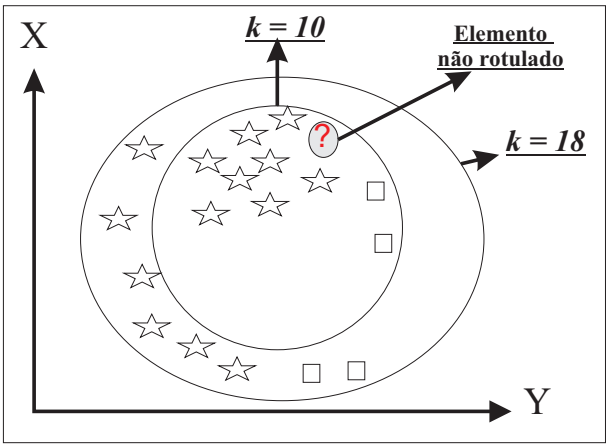

Figura 11 – Ilustração do algoritmo KNN

Fonte: Autoria Própria

parâmetro de acordo com as características do banco de dados. Para rede neural os elementos de maior influência nos resultados são: raio de aprendizagem, tempo de treinamento (ou critério de parada) e quantidade de camadas ocultas (ou arquitetura da rede). Para SVM, o Kernel e a forma de aprendizagem são os essenciais elementos de configuração. No ajuste do M5p, o número de instâncias mínimas das folhas é uma característica relevante. Por fim, o valor de K e a estratégia de busca são requisitos para o funcionamento do KNN (ou IBK).

Além de fornecer algoritmos relevantes, o WEKA dispõe de quatro métodos de validação dos modelos que definem como o banco de dados será particionado em conjunto de treinamento e de validação: no *Use Training Set*, o conjunto de treinamento e de validação são os mesmos, formados por toda a base de dados disponível; no *Supplied Test Set*, a uma nova base de dados externa pode ser carregada via arquivo para validação do modelo; no *Percentage Split*, um valor percentual entre 0 e 100 é fornecido para ser conjunto de treinamento e o remanescente é usado como validação. Normalmente, recomenda-se definir dois terços do banco de dados para treinamento e um terço para validação, o que é equivalente a cerca de 66% do banco de dados (KOHAVI, 1995). No *Cross-validation*, os dados são particionados randomicamente em F subconjuntos iguais conhecido como *folds*. A cada iteração, uma das partições $f1$ é utilizada para treinamento e a remanescente como conjunto de validação. O valor de F é definido pelo especialista, normalmente é recomendado definir 10 partições, pois assim um erro confiável é estimado com baixo viés e variância (HAN; KAMBER; PEI, 2012; WITTEN et al., 2017).

Dentre essas abordagens, a mais popular e recomendada é a *Cross-validation* para uma validação rigorosa e confiável. Porém, se há enormes volumes de dados disponíveis, a estratégia de 1/3 para validação e 2/3 para treinamento também é válida.

Em nossa abordagem implicamos diversos recursos relevantes de mineração de dados para contornar as limitações dos modelos desenvolvidos na literatura. Tais recursos podem fornecer melhorias significativas nas previsão da velocidade do vento, as quais são discutidas adiante.

4 ESTUDOS DE CASOS

Com o objetivo de validar a abordagem proposta, dois estudos foram realizados com duas bases de dados e diversos modelos de previsão da velocidade do ventos foram construídos. O sistema de gerenciamento de bancos de dados PostgreSQL foi usado para gerenciar os dados.

4.1 ESTUDO DE CASO A - PCD DE PETROLINA

No primeiro estudo, um banco de dados foi construído a partir dos dados da rede SONDA. Os dados foram de uma plataforma automática de coleta de dados (PCD) localizada no município de Petrolina, no Pernambuco: latitude 09° 04’ 08"Sul; longitude 40° 19’ 11"Oeste; e 387 metros de altitude. Os dados adquiridos são registros de 2009 até 2016 a cada minuto e possuem os atributos ano, dia do ano, minuto do dia, temperatura, umidade relativa, velocidade do vento e direção do vento a 10 metros do solo.

Neste estudo, foi definido 2/3 do banco de dados para construção dos modelos e 1/3 foi usado para validação, pois havia cerca de 4 milhões de registros disponíveis. Os quatros algoritmos utilizados são executados para construção de modelos de previsão horária e diária: NN, SVM, M5p e KNN.

Todas as implementações foram realizadas através de um computador portátil compatível com um processador Intel de oito núcleos, oito gigabytes de memória RAM e sistema operacional Windows 7 de 64 bits.

4.1.1 Pré-processamento e Transformação

Seguindo a abordagem proposta por esse trabalho, foram utilizados scripts em linguagem Python, comandos SQL e gráficos gerados através do Excel para detectar e sinalizar anomalias dentro do banco de dados. Diversas abordagens de processamento foram implementadas, no entanto, uma relatada adiante se mostrou mais relevante.

Além dos dados, a rede SONDA fornece um arquivo para sinalizar possíveis anomalias nos seus repositórios. A sinalização é determinada através de três critérios estabelecidos com base em informações do Centro de Recurso Meteorológico do Canadá (do inglês *Meteorological*

Resource Center - MRC) [1] e nas normais climatológicas do INMET [2]: (1) dados são sinalizados através de limiares se considerados fisicamente impossíveis; (2) dados são suspeitos se a variação do atributo é extremamente rara durante um certo período de tempo; por último, (3) dados são detectados quando há uma evolução temporal não condizente com o valor esperado para a variável num período de tempo. Esses critérios são detalhados para cada variável (VR) na Tabela 4. Para temperatura (TR), velocidade do vento (VV) e direção do vento (DV) os três critérios foram considerados, enquanto que para umidade relativa (UR) somente um critério foi determinado e pressão do ar (PA) dois deles.

O critério que lida com dados fisicamente impossíveis para temperatura e pressão do ar não foram sinalizados de forma adequada no arquivo fornecido pela SONDA, pois são determinados através das características locais da região. Para temperatura foi definido um limite de 15 a 40 °C em razão do estudo em (RAMOS et al., 2011a). Para pressão do ar foi definido um limiar de 950 até 980 mb em razão da altitude de 387 metros. Ambos os limiares e demais critérios de processamento mostraram-se relevantes, após diversos testes realizados com os dados da PCD de Petrolina.

Tabela 4 – Critérios estabelecidos para checagem dos dados segundo as normais climatológicas do INMET e do MRC

VR	Fisicamente Impossível	Extremamente Raro	Evolução Temporal
TR	min e máx para o local	variação $< 5°$ num período de 1 h	variação $> 0{,}5°$ num período de 12 h consecutivas
UR	min = 0 e máx = 100%	não aplicado	não aplicado
PA	min e máx segundo a altitude	variação < 6 num período de 3 h consecutivas	não aplicado
VV	min = 0 e máx = 25 m/s	variação $> 0{,}1$ num período de 3 h consecutivas	variação $> 0{,}5$ num período de 12 h consecutivas
DV	min: 0° e max: 360°	variação $> 1°$ num período de 3 h consecutivas	variação $> 10°$ num período de 18 h consecutivas

Fonte: Adaptada de MRC e INMET

A sinalização nos dados foi expressa através de três valores inteiros: “2” aponta que o atributo é suspeito de incorreção; “5” se o algoritmo não conseguiu executar por algum motivo; e “9” se o atributo for considerado de boa qualidade e atravessar todos os critérios estabelecidos. A sequência do código de erro corresponde respectivamente ao primeiro, ao segundo e ao terceiro algoritmo. Em outras palavras, se uma sinalização para um atributo *D1* de temperatura for “925”. Isto significa que *D* é fisicamente possível, porém *D1* tem uma variação rara. Como o algoritmo de evolução temporal não executou, o especialista deve verificar manualmente e tomar decisões relacionadas à *D1*.

Para exemplificar a decisão do especialista, considere um vetor X composto por oito elementos de dados de temperatura [30,10; 30,10; 30,20; 30,10; 3,01; 0,00; 39,00; 0,00] e sua sinalização X’ [999; 999; 999; 999; 222; 255; 922; 255]. A sinalização indica que o quinto, sexto, sétimo e oitavo elemento desse vetor são atributos suspeitos. Desse modo, o especialista analisa e percebe através de recursos visuais que o quinto elemento está incorreto devido à um problema de transmissão, na qual em vez de 30,1 o dado foi expresso 3,1. Nesses casos, foram corrigidos

[1] http://www.webmet.com

[2] http://sonda.ccst.inpe.br/infos/validacao.html

manualmente cada um dos atributos no banco de dados usando recursos do Excel. Nos demais casos, os dados suspeitos (0,00; 39,00; 0,00) foram ignorados através de um comando SQL pois provavelmente são erros de medição dos instrumentos.

Com os dados pré-processados, duas agregações foram consideradas usando a média dos atributos para construção dos modelos de previsão propostos: (1) dados foram transformados de 1 minutos para 1 hora e o banco de dados foi reduzido de 3683510 para 61934 registros; (2) dados foram convertidos de 1 minuto para 1 dia, nesse caso, a redução foi ainda maior de 3683510 para 2595 registros.

Além disso, quatro generalizações foram consideradas com o objetivo de alcançar um maior precisão e performance nos resultados: (1) um novo atributo mês nominal foi adquirido baseado no dia do ano, ou seja, dia do ano de 1 a 31, então mês é "Janeiro"; (2) da mesma forma, um novo atributo mês numérico foi gerada na qual dia do ano entre 1 e 31 então mês numérico é "1"; (3) um atributo hora do dia foi conseguido baseado no minuto do dia, o que significa que minuto do dia entre 1 e 60, a hora é "1"; (4) o atributo direção do vento numérico deu origem ao novo atributo direção nominal baseado nas oito classes da rosa dos ventos: Norte, Sul, Oeste, Leste, Noroeste, Nordeste, Sudeste, Sudoeste; por exemplo, se a direção do vento for maior do que 337,5° ou menor do que 22,5° então o novo atributo direção nominal recebe "Norte".

Para finalizar a transformação dos dados, normalizamos todos os atributos considerando duas casas decimais após a vírgula, ou seja, um atributo velocidade do vento com valor de 3,5453 tornou-se 3,54.

Todos os critérios de processamento e transformação nos dados para cada modelo foram implementados através de um único comando SQL: no apêndice A a descrição para os modelos horários é apresentada; apêndice B tem o código referente aos modelos diários. Ambos os comandos são capazes de gerar um arquivo no formato CSV que pode ser interpretado pelo WEKA, fornecendo uma maneira flexível e ágil para implementar técnicas de processamento e transformação.

4.1.2 Construção dos modelos

De acordo com a abordagem de mineração de dados elaborada neste trabalho, na etapa de construção dos modelos três atividades são realizadas: definição do conjunto de entrada, ajuste do algoritmo e validação.

Neste estudo, foi selecionado o algoritmo M5p para definir o conjunto de entrada dos modelos pois se apresentou superior a MLP, SVM e KNN em termos de performance, levando menos de cinco segundos para cada um dos 63 testes realizados. Na definição dos atributos de entrada dos modelos horários, o coeficiente de correlação *R* foi analisado para os seis atributos

numéricos dessa proposta, conforme apresentado na Tabela 5.

Tabela 5 – Análise do conjunto de entrada para os modelos de previsão da PCD de Petrolina através de *R*

n°	Entrada	R	n°	Entrada	R	n°	Entrada	R
1	HD	0.5448	22	HD, MS, TP	0.7530	43	HD, MS, TP, PA	0.8040
2	MS	0.3169	23	HD, MS, UR	0.7740	44	HD, MS, TP, DV	0.8136
3	TP	0.3199	24	HD, MS, PA	0.7413	45	HD, MS, UR, PA	0.8139
4	UR	0.4771	25	HD, MS, DV	0.7748	46	HD, MS, UR, DV	0.8243
5	PA	0.3556	26	HD, TP, UR	0.7721	47	HD, MS, PA, DV	0.8113
6	DV	0.5079	27	HD, TP, PA	0.7612	48	HD, TP, UR, PA	0.8204
7	HD, MS	0.6545	28	HD, TP, DV	0.7568	49	HD, TP, UR, DV	0.8225
8	HD, TP	0.6288	29	HD, UR, PA	0.7866	50	HD, TP, PA, DV	0.8167
9	HD, UR	0.7064	30	HD, UR, DV	0.7813	51	HD, UR, PA, DV	0.8294
10	HD, PA	0.6654	31	HD, PA, DV	0.7748	52	MS, TP, UR, PA	0.7961
11	HD, DV	0.7150	32	MS, TP, UR	0.6890	53	MS, TP, UR, DV	0.7712
12	MS, TP	0.5621	33	MS, TP, PA	0.7455	54	MS, TP, PA, DV	0.8032
13	MS, UR	0.5968	34	MS, TP, DV	0.7122	55	MS, UR, PA, DV	0.7855
14	MS, PA	0.5141	35	MS, UR, PA	0.7297	56	TP, UR, PA, DV	0.8135
15	MS, DV	0.5953	36	MS, UR, DV	0.7119	57	HD, MS, TP, UR, PA	0.8348
16	TP, UR	0.5997	37	MS, PA, DV	0.6800	58	HD, MS, TP, UR, DV	0.8492
17	TP, PA	0.6926	38	TP, UR, PA	0.7629	59	HD, MS, TP, PA, DV	0.8436
18	TP, DV	0.6366	39	TP, UR, DV	0.7150	60	HD, MS, UR, PA, DV	0.8496
19	UR, PA	0.6935	40	TP, PA, DV	0.7739	61	HD, TP, UR, PA, DV	0.8521
20	UR, DV	0.6662	41	UR, PA, DV	0.7627	62	MS, TP , UR, PA, DV	0.8348
21	PA, DV	0.6375	42	HD, MS, TP, UR	0.8143	63	HD, MS, TP , UR, PA, DV	0.8687

Fonte: Autoria Própria

Em geral, a maioria dos conjuntos que combinaram quatro, cinco e seis atributos mostraram-se relevantes para previsão da velocidade do vento na região de estudo, pois tiveram valores maiores do que 0,8000 para *R*. Os modelos que combinaram uma quantidade inferior de atributos tiveram *R* abaixo de 0,8000, o que significa que terão resultados inferiores em termos de precisão.

Um combinação que destacou-se foi o modelo n° 7, com dois dos atributos ignorados nos trabalhos apresentados na Tabela 5 (hora do dia e mês), que alcançou uma correção moderada com 0,6545 para *R*. No entanto, o melhor conjunto foi o modelo n° 63 com todas as variáveis que se apresentaram relevantes para previsão do vento na região de estudo. Porém, se as variáveis estiverem indisponíveis por algum motivo, o especialista pode conferir e utilizar o melhor conjunto possível baseado nos valores de *R*.

Assim que foi definido o conjunto formado pelos seis atributos numéricos, ajustamos os parâmetros de cada algoritmo e selecionamos alguns atributos generalizados no formato nominal para construir os modelos horários. Diversos testes foram implementados para ajustar os parâmetros de cada um dos quatro algoritmos propostos neste trabalho. Apenas os resultados dos modelos horários mais relevantes são apresentados na Tabela 6. As cinco métricas estáticas foram apresentadas, porém consideramos apenas o MAE para comparação de modelos desenvolvidos a partir de uma mesma base de dados e SMAPE para comparação de modelos com base de dados distintas.

Os quatro modelos desenvolvidos através das redes neurais MLP alcançaram resultados de precisão relevantes (n° 1 ao 4). Em especial, o modelo n° 4 que combinou na forma nominal

Tabela 6 – Resultados dos modelos relevantes para previsão horária de ventos na PCD de Petrolina

n°	Algoritmo	Entrada	Configuração	MAE	RMSE	MSE	MAPE	SMAPE	Tempo(s)
1	MLP	HD, MS, TP, UR, PA, DV	RA = 0,3; TT = 500; CO = aut	0,6464	0,8039	0,6464	34,40	7,04	21,38
2	MLP	HD, MS_n, TP, UR, PA, DV	RA = 0,3; TT = 500; CO = aut	0,5869	0,7499	0,5624	26,36	6,49	102,2
3	MLP	HD, MS, TP, UR, PA, DV_n	RA = 0,3; TT = 500; CO = aut	0,5975	0,7612	0,5795	27,32	6,59	65,36
4	MLP	HD, MS_n, TP, UR, PA, DV_n	RA = 0,3; TT = 500; CO = aut	0,5737	0,7316	0,5352	25,21	6,34	184,1
5	M5p	HD, MS, TP, UR, PA, DV	Instância mínima = 4	0,5552	0,7179	0,5155	24,09	6,09	6,84
6	M5p	HD, MS_n, TP, UR, PA, DV	Instância mínima = 4	0,5556	0,7191	0,5171	23,97	6,09	8,86
7	M5p	HD, MS, TP, UR, PA, DV_n	Instância mínima = 4	0,5678	0,7341	0,5389	24,45	6,22	7,81
8	M5p	HD, MS_n, TP, UR, PA, DV_n	Instância mínima = 4	0,5683	0,7355	0,5409	24,46	6,22	10,14
9	SVM	HD, MS, TP, UR, PA, DV	PolyKernel	0,6396	0,8114	0,6584	26,45	6,67	3259,25
10	KNN	HD, MS, TP, UR, PA, DV	k = 13, KDTree	0,5495	0,7046	0,4965	23,03	6,05	17,51
11	KNN	HD, MS_n, TP, UR, PA, DV	k = 13, KDTree	0,5513	0,7085	0,5021	23,17	6,05	16,12
12	KNN	HD, MS, TP, UR, PA, DV_n	k = 20, KDTree	0,5558	0,7143	0,5102	23,64	6,11	16,14
13	KNN	HD, MS_n, TP, UR, PA, DV_n	k = 16, KDTree	0,5559	0,7187	0,518	23,86	6,15	15,55

Fonte: Autoria Própria

os novos atributos mês (MS_n) e direção do vento (DV_n) para atingir os resultados mais relevantes da categoria: 0,5737 para MAE, 0,7316 para RMSE, 0,5352 para MSE, 25,21% para MAPE, 6,34% para SMAPE. A rede neural MLP levou 184 segundos na execução quando foi configurada com 0,3 para o Raio de Aprendizagem (RA), 500 para o tempo de treinamento (TT) e as camadas ocultas (CO) foram definidas de forma automática através de testes automáticos realizados no WEKA que minimizam o erro esperado. A performance da MLP variou de acordo com a configuração de parâmetros. O tempo de treinamento foi o parâmetro que mais interferiu nos resultados de precisão e performance. Observou-se que valores mais altos do que 500 podem fornecer resultados de precisão um pouco melhores. No entanto, a performance do modelo diminui consideravelmente. Já os valores abaixo de 500 tornavam a execução do algoritmo mais rápida porém forneciam resultados de precisão bem inferiores. O valor 500 para TT foi um número ajustado exaustivamente que manteve um bom nível de performance e precisão.

Os modelos gerados através do M5p (n° 5 ao 8) também foram relevantes superando moderadamente todos os modelos gerados via MLP em relação à precisão dos resultados. O modelo n° 5, configurado com instância mínima 4, foi o mais relevante da categoria, atingindo 0,5552 para MAE, 0,7179 para RMSE, 0,5155 para MSE, 24,09% para MAPE, 6,09% para SMAPE e apenas 6,84 segundos foram levados para construção do modelo. O modelo n° 8 que combinou dados nominais teve os resultados de precisão relativamente mais baixos de sua categoria. Falando de performance, M5p mostrou superioridade e em menos de 10 segundos todos os seus modelos executaram previsões horárias.

Não obstante, SVM foi o único algoritmo inexequível para este estudo. Embora uma variedade de configurações tenha sido testada, o melhor modelo n° 9 que foi construído com um Kernel polinomial denominado de PolyKernel, levou 3259,25 segundos, o que mostrou-se inviável para geração de modelos horários de previsão. Possivelmente, SVM poderá trazer bons resultados em estudos com volume de dados reduzidos, considerado que o alto tempo nesses experimentos se deu em razão das 61934 instâncias de dados.

Para nossa surpresa, o algoritmo KNN alcançou os melhores resultados de precisão para

previsão na PCD de Petrolina. O modelo soberano n° 10 que combinou seis atributos numéricos, teve 13 partições e foi configurado com um algoritmo de busca em árvore denominado de KDTree (FREIDMAN; BENTLEY; FINKEL, 1977), alcançando 0,5495 para MAE, 0,7046 para RMSE, 0,4965 para MSE, 23,03% para MAPE, 6,05% para SMAPE e levando 17,31 segundos em sua execução.

Em geral, para construção dos modelos horários usando os dados das PCD de petrolina, de acordo com a métrica MAE: KNN com 0,5495 para MAE foi um pouco superior ao M5p que teve 0,5552 para MAE e foi relativamente superior a MLP com 0,5737 para MAE, falando em relação aos resultados de precisão. Já o SVM foi totalmente inviável nos experimentos levando quase uma hora para construção de um modelo de previsão horária.

Além dos modelos horários, neste estudo foi proposta a construção de modelos para previsão diária na PCD de Petrolina. Novamente, uma série de experimentos e análises foram elaboradas e os modelos mais relevantes para previsão diária são apresentados na Tabela 7. Dessa vez, foram considerados sete atributos em vez de oito: mês (nominal e numérico), temperatura, pressão, velocidade do vento e direção do vento (nominal e numérico). Os dados usados para os modelos de previsão diária foram a média dos valores diários de cada um desses atributos, sendo desprezado o atributo "hora do dia".

Tabela 7 – Resultado dos modelos relevantes para previsão diária de ventos na PCD de Petrolina

n°	Algoritmo	Entrada	Configuração	MAE	RMSE	MSE	MAPE	SMAPE	Tempo(s)
1	MLP	MS, TP, UR, PA, DV	RA = 0,5; TT = 500; CO = aut	0,3669	0,4712	0,2221	14,16	3,59	0,77
2	MLP	MS_n, TP, UR, PA, DV	RA = 0,5; TT = 500; CO = aut	0,3484	0,4495	0,2021	13,70	3,44	3,78
3	MLP	MS, TP, UR, PA, DV_n	RA = 0,5; TT = 500; CO = aut	0,3708	0,4765	0,2270	14,29	3,63	1,52
4	MLP	MS_n, TP, UR, PA, DV_n	RA = 0,5; TT = 500; CO = aut	0,3487	0,4517	0,2040	13,92	3,47	5,58
5	M5p	MS, TP, UR, PA, DV	Instância mínima = 4	0,3558	0,4615	0,2130	13,85	3,49	0,29
6	M5p	MS_n, TP, UR, PA, DV	Instância mínima = 4	0,3477	0,4461	0,1990	13,69	3,44	0,24
7	M5p	MS, TP, UR, PA, DV_n	Instância mínima = 4	0,3579	0,4591	0,2108	13,98	3,52	0,17
8	M5p	MS_n, TP, UR, PA, DV_n	Instância mínima = 4	0,3548	0,4540	0,3061	13,81	3,49	0,28
9	SVM	MS, TP, UR, PA, DV	Kernel Puk	0,3431	0,4388	0,1926	13,50	3,37	6,65
10	SVM	MS_n, TP, UR, PA, DV	Kernel Puk	0,3530	0,4466	0,1995	13,91	3,47	7,16
11	SVM	MS, TP, UR, PA, DV_n	Kernel Puk	0,3406	0,4368	0,1908	13,35	3,34	6.49
12	SVM	MS_n, TP, UR, PA, DV_n	NormalizedPoly Kernel	0,3448	0,4463	0,1992	13,56	3,40	7,07
13	KNN	MS, TP, UR, PA, DV	k = 12, KDTree	0,3578	0,4537	0,2059	13,59	3,50	0,70
14	KNN	MS_n, TP, UR, PA, DV	k = 12, KDTree	0,3641	0,4607	0,2123	13,78	3,56	0,61
15	KNN	MS, TP, UR, PA, DV_n	k = 19, KDTree	0,3604	0,4632	0,2100	13,85	3,54	0,66
16	KNN	MS_n, TP, UR, PA, DV_n	k = 11, KDTree	0,3733	0,4751	0,2257	14,30	3,67	0,66

Fonte: Autoria Própria

No geral, todos os algoritmos alcançaram excelentes resultados de performance, considerando que em menos de 8 segundos construíram um modelo para previsão diária a partir de 1713 registros de dados.

Os modelos de previsão construídos através das redes neurais MLP (n° 1 ao 4) forneceram relevantes resultados quando configurado com 0,5 para raio de aprendizagem, 500 para tempo de treinamento e camadas ocultas definidas de forma automática através de testes que minimizam o erro. Sobretudo, usando mês nominal e quatro atributos numéricos, o modelo n° 2 que atingiu 0,3484 para MAE, 0,4495 para RMSE, 0,2021 para MSE, 13,70% para MAPE e 3,44% para

SMAPE. Esse modelo foi minimamente superior ao modelo n° 4 que teve mês e direção do vento representados na forma nominal. Percebeu-se que a rede neural usando dados nominais levou uma vantagem notória na precisão de resultados em relação aos modelos que usaram somente dados numéricos.

Os modelos gerados a partir do M5p também alcançaram resultados relevantes para previsão diária que foram bem compatíveis com os modelos gerados via MLP. No entanto, o modelo n° 6 ainda foi muito pouco superior à todos os modelos gerados via MLP, atingindo 0,3477 para MAE, 0,4461 para RMSE, 0,1990 para MSE, 13,69% para MAPE e 3,44% para SMAPE.

Os modelos de SVM que foram desprezados nas previsões horárias em razão da baixa performance. Todavia, eles foram os mais relevantes para previsão diária da velocidade do vento. O modelo n° 11, configurado com o Kernel PUK (ÜSTÜN; MELSSEN; BUYDENS, 2006), superou um pouco todos os modelos em termos de resultados de precisão, alcançando valores mínimos como 0,3406 no MAE, 0,4368 no RMSE, 0,1908 no MSE, 13,35% no MAPE e 3,34% no SMAPE.

Outrossim, os modelos gerados através do algoritmo KNN também tiveram bons resultados, mas ficaram um pouco aquém dos demais modelos na precisão. Dentre os modelos construídos com KNN, o modelo n° 13 configurado com 12 vizinhos próximos e cinco atributos numéricos foi o que alcançou melhores resultados com 0,3578 para MAE, 0,4537 para RMSE, 0,2059 para MSE, 13,59% para MAPE e 3,50% para SMAPE.

Mediante o exposto, os modelos para previsão diária, usando os dados das PCD de petrolina, tiveram resultados relevantes e bem equivalentes. Analisando o MAE dos melhores modelos: o SVM com 0,3406 para MAE (n° 11) alcançou uma vantagem mínima acima do M5p que teve 0,3477 para MAE (n° 6); MLP atingiu 0,3484 e ficou minimamente atrás desses modelos, porém um pouco superior ao KNN que teve 0,3578 para MAE.

4.2 ESTUDO DE CASO B - TURBINA EÓLICA

No primeiro estudo de caso foram alcançados resultados importantes nos modelos para previsão diária e horária usando dados da PCD de Petrolina. Contudo, manifestou-se o interesse de construir outros modelos de previsão usando dados de uma turbina eólica, tendo em vista que nossa abordagem tem o objetivo de auxiliar os operadores de energia.

As indústrias eólicas, principalmente na região Nordeste do Brasil, enfrentam uma forte competição no mercado de eletricidade. Operadores de energia têm produzido eletricidade de forma transparente aos seus concorrente com a finalidade de evitar uma competição direta em

uma determinada região. Em razão disso, conseguir dados de uma turbina eólica foi um desafio.

Felizmente, os bons resultados que alcançamos no primeiro estudo e uma forte parceria da FUNCEME foram dois dos fatores que conduziram uma interação com empresas de energia, que se mostraram interessadas em nosso trabalho. Uma dessas empresas nos forneceu uma boa quantidade de dados sigilosos de uma turbina eólica. Nesse ínterim, conquistamos registros de 02 de fevereiro de 2013 até 15 de novembro de 2016 disponibilizados a cada 10 minutos, tais como pressão (hPa), temperatura (°C) e umidade do ar (%), velocidade do vento (a 50, 83 e 85 m em m/s) e direção do vento (a 50m e 85m) assim como uma data e um minuto do dia expressos através de valores inteiros.

Neste estudo propomos construir modelos para previsão horária, diária, três dias à frente e semanal. Em outras palavras, vamos avaliar nossa abordagem para previsão de prazo ultra curto, curto, médio e longo. Queremos avaliar se a proposta é capaz de fornecer resultados relevantes em todos os problemas de previsão. Para validação dos modelos, determinamos a técnica *Cross-Validation* configurada com 10 partições, pois o volume do dados foi reduzido em relação ao primeiro estudo.

Novamente, todas as implementações foram realizadas através de um computador portátil compatível com um processador Intel de oito núcleos, oito gigabytes de memória RAM e sistema operacional Windows 7 de 64 bits.

4.2.1 Pré-processamento e Transformação

Para alcançar um nível adequado de processamento, precisamos conhecer a área de estudo, porém não tivemos nenhum conhecimento sobre a região de origem dos dados. No entanto, os dados que foram fornecidos atravessaram uma análise de processamento na qual um código de erro é formado adicionando os seguintes erros individuais:

- (0) nenhum erro;
- (1) pressão do ar (barômetro);
- (2) velocidade do vento a 85m;
- (4) direção do vento a 83m;
- (8) velocidade de vento a 50m;
- (16) direção do vento a 50m;
- (32) velocidade do vento a 83;

- (64) temperatura;
- (128) umidade do ar.

Isto significa que registros expressos com um código de erro “131” são suspeitos de inconsistências para pressão do ar, velocidade do vento a 85 metros e umidade do ar, que são a soma de 1, 2 e 128, respectivamente. De fato, registros com códigos de erros maiores do que zero têm atributos considerados inconsistentes. Em razão disso, um processamento nos dados foi realizado considerando apenas registros consistentes, ou seja, aqueles que tiveram código de erro igual a zero.

O processamento fornecido mostrou-se relevante através de uma análise de gráficos e consultas SQL. Após o processamento, todos os atributos tiveram uma variação pertinente para os três anos de dados, conforme ilustrado na Figura 12: temperatura do ar variou no intervalo de 21,1 a 33,2 °C; umidade relativa na faixa de 34,4 a 99,6%; pressão atmosférica de 996 a 115 hpa; velocidade do vento a 50m de 0,2 a 16,6 m/s e a 80m de 0,2 a 18,5 m/s. A umidade do ar foi relativamente alta, o que indica que os dados fornecidos sejam de uma turbina eólica instalada próximo ao litoral. A pressão atmosférica teve uma variação predominantemente de 1000 a 1005 hpa para 80% da sequência dos dados, enquanto que nos 20% restantes, a pressão variou de 110 a 115 hpa, talvez em função da redução da altura do instrumento (barômetro). Observou-se que os valores de velocidade do vento a 50 m foram relativamente mais baixos do que os de 85m, o que pode ser dar em razão da maior influência de efeitos físicos de montanhas e obstáculos. No geral, os dados apresentaram-se consistentes após o processamento.

Na etapa de transformação, quatro agregações foram consideradas com base na média aritmética para cada um dos quatros modelos de previsão propostos. Os dados processados foram convertidos de 10 minutos: (i) para uma hora através da média de valores e 181796 registros tornaram-se 30640; (ii) para um dia com uma redução de 181796 para 1301 registros; (iii) para 3 dias com uma redução de 181796 para 516 registros; (iv) por fim, agregamos os atributos de 10 minutos para alcançar registros semanais (redução de 181796 para 174 registros).

Ainda na etapa de transformação, sete generalizações foram consideradas: (1) O mês numérico foi extraído do atributo data, ou seja, se data foi “20141012” então mês numérico foi “10”; (2) o mês na forma nominal foi alcançado através do mês numérico, o que significa que se mês numérico é “1” então mês nominal é “Janeiro”; (3) O atributo dia do mês foi adquirido a partir da data inteira, ou seja, para data “20141012” o dia do mês é “12”; (4) um novo atributo número da semana do mês foi generalizado a partir do dia do mês, ou seja, se dia do mês for entre 1 e 8 então semana do mês é “1”; (5) um novo atributo três dias do mês foi obtido através do dia do mês, ou seja, se dia do mês for entre 1 e 3, o atributo três dias do mês é “1”; (6) o atributo hora do dia foi generalizado a partir do minuto do dia, sendo assim o minuto do dia “110000” então a hora é “11”; (7) um atributo direção nominal foi obtido a partir do direção inteiro seguindo a ideia das oito classes da rosa dos ventos, ou seja, para direção maior do que 112,5° e menor ou

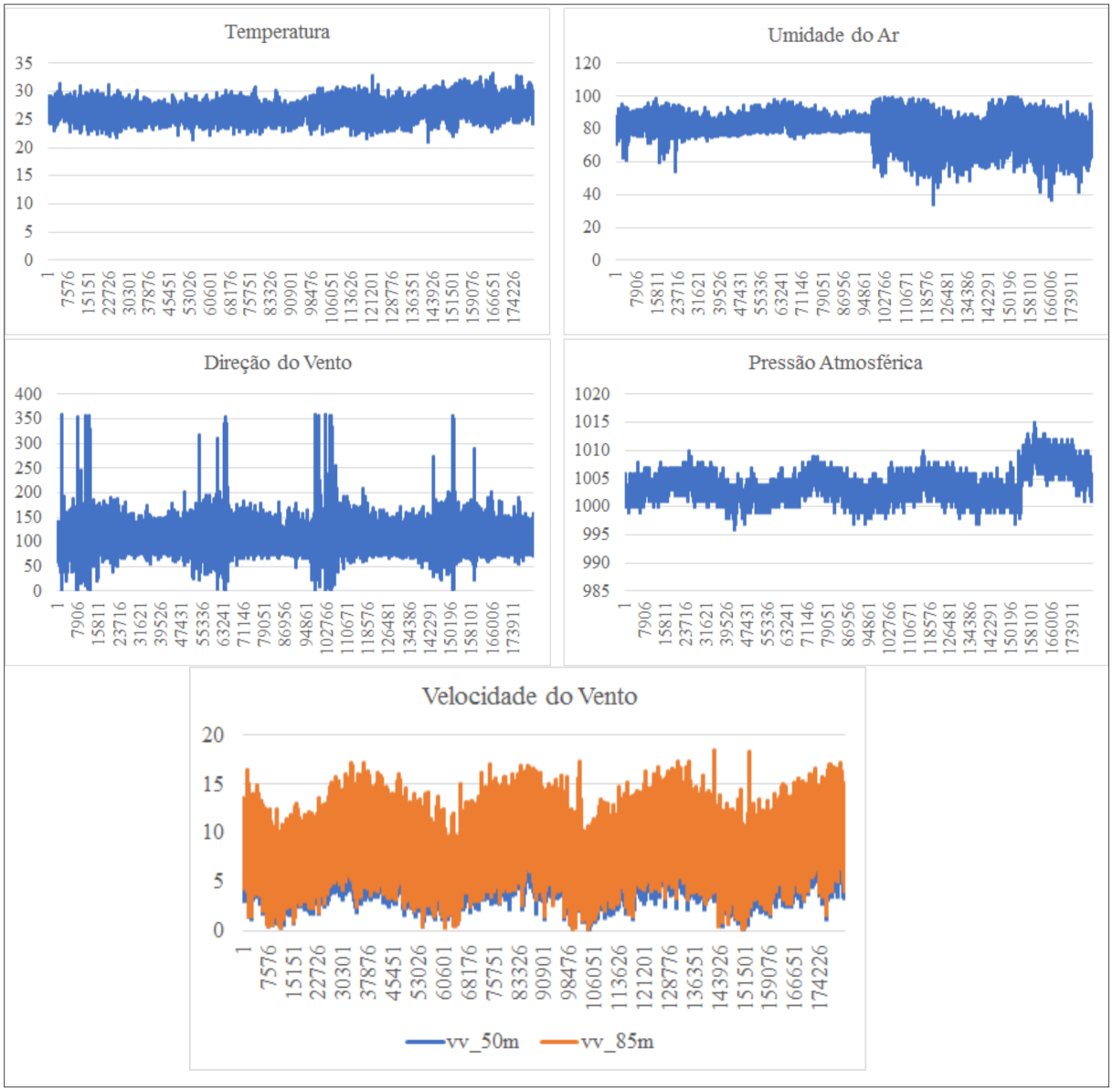

Figura 12 – Gráficos gerados a partir de dados processados de uma turbina a cada 10 minutos

Fonte: Autoria Própria

igual do que 157,5° então direção nominal é “sudeste”. Para finalizar a transformação dos dados, todos os atributos foram normalizados considerando apenas duas casas decimais após a vírgula.

Os comandos em linguagem SQL que foram utilizados para processamento e transformação nos dados são apresentados: no apêndice C para os modelos horários; no apêndice D para os modelos diários; no apêndice E para os modelos de previsão três dias à frente; e no apêndice F para os modelos de previsões semanais.

4.2.2 Construção dos modelos

De acordo com a abordagem proposta, três atividades foram realizadas na construção dos modelos: definição dos atributos, ajuste no algoritmo e validação.

No primeiro estudo de caso, o algoritmo M5p foi usado para definir o conjunto de entrada mais confiável. Entretanto, vimos que os resultados de precisão e performance variaram entre os algoritmos. Por isso, ao invés de determinar apenas um algoritmo em um único volume de dados, uma análise mais criteriosa para cada algoritmo (MLP, SVM, M5p e KNN) e cada intervalo de previsão (horário, diário, três dias e semanal) foi realizada. Em outras palavras, foi avaliado se o conjunto de entrada mais adequado pode variar entre os algoritmos e entre os diferentes volumes de dados usando o mesmo banco de dados.

Para validar o conjunto de entrada dos modelos horários foram considerados os sete atributos numéricos: mês (MS), numero da semana (NS), hora do dia (HD), pressão atmosférica (PA), umidade relativa (UR) e temperatura do ar (TP), velocidade (VV) e direção do vento (DV). Diferente do primeiro estudo, consideramos um novo atributo NS que foi gerado a partir de uma generalização com o objetivo de avaliar sua viabilidade nas previsões. Muitos experimentos foram realizados, porém, como no primeiro estudo de caso, os conjunto mais relevantes foram os que tiveram maiores números de atributos. Por isso, os conjuntos de entrada mais confiáveis para os modelos de previsão horária gerados através dos algoritmos MLP, M5p e KNN são apresentados na Figura 8.

Tabela 8 – Análise do conjunto de entrada mais confiável para previsão horária de uma turbina eólica através do coeficiente *R*

n°	**Entrada do modelo**	***R***		
		MLP	**M5p**	**KNN**
1	MS, TP, UR, PA, DV, NS	0,8356	0,8827	0,8971
2	HD, TP, UR, PA, DV, NS	0,7716	0,8496	0,8522
3	HD, MS, UR, PA, DV, NS	0,8345	0,8941	0,9026
4	HD, MS, TP, PA, DV, NS	0,8282	0,8959	0,9021
5	HD, MS, TP, UR, DV, NS	0,8346	0,8943	0,9030
6	HD, MS, TP, UR, PA, NS	0,8327	0,8768	0,8964
7	HD, MS, TP, UR, PA, DV	0,8380	0,8949	0,8998
8	HD, MS, TP, UR, PA, DV, NS	0,8353	0,8965	0,9112

Fonte: Autoria Própria

O SVM foi novamente inviável para previsão horária, pois teve um alto custo operacional levando cerca de 2754.28 segundos em sua execução para os 30640 registros horários. De fato, o custo operacional desse algoritmo tem-se apresentado alto para enormes conjuntos de dados.

O melhor conjunto para a NN MLP foi o n° 7 que combinou apenas seis variáveis e desprezou o atributo número da semana atingindo valores de R um pouco superior ao conjunto n° 8. M5p e KNN tiveram como melhor combinação de entrada o conjunto n° 8, com todos

os atributos disponíveis com *R* de 0,8965 e 0,9112, nesta ordem. Portanto, o novo atributo NS mostrou-se relevante para as previsões horárias usando os algoritmos M5p e KNN.

Em síntese, todos modelos de seis e sete variáveis, com exceção do modelo n° 2, tiveram valores correspondentes e poderiam ser usados para previsão horárias em razão dos altos valores de *R*. Eventualmente, forneceriam resultados de precisão relativamente inferiores devido aos valores mais baixos para *R*.

Nossa hipótese que a entrada para cada algoritmo poderia ser diferente para fornecer melhores resultados foi testada e validada. De fato, a entrada da MLP foi diferente do M5p e KNN: MLP combinou apenas seis atributos e teve a correção superior à sua combinação com sete. Isto significa que os algoritmos podem fornecer resultados relativamente melhores de performance e precisão em razão do maior *R* e menor número de atributos na entrada da construção do modelo.

Definida a entrada para cada algoritmo, uma série de experimentos foi realizada usando MLP, M5p e KNN. Os resultados dos modelos de previsão horária mais relevantes para cada categoria de algoritmo são apresentados na Tabela 9.

Tabela 9 – Resultados dos modelos relevantes para previsão horária de uma turbina eólica

n°	Algoritmo	Entrada	Configuração	MAE	RMSE	MSE	MAPE	SMAPE	Tempo(s)
1	MLP	HD, MS, TP, UR, PA, DV	RA = 0,3; TT = 1200; CO = aut	1,1581	1,4669	2,1520	16,14	4,07	23,95
2	MLP	HD, MS_n, TP, UR, PA, DV	RA = 0,3; TT = 1200; CO = aut	1,0443	1,3383	1,7912	14,43	3,68	105,41
3	MLP	HD, MS, TP, UR, PA, DV_n	RA = 0,3; TT = 1200; CO = aut	1,0504	1,3462	1,8125	14,56	3,71	69,85
4	MLP	HD, MS_n, TP, UR, PA, DV_n	RA = 0,3; TT = 1200; CO = aut	0,9959	1,2750	1,6256	13,77	3,52	176,59
5	M5p	HD, MS, TP, UR, PA, DV, NS	Instância mínima = 4	0,8809	1,1573	1,3395	12,41	3,16	2,63
6	M5p	HD, MS_n, TP, UR, PA, DV, NS	Instância mínima = 4	0,8827	1,1578	1,3405	12,52	3,17	3,77
7	M5p	HD, MS, TP, UR, PA, DV_n, NS	Instância mínima = 4	0,9188	1,1993	1,4384	12,89	3,29	3,06
8	M5p	HD, MS_n, TP, UR, PA, DV_n, NS	Instância mínima = 4	0,9204	1,1982	1,4358	12,92	3,29	4,36
9	KNN	HD, MS, TP, UR, PA, DV, NS	k = 4, KDTree	0,8446	1,1122	1,2371	11,74	3,02	1,04
10	KNN	HD, MS_n, TP, UR, PA, DV, NS	k = 4, KDTree	0,8426	1,1083	1,2284	11,74	3,01	1,04
11	KNN	HD, MS, TP, UR, PA, DV_n, NS	k = 6, KDTree	0,8749	1,1471	1,3159	12,28	3,14	1,06
12	KNN	HD, MS_n, TP, UR, PA, DV_n, NS	k = 6, KDTree	0,8735	1,1432	1,3074	12,29	3,13	1,06

Fonte: Autoria Própria

De modo geral, tivemos novamente excelentes resultados de performance em todos os modelos desenvolvidos com MLP, M5p e KNN (n° 1 ao 12). A MLP foi o algoritmo não tão rápido em relação aos demais. No entanto, em menos de três minutos conseguiu fazer uma previsão horária a partir de 30640 registros quando foi configurada com 0,3 para raio de aprendizagem, 1200 para tempo de treinamento e camadas ocultas definidas de forma automática. Por outro lado, M5p e KNN alcançaram uma execução em um tempo ainda mais inferior: em menos de 5 segundos foram capazes de construir modelos para previsão horária.

Os modelos gerados através da MLP alcançaram resultados relevantes. O modelo n° 4 (superior da categoria) quando configurado com um raio de aprendizagem 0,3, tempo de treinamento 1200 e camadas ocultas definidas de forma automática, alcançou 0,9959 para MAE 1,2750 para RMSE, 1,6256 para MSE, 13,77% para MAPE% e 3,52 para SMAPE. Nota-se que a rede neural MLP, como no primeiro estudo, forneceu os melhores resultados com dados no formato nominal.

O algoritmo M5p produziu quatro modelos relevantes. O modelo n° 5 foi o mais satisfatório da categoria que combinou apenas atributos numéricos e teve erros reduzidos como 0,8809 no MAE, 1,1573 no RMSE, 1,3395 no MSE, 12,41% no MAPE e 3,16% no SMAPE.

Embora M5p e MLP consigam alcançar bons resultados, o KNN foi superior atingindo uma maior precisão nos resultados dos modelos horários. Principalmente, o modelo n° 10 que combinou o atributo mês na forma nominal e analisou os 4 vizinhos próximos, atingindo 0,8426 para MAE, 1,1083 para RMSE, 1,2284 para MSE, 11,74% para MAE e 3,01% para SMAPE.

De forma resumida, numa análise de resultados dos modelos mais relevantes em termos de MAE: o KNN com 0,8426 foi superior ao M5p com 0,8809; MLP com 0,9959 teve resultado abaixo de ambos os algoritmos; SVM mostrou-se inviável mais uma vez devido ao alto custo operacional proporcionado em razão do grande volume de dados.

Além dos modelos horários, foram desenvolvidos os modelos para previsão diária. Em primeiro lugar, experimentos foram executados com a finalidade de definir os conjuntos de entrada mais relevantes através da análise de *R*, conforme apresentado na Tabela 10. O atributo "hora do dia" foi ignorado, pois os valores usados nos modelos diários foram a média diária dos atributos.

Tabela 10 – Análise do conjunto de entrada mais confiável para previsão diária de uma turbina eólica através de *R*

n°	**Entrada**	**R**			
		MLP	**SVM**	**M5P**	**KNN**
1	TP, UR, PA, DV, NS	0,7119	0,8198	0,7873	0,7862
2	MS, UR, PA, DV, NS	0,8251	0,8802	0,8686	0,8660
3	MS, TP, PA, DV, NS	0,8553	0,8811	0,8740	0,8654
4	MS, TP, UR, DV, NS	0,8195	0,8852	0,8783	0,8762
5	MS, TP, UR, PA, NS	0,8174	0,8844	0,8757	0,8688
6	MS, TP, UR, PA, DV	0,8169	0,8916	0,8726	0,8792
7	MS, TP, UR, PA, DV, NS	0,7941	0,8900	0,8746	0,8697

Fonte: Autoria Própria

Para os modelos diários, os maiores coeficientes não foram os que combinaram o maior número de atributos. Para MLP a entrada n° 3 com cinco atributos foi a mais relevante que ignorou a umidade relativa e alcançou 0,8553 de *R*. A entrada do SVM e do KNN (n° 6) que desprezou o atributo número da semana atingiu os mais altos valores para R: 0,8916 e 0,8782, respectivamente. Entretanto, M5p teve melhores valores quando ignorou a pressão do ar (n°4). De fato, essa análise da entrada do modelo mostrou-se eficiente, pois antes do desenvolvimento dos modelos, foi determinado o conjunto de entrada mais favorável para cada um dos algoritmos, de modo a alcançar um maior nível de precisão e performance nos resultados.

Assim que foram definidos os atributos de entrada para cada algoritmo, ajustou-se os parâmetros dos métodos para construção dos modelos diários. Mais uma vez, vários modelos foram construídos, os mais relevantes por categoria de algoritmo são apresentados na Tabela 11.

Tabela 11 – Resultados dos modelos relevantes para previsão diária de uma turbina eólica

n°	Algoritmo	Entrada	Configuração	MAE	RMSE	MSE	MAPE	SMAPE	Tempo(s)
1	MLP	MS, TP, PA, DV, NS	RA = 0,4; TT = 3000; CO = aut	0,7299	0,9600	0,9216	9,80	2,49	2,26
2	MLP	MS_n, TP, PA, DV, NS	RA = 0,4; TT = 3000; CO = aut	0,7365	0,9597	0,9211	9,88	2,51	9,34
3	MLP	MS, TP, PA, DV_n, NS	RA = 0,4; TT = 3000; CO = aut	0,7420	0,9696	0,9401	9,88	2,50	4,40
4	MLP	MS_n, TP, PA, DV_n, NS	RA = 0,4; TT = 3000; CO = aut	0,7212	0,9427	0,8887	9,64	2,44	13,10
5	M5p	MS, TP, UR, DV, NS	Instância mínima = 4	0,6859	0,8998	0,8097	9,22	2,34	0,09
6	M5p	MS_n, TP, UR, DV, NS	Instância mínima = 4	0,6971	0,9067	0,8221	9,38	2,38	0,15
7	M5p	MS, TP, UR, DV_n, NS	Instância mínima = 4	0,6845	0,8968	0,8043	9,21	2,35	0,10
8	M5p	MS_n, TP, UR, DV_n, NS	Instância mínima = 4	0,6901	0,9003	0,8106	9,25	2,36	0,13
9	SVM	MS, TP, UR, PA, DV	Kernel Puk	0,6474	0,8536	0,7286	8,64	2,20	1,06
10	SVM	MS_n, TP, UR, PA, DV	Kernel Puk	0,6778	0,8915	0,7949	9,15	2,33	1,33
11	SVM	MS, TP, UR, PA, DV_n	Kernel Puk	0,6740	0,8875	0,7877	9,08	2,30	1,06
12	SVM	MS_n, TP, UR, PA, DV_n	Kernel Puk	0,6811	0,8974	0,8053	9,24	2,34	1,51
13	KNN	MS, TP, UR, PA, DV	k = 14, KDTree	0,6858	0,9042	0,8175	9,05	2,34	1,01
14	KNN	MS_n, TP, UR, PA, DV	k = 11, KDTree	0,6848	0,9067	0,8221	9,04	2,34	1,01
15	KNN	MS, TP, UR, PA, DV_n	k = 12, KDTree	0,6804	0,8952	0,8015	9,06	2,33	1,00
16	KNN	MS_n, TP, UR, PA, DV_n	k = 8, KDTree	0,6883	0,8964	0,8036	9,16	2,35	1,00

Fonte: Autoria Própria

Em síntese, todos os modelos para previsão diária do n° 1 ao 16 alcançaram bons resultados, sobretudo, a performance que foi favorável na qual 13,10 segundos foi o maior tempo para execução dos algoritmos.

Os modelos gerados através da MLP tiveram resultados equivalentes e relevantes (n ° 1 ao 4). Sobretudo o n° 4 que foi o mais relevante dentre os modelos gerados via MLP, combinou novamente mês e direção do vento na forma nominal para atingir 0,7212 para MAE, 0,9427 para RMSE, 0,8887 para MSE, 9,64% para MAPE e 2,44% para SMAPE. Apenas foram alcançados esses resultados quando definiu-se 0,4 para raio de aprendizagem e 3000 para tempo de treinamento na configuração de parâmetros da MLP.

Embora os resultados da MLP sejam relevantes, M5p conseguiu superá-los. Inclusive, o modelo n° 7 alcançou os melhores resultados de sua categoria com 0,6845 no MAE, 0,8968 no RMSE, 0,8043 no MSE, 9,21% no MAPE e 2,35% no SMAPE. Esse modelo, ao contrário dos modelos horários, teve apenas a direção do vento expresso no formato nominal e os demais atributos foram numéricos.

O modelo horário mais relevante desse estudo foi o n° 9, construído a partir do SVM. Esse modelo teve a direção expressa no formato nominal e foi configurado com o Kernel Puk para atingir 0,6474 para MAE, 0,8536 para RMSE, 0,7286 para MSE, 8,64% para MAPE e 2,20% para SMAPE.

Como a MLP, os modelos gerados através do KNN tiveram resultados equivalentes e relevantes (n ° 13 ao 16). O modelo mais convecedor foi o n ° 15 que foi construído através da análise de 12 vizinhos próximos e um algoritmo de busca em árvore, atingindo 0,6804 para MAE, 0,8952 para RMSE, 0,8015 para MSE, 9,06% para MAE, 2,33% para SMAPE. Dessa vez, a melhor combinação do KNN teve a direção expressa no formato nominal.

No geral, SVM, M5p e KNN tiveram resultados correspondentes, enquanto que a rede neural MLP forneceu resultados um pouco abaixo desses modelos em precisão. Na análise

individual de MAE, os melhores modelos gerados com 1301 registros de dados diários de uma turbina eólica: SVM com 0,6474 para MAE foi superior ao KNN com 0,6804, que foi levemente superior ao M5p com 0,6845; MLP ficou abaixo dos demais modelos com 0,7212 para MAE .

Nossa proposta mais um vez mostrou-se relevante para previsão horária e diária (ou estimativas de prazo ultra curto e curto). Em sumo, foi proposto neste estudo executar nossa abordagem para construir modelos para previsão três dias à frente e semanal (ou de prazo médio e longo). Desse modo, uma análise de *R* foi realizada para definir o conjunto mais relevante para cada algoritmo, conforme apresentado na Tabela 12.

Tabela 12 – Análise do conjunto de entrada mais confiável para previsão a cada três dias através coeficiente *R*

n°	**Entrada**	**R**			
		MLP	**SVM**	**M5P**	**KNN**
1	TP, UR, PA, DV, NS	0,7389	0,8316	0,8133	0,7772
2	MS, UR, PA, DV, NS	0,7835	0,8872	0,8767	0,8694
3	MS, TP, PA, DV, NS	0,8082	0,8863	0,8789	0,8493
4	MS, TP, UR, DV, NS	0,8015	0,8935	0,8819	0,8655
5	MS, TP, UR, PA, NS	0,7907	0,8917	0,8810	0,8604
6	MS, TP, UR, PA, DV	0,8019	0,9030	0,8818	0,8897
7	MS, TP, UR, PA, DV, NS	0,8037	0,9016	0,8810	0,8681

Fonte: Autoria Própria

Por analogia à análise dos conjuntos diários, nenhum dos algoritmos que combinou todos os atributos disponíveis tiveram maiores valores de *R* para previsão três dias à frente. Isto significa que o atributo em excesso, se considerado no modelo, teria reduzido os resultados de performance e precisão. O conjunto mais relevante para MLP foi o n ° 3 que ignorou a umidade relativa dentre os atributos disponíveis atingindo *R* de 0,8082. Em oposição, o conjunto mais promissor para SVM e M5p foi o n° 6 que ignorou o atributo número da semana atingindo 0,9030 e 0,8897, nesta ordem. Já o n° 4 apresentou-se como um conjunto mais favorável para M5p atingindo valor de 0,8819 para *R*.

Definidos os melhores conjuntos numéricos para cada algoritmo, é hora de ajustar os algoritmos, incrementar dados nominais e validar os modelos. Diversos combinações foram testadas para a base de dados. Os resultados das previsões mais relevantes para os próximos três dias são apresentados na Tabela 13.

Todos os modelos apresentaram uma performance muito favorável e em menos de 5 segundos os algoritmos executaram previsões da velocidade do vento para os próximos três dias. Isso se deu em razão do conjunto de dados que foi reduzido na etapa de transformação para as previsão a cada três dias (516 registros).

Os modelos gerados através da MLP alcançaram bons resultados quando configurados com 0,6 no raio de aprendizagem, 3000 no tempo de treinamento e camadas ocultas definidas de forma automática. Pela primeira vez, a MLP forneceu os melhores resultados quando ignorou

Tabela 13 – Resultados dos modelos para previsão de uma turbina eólica três dias à frente

n°	Algoritmo	Entrada	Configuração	MAE	RMSE	MSE	MAPE	SMAPE	Tempo(s)
1	MLP	MS, TP, PA, DV, NS	RA = 0,6; TT = 3000; CO = aut	0,6570	0,8796	0,7737	8,71	2,20	0,91
2	MLP	MS_n, TP, PA, DV, NS	RA = 0,6; TT = 3000; CO = aut	0,6773	0,8847	0,7828	9,02	2,28	3,69
3	MLP	MS, TP, PA, DV_n, NS	RA = 0,6; TT = 3000; CO = aut	0,7903	1,0363	1,0740	10,62	2,68	1,36
4	MLP	MS_n, TP, PA, DV_n, NS	RA = 0,6; TT = 3000; CO = aut	0,6608	0,8703	0,7574	9,18	2,26	5,04
5	M5p	MS, TP, UR, DV, NS	Instância mínima = 4	0,6482	0,8593	0,7385	8,65	2,19	0,16
6	M5p	MS_n, TP, UR, DV, NS	Instância mínima = 4	0,6492	0,8598	0,7392	8,61	2,19	0,16
7	M5p	MS, TP, UR, DV_n, NS	Instância mínima = 4	0,6414	0,8533	0,7282	8,58	2,17	0,04
8	M5p	MS_n, TP, UR, DV_n, NS	Instância mínima = 4	0,6504	0,8496	0,7218	8,67	2,20	0,06
9	SVM	MS, TP, UR, PA, DV	Kernel Puk	0,6075	0,7848	0,6160	8,05	2,06	0,16
10	SVM	MS_n, TP, UR, PA, DV	Kernel Puk	0,6713	0,8554	0,7317	8,92	2,27	0,16
11	SVM	MS, TP, UR, PA, DV_n	Kernel Puk	0,6317	0,8047	0,6476	8,35	2,12	0,12
12	SVM	MS_n, TP, UR, PA, DV_n	Kernel Puk	0,6555	0,8544	0,7301	8,71	2,21	0,19
13	KNN	MS, TP, UR, PA, DV	k = 9, KDTree	0,6529	0,8417	0,7085	8,53	2,20	1,00
14	KNN	MS_n, TP, UR, PA, DV	k = 5, KDTree	0,6783	0,8899	0,7919	8,89	2,30	1,00
15	KNN	MS, TP, UR, PA, DV_n	k = 12, KDTree	0,6599	0,8449	0,7139	8,74	2,24	1,00
16	KNN	MS_n, TP, UR, PA, DV_n	k = 5, KDTree	0,6783	0,8850	0,7832	8,96	2,30	1,00

Fonte: Autoria Própria

a combinação de direção do vento e mês no formato nominal. Provavelmente, em função do menor conjunto de dados que foi utilizado para treinamento desse modelo. O modelo n° 1 usou apenas dados numéricos e alcançou os melhores resultados em sua categoria, tais como 0,6570 para MAE, 0,8796 para RMSE, 0,7737 para MSE, 8,71% para MAPE e 2,20% para SMAPE. Contudo, esses resultados foram apenas um pouco superiores ao modelo (n° 4) que teve mês e direção no formato nominal e atingiu 0,6608 para MAE, 0,8703 para RMSE, 0,7737 para MSE, 0,7574, 9,18% para MAPE e 2,26% para SMAPE. O RMSE do modelo com a combinação mês e direção do vento ainda conseguiu ser inferior em termos de RMSE e MSE. Isto significa que o modelo n° 4, gerado com dados mês e direção do vento na forma nominal, teve um maior número na média absoluta de erros porém, foram imprecisões menos discrepantes.

O modelo n° 7 gerado via M5p foi o que alcançou um dos melhores resultados gerais que foram levemente superiores aos demais modelos da categoria M5p com 0,6414 para MAE, 0,8533 para RMSE, 0,7282 para MSE, 8,58% para MAPE e 2,17% para SMAPE. Esse modelo teve a combinação da direção do vento na forma nominal, o que é análogo ao modelo soberano de previsão diária na categoria M5p. Nota-se que quando o conjunto de dados é reduzido, M5p fornece melhores resultados com atributos nominais em seus modelos. Por outro lado, em conjuntos mais amplos, os dados no formato numérico apresentaram melhores resultados.

O modelo mais relevante foi o n° 9, construído através do SVM com o Kernel Puk, combinou apenas atributos numéricos para alcançar 0,6075 para MAE, 0,7848 para RMSE, 0,6160 para MSE, 8,05% para MAPE e 2,06% para SMAPE. Essa combinação de entrada foi a mesma usada no modelo diário mais satisfatório construído através do SVM.

O algoritmo KNN também apresentou resultados relevantes que foram superiores à MLP. O modelo n° 13, que foi configurado com 13 partições e um algoritmo de busca em árvore, forneceu os melhores resultados da categoria, tais como 0,6529 para MAE, 0,8417 para RMSE, 0,7085 para MSE, 8,53% para MAPE e 2,20% para SMAPE. Dessa vez, a melhor combinação do KNN teve todos os atributos numéricos.

Os resultados ainda poderiam ser melhores se uma maior quantidade de dados fosse utilizada para construção dos modelos nas previsão a cada três dias à frente. No entanto, bons resultados foram alcançados em todos os modelos.

Na comparação individual através de MAE dos melhores modelos por categoria: SVM com 0,6075 foi superior ao M5p que alcançou 0,6414; No entanto, M5p teve uma ligeira vantagem sobre MLP e KNN que tiveram a média de erros absolutos: 0,6570 e 0,6529, respectivamente.

Neste ponto, propõe-se a construção dos modelos de previsão semanal. Segundo a abordagem proposta neste trabalho três etapas são executadas: definição dos atributos de entrada, ajuste do algoritmo e validação. Na Tabela 14 são apresentadas as combinações mais relevantes dos modelos semanais para cada algoritmo. Dessa vez, a MLP teve todos os atributos disponíveis como conjunto de entrada mais favorável, atingindo 0,8442 para *R*. Tanto SVM como KNN tiveram a combinação n° 6 a mais relevante com *R* de 0,9146 e 0,9003, na devida ordem. Por outro lado, a combinação n° 4 foi a mais confiável para M5p que desprezou a pressão atmosférica para atingir 0,8917 de *R*.

Tabela 14 – Análise do conjunto de entrada mais confiável para previsão semanal através do coeficiente *R*

n°	**Entrada**	**R**			
		MLP	**SVM**	**M5P**	**KNN**
1	TP, UR, PA, DV, NS	0,7846	0,8192	0,8050	0,7433
2	MS, UR, PA, DV, NS	0,8157	0,8984	0,8820	0,8593
3	MS, TP, PA, DV, NS	0,7666	0,9032	0,8831	0,8419
4	MS, TP, UR, DV, NS	0,8434	0,8920	0,8917	0,8540
5	MS, TP, UR, PA, NS	0,7948	0,8952	0,8851	0,8450
6	MS, TP, UR, PA, DV	0,8405	0,9146	0,8890	0,9003
7	MS, TP, UR, PA, DV, NS	0,8442	0,9003	0,8878	0,8588

Fonte: Autoria Própria

Os resultados dos modelos mais relevantes para previsão semanal são apresentados na Tabela 14. Embora a quantidade de dados tenha sido também desfavorável, alcançou-se excelentes resultados nos modelos de previsão semanal. Em menos de 2 segundos os algoritmos executaram a partir dos 174 registros semanais.

O modelo soberano da categoria MLP foi o n° 4 que teve mais uma vez o mês e a direção do vento no formato nominal atingindo 0,5573 para MAE, 0,7339 para RMSE, 0,5386 para MSE, 7,42% para MAPE e 1,86% para SMAPE. Esse modelo foi configurado com um raio de aprendizagem 0,6, tempo de treinamento 2000 e camadas ocultas definidas automaticamente através de experimentos que minimizaram o erro.

O modelo n° 6, gerado através do M5p que combinou mês no formato nominal, foi o melhor modelo da categoria com 0,5546 para MAE, 0,7491 para RMSE, 0,5612 para MSE, 7,47% para MAPE e 1,88% para SMAPE.

O modelo mais relevante para previsão semanal foi o n° 9 que combinou apenas atributos numéricos e alcançou os melhores resultados tais como 0,5113 para MAE, 0,6946 para RMSE, 0,4825 para MSE, 6,81% para MAPE e 1,72% para SMAPE. Esse modelo foi construído a partir da técnica SVM configurada com o Kernel Puk.

Para os modelos semanais, o KNN ficou abaixo dos demais algoritmos, talvez em função da baixa quantidade de dados. De fato, o algoritmo mais relevante para previsão semanal dessa categoria foi o n° 13, que configurado com 6 partições e um algoritmo de busca em árvore atingiu 0,5900 para MAE, 0,7511 para RMSE, 0,5642 para MSE, 8,83% para MAPE e 1,99% para SMAPE.

No geral, em termos de resultados de precisão na análise individual de MAE: SVM com 0,5113 foi moderadamente superior ao M5p e MLP que alcançaram resultados equivalentes 0,5546 e 0,5573, respectivamente; já o KNN que atingiu 0,5900 e foi relativamente inferior aos demais modelos. Certamente, esses resultados foram relevantes mas poderiam ser superiores se uma maior série temporal de dados fosse considerada.

Tabela 15 – Resultados dos modelos para previsão semanal de uma turbina eólica

n°	Algoritmo	Entrada	Configuração	MAE	RMSE	MSE	MAPE	SMAPE	Tempo(s)
1	MLP	MS, TP, UR, PA, DV, NS	RA = 0,3; TT = 500; CO = aut	0,6786	0,8725	0,7612	8,92	2,24	0,07
2	MLP	MS_n, TP, UR, PA, DV, NS	RA = 0,6; TT = 2000; CO = aut	0,5699	0,7508	0,5637	7,64	1,91	1,01
3	MLP	MS, TP, UR, PA, DV_n, NS	RA = 0,6; TT = 2000; CO = aut	0,7919	0,9835	0,9673	10,40	2,60	0,31
4	MLP	MS_n, TP, UR, PA, DV_n, NS	RA = 0,6; TT = 2000; CO = aut	0,5573	0,7339	0,5386	7,42	1,86	1,20
5	M5p	MS, TP, UR, DV, NS	Instância mínima = 4	0,5964	0,7747	0,6001	8,01	2,01	0,03
6	M5p	MS_n, TP, UR, DV, NS	Instância mínima = 4	0,5546	0,7491	0,5612	7,47	1,88	0,03
7	M5p	MS, TP, UR, DV_n, NS	Instância mínima = 4	0,6178	0,8010	0,6416	8,32	2,09	0,03
8	M5p	MS_n, TP, UR, DV_n, NS	Instância mínima = 4	0,5651	0,7651	0,5855	7,60	1,91	0,03
9	SVM	MS, TP, UR, PA, DV	Kernel Puk	0,5113	0,6946	0,4825	6,81	1,72	0,04
10	SVM	MS_n, TP, UR, PA, DV	PolyKernel	0,5808	0,7621	0,5808	7,96	2,00	0,03
11	SVM	MS, TP, UR, PA, DV_n	Kernel Puk	0,5584	0,7263	0,5275	7,23	1,82	0,03
12	SVM	MS_n, TP, UR, PA, DV_n	PolyKernel	0,5921	0,7577	0,5742	8,08	2,01	0,03
13	KNN	MS, TP, UR, PA, DV	k = 6, KDTree	0,5900	0,7511	0,5642	8,83	1,99	0,03
14	KNN	MS_n, TP, UR, PA, DV	k = 10, KDTree	0,6233	0,8191	0,6710	8,39	2,13	0,03
15	KNN	MS, TP, UR, PA, DV_n	k = 4, KDTree	0,6311	0,8162	0,6662	8,44	2,13	0,03
16	KNN	MS_n, TP, UR, PA, DV_n	k = 9, KDTree	0,6245	0,7948	0,6317	8,43	2,12	0,03

Fonte: Autoria Própria

4.3 RESULTADOS E DISCUSSÃO

Esta abordagem mostrou-se relevante para desenvolvimento de modelos de prazo ultra curto, curto, médio e longo. No geral, obtivemos excelentes resultados em todos os seis modelos construídos nos dois estudos de casos realizados.

Comparar e definir um algoritmo geral para previsão de velocidade do vento foi desafiador pois são muitos fatores que devem ser levado em consideração. Além do mais foram usados quatro algoritmos robustos que podem fornecer excelentes resultados em uma dada situação. Definiu-se três critérios para determinar o melhor algoritmo: (i) tempo de execução, o tempo

dado em segundos levado para execução do algoritmo até o fornecimento do modelo de previsão; (ii) resultados de precisão, o resultado mais baixo para a métrica MAE; (iii) por fim, a quantidade de parâmetros (ou tempo levado) para um ajuste confiável no algoritmo.

MLP, M5p e KNN foram algoritmos relevantes que tiveram performance significativas na execução de enormes conjuntos de dados. Em contraste, SVM mostrou-se inviável nos dois primeiros estudo para construção de modelos de previsão horária a partir de enormes conjuntos de dados. Nos conjunto de dados menores, SVM teve um tempo de execução compatível com os demais modelos. Tomando por base os dois experimentos, foi classificada a performance na seguinte ordem: M5p, KNN, MLP e SVM.

Nos resultados de precisão, foram desenvolvidos seis modelos: dois no primeiro estudo e quatro no segundo estudo. SVM mostrou soberania e ótima capacidade de generalização em quatro dos modelos de previsão construídos a partir de 2595, 1301, 515 e 174 registros. Em contrapartida, KNN teve resultados superiores nos dois outros modelos para bases de dados maiores: 61934 e 30640 registros. M5p ficou em segundo lugar na precisão de cinco dos seis modelos de previsão construídos. Já a MLP ficou abaixo dos demais algoritmos na maioria dos modelos. Baseado nos dois estudo de casos, classificamos a precisão dos modelos na seguinte ordem: SVM, M5p, KNN e MLP.

Em relação à quantidade de parâmetros, SVM demanda a configuração do Kernel o que influência diretamente os resultados. M5p necessita de ajuste no limiar das folhas pois limita a estrutura do modelo de saída da árvore. KNN tem um custo moderado pois é necessário encontrar o valor exato para K, além de definir um algoritmo de busca relevante. Uma desvantagem da MLP é o alto número de parâmetros de configuração antes da execução, o que demanda um tempo significativo para um ajuste favorável. Sendo assim, classificamos o ajuste dos parâmetros dos modelos na seguinte ordem: M5p, SVM, KNN e MLP.

Tomando por base os três critérios estabelecidos, o M5p foi o melhor algoritmo para previsão da velocidade do vento, pois teve uma configuração simples, foi o mais rápido e forneceu resultados relevantes de precisão em todos os modelos construídos. Contudo, o especialista pode testar o KNN para conjuntos enormes de dados e SVM em conjuntos reduzidos com o objetivo de alcançar resultados relativamente melhores.

Os resultados da distribuição dos seis modelos mais relevantes são apresentados na Figura 13. O valor atual e o valor predito da velocidade do vento podem ser analisados. No geral, pode-se assumir numa análise qualitativa que os seis modelos construídos nos dois estudos de casos tiveram resultados relevantes, principalmente, os resultados que são apresentados nos itens (B), (D), (E) e (F) que mostraram contínua a proximidade entre valores análogos. De fato, se for considerado que há uma boa série de dados disponíveis para fazer previsões de longo prazo, as previsões a cada uma hora tende a ser mais difícil que as diárias, que tende a ser mais complexa do que as de 3 dias, e assim por diante.

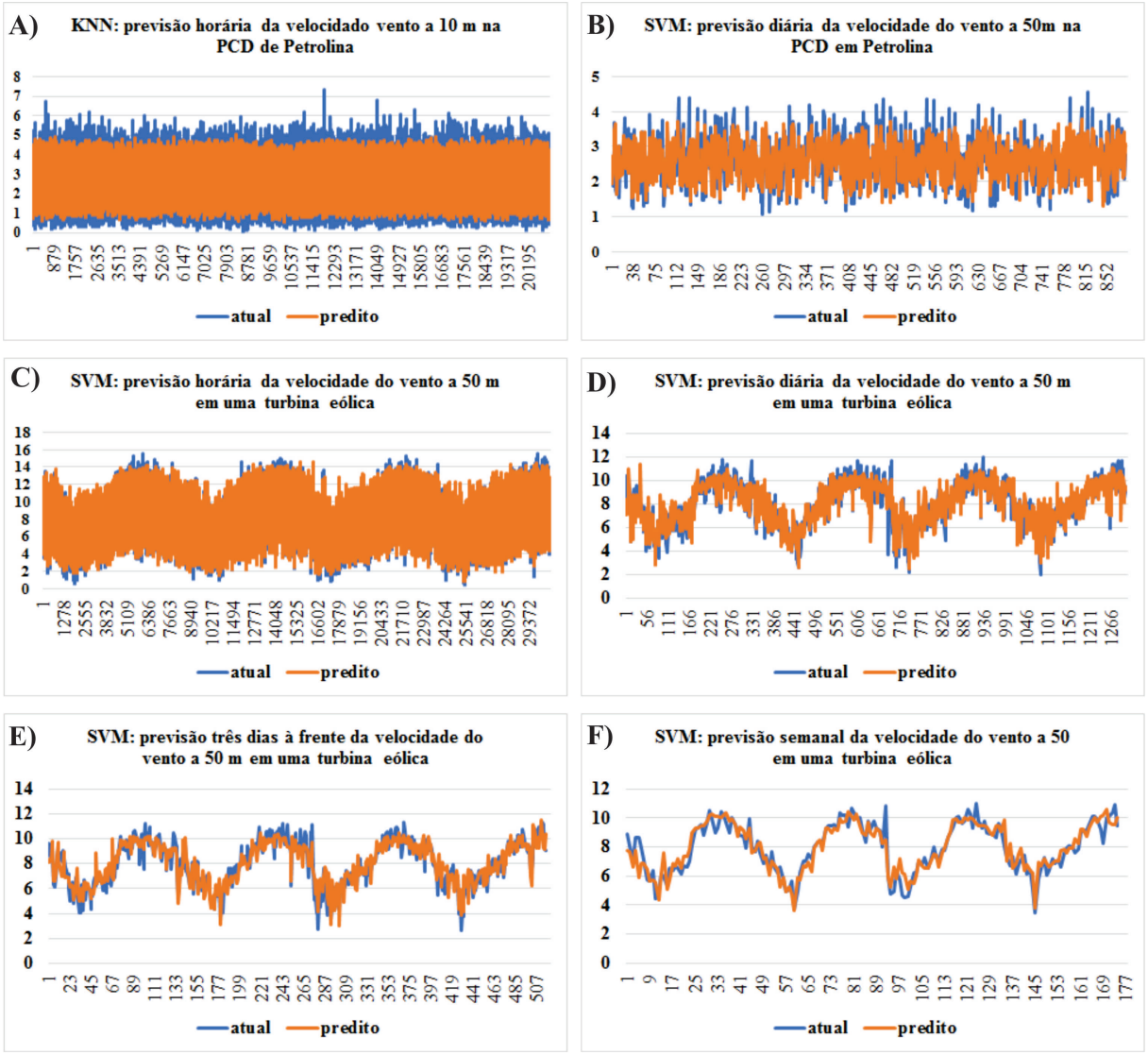

Figura 13 – Resultados dos seis modelos mais relevantes construídos nos dois estudos de casos

Fonte: Autoria Própria

A comparação de modelos para previsão do vento é complexa pois os resultados são influenciados por muitos fatores, como volume de dados na validação, projeções de sinais, critério de validação, conjunto de entrada disponível, amplitude da variação da velocidade do vento, objetivo da previsão do modelo, altura do instrumento de medição, aspectos físicos do terreno e configuração exata do algoritmo. Portanto, assumir que um modelo é melhor do que o outro é desafiador em razão do número de situações que devem ser levados em consideração.

No entanto, discuti-se adiante uma comparação geral dos seis modelos mais relevantes elaborados através desta abordagem com alguns modelos de destaques extraídos de trabalhos na literatura, conforme apresentado na Tabela 16.

Analisando todas as métricas estatísticas dos modelos de prazo curto do n° 1 ao 16, pode-se observar que nosso modelo n° 7 apresentou-se com valores menores na maioria das métricas estatísticas em relação aos modelos de previsão horária da velocidade do vento, tais

Tabela 16 – Uma comparação geral de resultados de modelos para previsão da velocidade do vento com diferentes intervalos de previsão e algoritmos

n°	Referência do modelo	IP	Algoritmo	MAE	RMSE	MSE	MAPE	SMAPE
1	(PINTO et al., 2015)	5m	SVM	0,7120	-	-	22,87	21,17
2	(YESILBUDAK; SAGIROGLU; COLAK, 2013)	10m	KNN	0,7400	-	-	7,08	-
3	(YESILBUDAK; SAGIROGLU; COLAK, 2013)	10m	KNN	0,8210	-	-	7,71	-
4	(YESILBUDAK; SAGIROGLU; COLAK, 2013)	10m	KNN	0,8000	-	-	7,48	-
5	(YESILBUDAK; SAGIROGLU; COLAK, 2013)	10m	KNN	1,0130	-	-	9,52	-
6	(LIU et al., 2014)	30m	SVM	0,7843	1,2125	-	17,80	-
7	PCD de Petrolina	1h	KNN	0,5495	0,7046	0,4965	23,03	6,05
8	Turbina Eólica	1h	KNN	0,8426	1,1083	1,2284	11,74	3,01
9	(LAHOUAR; SLAMA, 2014)	1h	SVM	0,8363	1,1800	-	-	-
10	(SALCEDO-SANZ et al., 2011)	1h	SVM	1,7823	-	-	-	-
11	(HU et al., 2016)	1h	NN	0,9305	1,2382	-	16,72	-
12	(HU et al., 2016)	1h	NN	0,9319	1,2435	-	14,95	-
13	(HU et al., 2016)	1h	RL	0,9267	1,2359	-	15,93	-
14	(HU et al., 2016)	1h	NN	0,9725	1,2685	-	26,49	-
15	(HU et al., 2016)	1h	NN	1,0037	1,2905	-	26,81	-
16	(HU et al., 2016)	1h	RL	0,9735	1,2662	-	26,26	-
17	PCD de Petrolina	1d	SVM	0,3406	0,4368	0,1908	13,35	3,34
18	Turbina Eólica	1d	SVM	0,6474	0,8536	0,7286	8,64	2,20
19	(RAMOS et al., 2011b)	1d	NN	1,9770	-	-	-	-
20	(FINAMORE et al., 2015)	1d	NN	-	-	3,1500	-	-
21	(FINAMORE et al., 2015)	1d	NN	-	-	3,4500	-	-
22	(DARAEEPOUR; ECHEVERRI, 2014)	1d	NN	0,9789	1,2984	-	-	-
23	(DARAEEPOUR; ECHEVERRI, 2014)	1d	NN	1,5864	2,2126	-	-	-
24	(MOHANDES; REHMAN; HALAWANI, 1998)	1d	NN	-	1,2400	-	-	-
25	Turbina Eólica	3d	SVM	0,6074	0,7848	0,6160	8,05	2,06
26	Turbina Eólica	7d	SVM	0,5113	0,6946	0,4825	6,81	1,72
27	(MOHANDES; REHMAN; HALAWANI, 1998)	30d	NN	-	1,8700	-	-	-
28	(MALIK; SAVITA, 2016)	30d	NN	-	-	0,4800	22,80	-

Fonte: Autoria Própria

como 0,5495 para MAE, 0,7046 para RMSE, 0,4965 para MSE. Este modelo foi construído usando dados a 10 metros do solo de uma PCD e validando os modelos com uma base de dados ampla. O modelo n° 7 foi superado somente na métrica MAPE em relação a alguns dos modelos. O MAPE não é uma métrica recomendada para comparação entre modelos com base de dados de volumes distintos, pois calcula um erro desbalanceado em função do volume de dados. Como o estudo foi realizado com 7 anos de dados e 1/3 foi usado na validação, o MAPE tendeu a crescer desproporcionalmente. Isto pode ser evidenciado se for analisado o SMAPE deste modelo, o qual alcançou 6,05%, ou se for visualizado o item A) da Figura 13 onde pode-se observar que os erros absolutos foram predominantemente positivos, ou seja, o valor predito foi menor do que o valor atual.

O modelo n° 8 do segundo estudo de caso com dados de uma turbina a 50 metros teve resultados extremamente relevantes tais como 0,8426 para MAE, 1,1083 para RMSE e 1,2284 para MSE. Esse modelo foi validado através da técnica *Cross-validation* e alcançou 0,8426 para MAE, 1,1083 para RMSE, 1,2284 para MSE, 11,74% para MAPE e 3,01% para SMAPE. O MAE, RMSE e MSE foram superiores no modelo n° 8 devido à maior amplitude da velocidade do vento a 50 metros. Se for analisado SMAPE, uma métrica que calcula os erros de forma balanceada em função da base de dados, observa-se que o modelo n° 8 foi superior ao modelo n°

7. De fato, a previsão da velocidade do vento tende a ser diretamente proporcional à altura da medição devido à maior influência de obstáculos (ou maior rugosidade do terreno) em função da proximidade do solo. Por isso, o modelo de previsão horária do segundo estudo foi superior ao primeiro.

Do mesmo modo, analisando os modelos de previsão de prazo curto, observou-se que os modelos para previsão diária n° 17 e 18 construídos no primeiro e segundo estudo tiveram valores menores para MAE, RMSE e MSE do que os modelos diários da literatura do n° 19 ao 25. Muitas vezes com valores bem distantes como no modelo n° 19 que alcançou apenas 1,9770 para MAE e nos modelos n° 20 e 21 que tiveram somente 3,1500 e 3,4500, respectivamente.

Em relação a comparação dos modelos construídos nesta proposta, novamente o modelo para previsão diária na turbina eólica se mostrou superior com 2,20% de SMAPE ao modelo da PCD de Petrolina que atingiu 3,34%. Salientar-se que os dados das PCDs são de alturas inferiores e sofrem com variações de sinais.

Os modelos n° 26 e 27 também mostraram resultados relevantes. Em razão da inexistência de modelos de previsão semanal e de três dias à frente, fizemos uma comparação com um modelo mensal da literatura. De fato, nosso modelo mostrou resultados superiores em termos de RMSE e equivalentes em termos de MSE com dois modelos mensais, n° 28 e 29, mesmo em uma comparação com um modelo de previsão semanal.

A maioria dos trabalhos da literatura que foram apresentados na Tabela 16, executam apenas um algoritmo em um repositório de dados, sem considerar os aspectos que são capazes de aprimorar a precisão e performance dos modelos. Em outras palavras, abordagens foram implementadas para um conjunto de dados específico sem qualquer processamento e checagem destes e somente resultados de precisão foram comparados para determinar o método mais confiável para aquela situação.

Certamente, os excelentes resultados dos modelos para previsão da PCD em Petrolina e da turbina eólica se deram devido ao conjunto de etapas eficientes e estratégicas que contemplam a abordagem proposta. Os modelos iniciais construídos tiveram resultados muito inferiores aos valores mais relevantes que alcançamos. Os modelos construídos atravessaram evoluções e aprimoramentos constantes proporcionando melhorias significativas nos resultados de precisão e performance. A checagem nos dados permitiu identificar atributos suspeitos que foram corrigidos ou ignorados. Os dados atravessaram um processo de agregação que aumentou a performance na execução dos algoritmos. A generalização dos dados nominais, que foi determinada a partir de atributos numéricos, aumentou o nível dos resultados na maioria dos modelos. O atributo número da semana extraído também de uma generalização, forneceu melhorias significativas nos resultados na maioria dos modelos. Tanto a agregação como a normalização aumentaram os resultados de performance dos modelos, pois permitiram a redução do conjunto de dados e otimizaram os cálculos complexos no treinamento dos modelos. Na construção do mesmo, o conjunto de entrada foi ajustado através de uma tática exaustiva, o que definiu o melhor conjunto

para cada algoritmo, proporcionando melhorias nos resultados. Por conseguinte, o ajuste nos parâmetros permitiu alcançar os mais favoráveis níveis de performance e precisão nos resultados. Além do mais, os excelentes resultados foram alcançados através de abordagens confiáveis e conhecidas na literatura: a estratégia 2/3 treinamento e 1/3 validação, assim como a abordagem *Cross-Validation*.

5 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Os operadores de energia enfrentam dificuldades para previsão da velocidade do vento devido à carência de recursos tecnológicos que podem otimizar as atividades operacionais nas indústrias eólicas. A quantidade significativa de informações armazenadas com o passar do tempo nos bancos de dados é um outro fator que deve ser considerado, tendo em vista que ferramentas devem dar suporte aos enormes repositórios existentes. Através de pesquisas bibliográficas foi detectado que uma quantidade significativa de modelos de previsão da velocidade são construídos, porém sem considerar aspectos relevantes que podem fornecer melhorias significativas nos resultados de precisão e performance.

Diante do exposto, nesta pesquisa é proposta uma abordagem de mineração de dados flexível e de baixo custo que pode ser utilizada pelas indústrias de energia para gerar modelos de previsão da velocidade do vento em diversos intervalos e a partir de grandes bancos de dados. A abordagem foi construída com base em fundamentos de mineração de dados e atravessou uma série de evoluções a partir de experimentos com dados reais.

Dois estudos foram realizados para validar a abordagem proposta: no primeiro estudo de caso com dados de velocidade do vento a 10m, surgiu uma dificuldade na construção dos modelos de previsões da velocidade do vento, em razão das inconsistências de dados da plataforma de coleta de dados de Petrolina. No entanto, com a abordagem considerando aspectos eficientes de processamento e fornecendo novos atributos generalizados, alcançamos resultados de precisão e performance importantes nos dois modelos de previsão construídos; no segundo estudo com dados de uma turbina eólica a 50m, provamos novamente que os modelos podem fornecer resultados mais relevantes se dados nominais fossem considerados ao invés de numéricos. Também mostramos que o conjunto de entrada para cada modelo varia de acordo com o algoritmo selecionado. Além disso, mostramos que os três novos atributos generalizados (mês, direção nominal e número da semana na forma numérica) e as agregações realizados promoveram melhorias significativas nos resultados de precisão e performance dos quatro modelos desenvolvidos.

Com base na literatura e em nossos experimentos, constatamos que não existe um único e melhor algoritmo global para previsão da velocidade do vento que possa ser aplicado em qualquer situação para fornecer os melhores resultados, devido ao fato dos padrões de ventos serem influenciados por muitos fatores. No entanto, existem algoritmos robustos contemplados em nossa abordagem que podem ser aplicados em uma dada situação para fornecer resultados relevantes, tais como a rede neural MLP, a estratégia de máquina de vetores de suporte e os algoritmos de árvore de decisão e K-vizinhos mais próximos.

Em relação às contribuições científicas deste trabalho, um projeto foi submetido e

aprovado, parcerias foram estabelecidas com a FUNCEME e com empresas de energia eólica, um artigo foi publicado pelo *International Journal of Engineering Research and Applications* e pode ser acessado em (FREITAS; SILVA; SAKAMOTO, 2018); duas outras propostas de artigos foram submetidas ao *Journal of Information Sciences* e ao *Journal of Renewable Energy*.

Algumas limitações foram observadas nesta proposta: (1) embora a abordagem de mineração de dados incorpore recursos relevantes para as previsões da velocidade do vento, os resultados estão diretamente relacionados à expertise do especialista; ele é quem deve compreender os dados e determinar as formas necessárias para processamento e transformação, bem como os algoritmos e suas configurações na construção dos modelos; (2) as constantes iterações envolvidas na abordagem podem demandar um tempo significativo para construir um modelo inicial em determinada região, principalmente em bancos de dados inconsistentes; (3) devido à complexidade da previsão da velocidade do vento, bases de dados que não atinjam um patamar de qualidade adequado comprometerão os resultados.

Possíveis aperfeiçoamentos e trabalhos futuros desta proposta incluem:

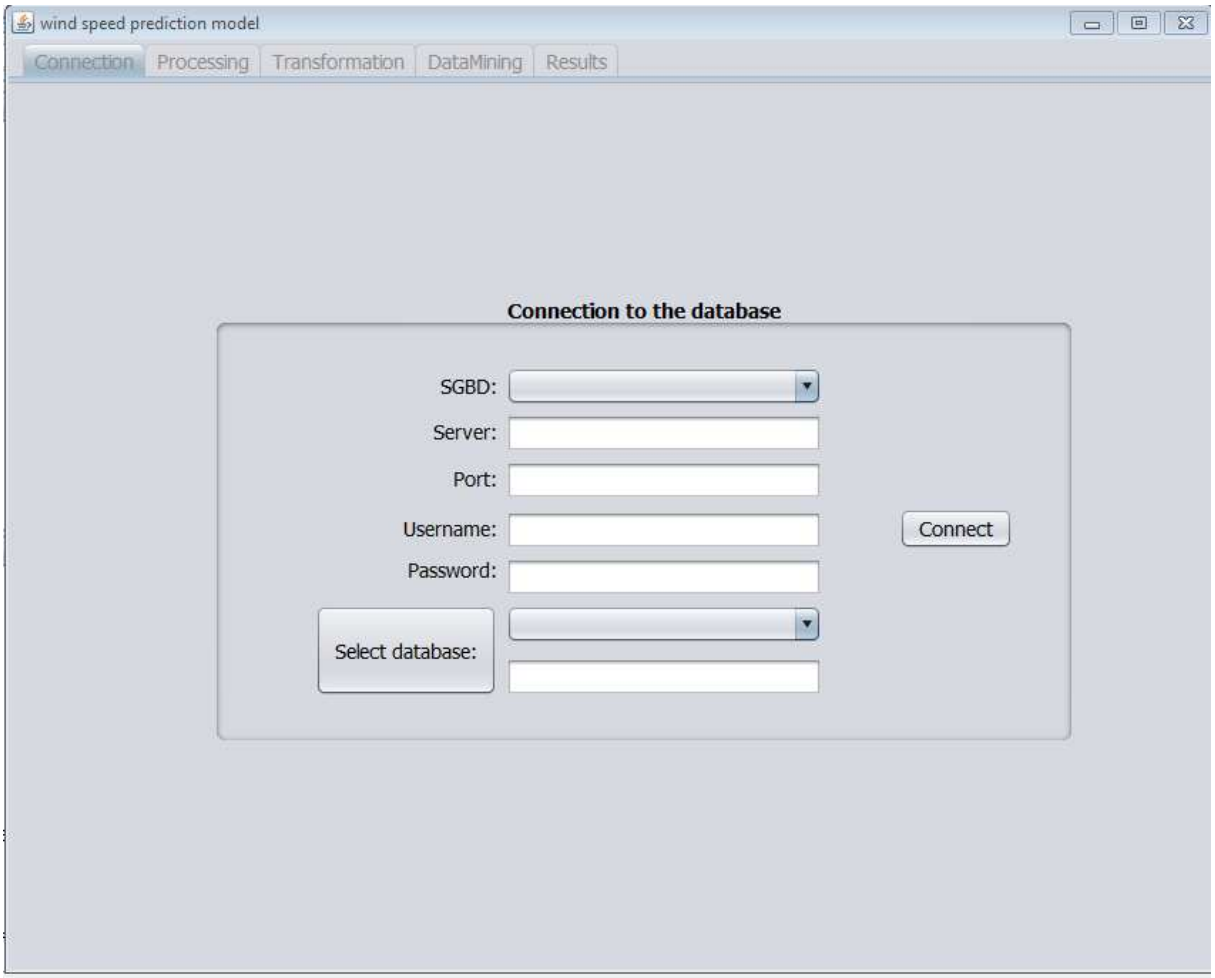

Figura 14 – Protótipo em desenvolvimento

Fonte: Autoria Própria

- Construir uma ferramenta completa e automatizada viabilizando inovações e vantagens operacionais; inclusive, um protótipo está sendo desenvolvido usando linguagem Java para acessar bases de dados, sinalizar dados suspeitos de inconsistência, pré-processar, transformar, minerar dados e visualizar resultados, conforme apresentado na Figura 14;
- Desenvolver um método híbrido para previsão na turbina eólica e na PCD de Petrolina; abordagens híbridas podem fornecer resultados melhores para um conjunto de dados particular que foi bem compreendido;
- Implementar e testar novos algoritmos para previsão da velocidade do vento;
- Testar novos atributos que podem influenciar o comportamento do vento;
- Executar novos experimentos a partir de novos dados de turbina eólica com diferentes alturas.

REFERÊNCIAS

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-Based Learning Algorithms. *Mach. Learn.*, v. 6, n. 1, p. 37–66, 1991. ISSN 08856125. Disponível em: <http://link.springer.com/10.1023/A:1022689900470>. Citado na página 40.

AHRENS, C. D.; HENSON, R. *Meteorology Today: An Introduction to Weather, Climate, and the Environment*. Eleventh. Boston, MA, US: Cengage Learning, 2016. 662 p. ISBN 9781305113589. Disponível em: <https://searchworks.stanford.edu/view/10968724>. Citado 4 vezes nas páginas 15, 21, 30 e 31.

AL-ODAN, H. A.; AL-DARAISEH, A. A. Open Source Data Mining tools. In: *2015 Int. Conf. Electr. Inf. Technol.* Marrakech, Morocco: IEEE, 2015. p. 369–374. ISBN 978-1-4799-7479-5. Citado na página 35.

BARBOUNIS, T. G.; THEOCHARIS, J. B. Locally recurrent neural networks for wind speed prediction using spatial correlation. *Inf. Sci. (Ny).*, v. 177, n. 24, p. 5775–5797, 2007. ISSN 00200255. Citado na página 26.

BARRY, R. G.; CHORLEY, R. J. *Atmosfera, Tempo e Clima*. 9 edt. ed. Porto Alegre, Brasil: Bookman, 2013. 495 p. ISBN 9788565837392. Citado na página 14.

BISHOP, C. M. *Neural Networks for Pattern Recognition*. 1 edt. ed. United States: Clarendon Press and Press, Oxford University, 1995. 504 p. ISBN 0198538642. Citado na página 38.

BLOMBERG, L. C.; HEMERICH, D.; RUIZ, D. D. A. Evaluating the performance of regression algorithms on datasets with missing data. *Int. J. Bus. Intell. Data Min.*, v. 8, n. 2, p. 105–131, nov 2013. ISSN 1743-8187. Citado na página 39.

BOUZGOU, H.; BENOUDJIT, N. Multiple architecture system for wind speed prediction. *Appl. Energy*, Elsevier Ltd, v. 88, n. 7, p. 2463–2471, jul 2011. ISSN 03062619. Citado na página 26.

BRAGA, A. d. P.; LUDERMIR, T. B.; CARVALHO, A. C. L. F. *Redes Neurais Artificiais : Teoria e Aplicações*. 1 edt. ed. Rio de Janeiro, Brasil: LTC, 2000. 262 p. Citado na página 37.

BREEZE, P. *Wind Power Generation*. Fisrt. San Diego, CA, US: Elsevier, 2016. 1–9 p. ISSN 18653529. ISBN 978-0-12-804038-6. Disponível em: <https://www.elsevier.com/books/wind-power-generation/breeze/978-0-12-804038-6>. Citado na página 19.

BURTON, T. et al. *Wind Energy Handbook*. England: Wiley, 2001. 643 p. ISBN 0-471-48997-2. Citado na página 20.

CADENAS, E.; RIVERA, W. Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA–ANN model. *Renew. Energy*, Elsevier, v. 35, n. 12, p. 2732–2738, dec 2010. ISSN 09601481. Citado na página 26.

CATALÃO, J.; POUSINHO, H.; MENDES, V. An Artificial Neural Network Approach for Short-Term Wind Power Forecasting in Portugal. In: *2009 15th Int. Conf. Intell. Syst. Appl. to Power Syst.* Curitiba, Brazil: IEEE, 2009. p. 1–5. ISBN 978-1-4244-5097-8. ISSN 1949-3029. Citado na página 20.

CATALÃO, J.; POUSINHO, H.; MENDES, V. Hybrid intelligent approach for short-term wind power forecasting in Portugal. *IET Renew. Power Gener.*, v. 5, n. 3, p. 251, may 2011. ISSN 17521416. Citado na página 20.

CHAI, T.; DRAXLER, R. R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.*, v. 7, n. 3, p. 1247–1250, jun 2014. ISSN 1991-9603. Citado 2 vezes nas páginas 33 e 34.

CHANG, G. W. et al. An improved neural network-based approach for short-term wind speed and power forecast. *Renew. Energy*, v. 105, p. 301–311, 2017. ISSN 18790682. Citado na página 26.

CHANG, W.-d.; SHIN, J. Missing Data Handling in Multi-Layer Perceptron. In: BOJKOVIC, Z. S. (Ed.). *Proc. 10th WSEAS Int. Conf. Comput.* Vouliagmeni, Athens, Greece: World Scientific and Engineering Academy and Society (WSEAS), 2006. p. 631–636. Disponível em: <http://www.wseas.us>. Citado na página 37.

CHEUNG, C.; LI, F. A quantitative correlation coefficient mining method for business intelligence in small and medium enterprises of trading business. *Expert Syst. Appl.*, v. 39, n. 7, p. 6279–6291, jun 2012. ISSN 09574174. Citado na página 33.

COLAK, I.; SAGIROGLU, S.; YESILBUDAK, M. Data mining and wind power prediction: A literature review. *Renew. Energy*, Elsevier Ltd, v. 46, p. 241–247, oct 2012. ISSN 09601481. Disponível em: <http://dx.doi.org/10.1016/j.renene.2012.02.015>. Citado 2 vezes nas páginas 17 e 20.

CORTES, C.; VAPNIK, V. Support-Vector Networks. *Mach. Learn.*, v. 20, n. 3, p. 273–297, 1995. ISSN 08856125. Citado na página 38.

DAMOUSIS, I. et al. A Fuzzy Model for Wind Speed Prediction and Power Generation in Wind Parks Using Spatial Correlation. *IEEE Trans. Energy Convers.*, v. 19, n. 2, p. 352–361, 2004. ISSN 0885-8969. Citado na página 26.

DARAEEPOUR, A.; ECHEVERRI, D. P. Day-ahead wind speed prediction by a Neural Network-based model. In: *ISGT 2014*. Washington, DC, USA: IEEE, 2014. p. 1–5. ISBN 978-1-4799-3653-3. Citado 2 vezes nas páginas 28 e 62.

EIA. *EIA projects 28% increase in world energy use by 2040*. 2017. 1 p. Disponível em: <https://www.eia.gov/todayinenergy/detail.aphp?id=32912>. Citado na página 17.

EL-FOULY, T.; EL-SAADANY, E.; SALAMA, M. One Day Ahead Prediction of Wind Speed and Direction. *IEEE Trans. Energy Convers.*, v. 23, n. 1, p. 191–201, 2008. ISSN 0885-8969. Citado na página 26.

EMEIS, S. *Wind Energy Meteorology*. 1. ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. 192 p. (Green Energy and Technology). ISBN 978-3-642-30522-1. Citado na página 20.

ERDEM, E.; SHI, J. ARMA based approaches for forecasting the tuple of wind speed and direction. *Appl. Energy*, Elsevier Ltd, v. 88, n. 4, p. 1405–1414, apr 2011. ISSN 03062619. Citado na página 26.

FAYYAD, U. Data mining and knowledge discovery in databases: implications for scientific databases. In: *Proceedings. Ninth Int. Conf. Sci. Stat. Database Manag. (Cat. No.97TB100150)*. Olympia, WA, USA: IEEE, 1997. p. 2–11. ISBN 0-8186-7952-2. Citado na página 24.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. *AI Mag.*, v. 17, n. 3, p. 37–54, nov 1996. Disponível em: <http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1230>. Citado na página 22.

FAZELPOUR, F.; TARASHKAR, N.; ROSEN, M. A. Short-term wind speed forecasting using artificial neural networks for Tehran, Iran. *Int. J. Energy Environ. Eng.*, v. 7, n. 4, p. 377–390, dec 2016. ISSN 2008-9163. Citado na página 26.

FINAMORE, A. R. et al. A day-ahead wind speed forecasting using data-mining model - a feed-forward NN algorithm. In: *2015 Int. Conf. Renew. Energy Res. Appl.* Palermo, Italy: IEEE, 2015. v. 5, p. 1230–1235. ISBN 978-1-4799-9982-8. Citado 2 vezes nas páginas 28 e 62.

FINAMORE, A. R. et al. A wind speed forecasting model based on artificial neural network and meteorological data. In: *2016 IEEE 16th Int. Conf. Environ. Electr. Eng.* Florence, Italy: IEEE, 2016. p. 1–5. ISBN 978-1-5090-2320-2. Citado na página 26.

FREIDMAN, J. H.; BENTLEY, J. L.; FINKEL, R. A. An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Trans. Math. Softw.*, v. 3, n. 3, p. 209–226, sep 1977. ISSN 00983500. Citado na página 47.

FREITAS, N. C. A. D.; SILVA, M. P. S.; SAKAMOTO, M. S. Wind Speed Forecasting: A Review. *Int. J. Eng. Res. Appl. (IJERA)*, v. 8, n. 1, p. 4–9, jan 2018. Disponível em: <http://www.ijera.com>. Citado 2 vezes nas páginas 25 e 66.

GAO, Z. et al. An overview on development of wind power generation. In: *2016 Chinese Control Decis. Conf.* Yinchuan, China: IEEE, 2016. p. 435–439. ISBN 978-1-4673-9714-8. Citado na página 17.

GARCÍA, S.; LUENGO, J.; HERRERA, F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Syst.*, Elsevier B.V., v. 98, p. 1–29, apr 2016. ISSN 09507051. Citado 2 vezes nas páginas 23 e 31.

GARDNER, M.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos. Environ.*, v. 32, n. 14-15, p. 2627–2636, aug 1998. ISSN 13522310. Citado na página 37.

GOLDSCHMIDT, R.; PASSOS, E. *Data Mining: um guia prático*. four. Rio de Janeiro, Brasil: Elsevier Campus, 2005. 261 p. ISBN 9788535218770. Citado na página 24.

GOODWIN, P.; LAWTON, R. On the asymmetry of the symmetric MAPE. *Int. J. Forecast.*, v. 15, n. 4, p. 405–408, oct 1999. ISSN 01692070. Citado 2 vezes nas páginas 33 e 34.

GUO, Z.-h. et al. A case study on a hybrid wind speed forecasting method using BP neural network. *Knowledge-Based Syst.*, Elsevier B.V., v. 24, n. 7, p. 1048–1056, oct 2011. ISSN 09507051. Citado na página 26.

GWEC. *Global Wind Statistics 2016*. Brussels, Belgium, 2017. 4 p. Disponível em: <http://gwec.net/publications/global-wind-report-2/>. Citado na página 18.

HALL, M. et al. The WEKA data mining software. *ACM SIGKDD Explor. Newsl.*, v. 11, n. 1, p. 10, nov 2009. ISSN 19310145. Citado na página 35.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining Concepts and Techniques*. third. Waltham, MA, US: Elsevier, 2012. 703 p. ISBN 978-0-12-381479-1. Disponível em: <http://www.sciencedirect.com/science/book/9780123814791>. Citado 6 vezes nas páginas 22, 23, 24, 25, 39 e 41.

HU, Q. et al. Short-Term Wind Speed or Power Forecasting With Heteroscedastic Support Vector Regression. *IEEE Trans. Sustain. Energy*, v. 7, n. 1, p. 241–249, jan 2016. ISSN 1949-3029. Disponível em: <http://ieeexplore.ieee.org/document/7335638/>. Citado 3 vezes nas páginas 26, 28 e 62.

IDC. *Data Growth, Business Opportunities, and the IT Imperatives*. 2014. 1–6 p. Disponível em: <https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm>. Citado na página 21.

JIANG, P.; WANG, Y.; WANG, J. Short-term wind speed forecasting using a hybrid model. *Energy*, Elsevier Ltd, v. 119, p. 561–577, jan 2017. ISSN 03605442. Disponível em: <http://dx.doi.org/10.1016/j.energy.2016.10.040>. Citado na página 26.

KANTARDZI, M. *Data mining: concepts, models, methods, and algorithms*. Second. Hoboken, New Jersey, US: John Wiley & Sons, 2011. 552 p. ISBN 978-0-470-89045-5. Disponível em: <http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470890452.html>. Citado na página 23.

KAUR, T.; KUMAR, S.; SEGAL, R. Application of artificial neural network for short term wind speed forecasting. In: *2016 Bienn. Int. Conf. Power Energy Syst. Towar. Sustain. Energy*. Bangalore, India: IEEE, 2016. p. 1–5. ISBN 978-1-4673-6660-1. Citado na página 26.

KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: . Montreal, Quebec, Canada: Morgan Kaufmann, 1995. p. 1137–1143. Citado na página 41.

LAHOUAR, A.; SLAMA, J. B. H. Wind speed and direction prediction for wind farms using support vector regression. In: *2014 5th Int. Renew. Energy Congr.* Hammamet, Tunisia: IEEE, 2014. p. 1–6. ISBN 978-1-4799-2195-9. Citado 3 vezes nas páginas 26, 27 e 62.

LAROSE, D. T.; LAROSE, C. D. *Discovering Knowledge in Data: an introduction to data mining*. Second. New Jersey, Canada: IEEE computer society, Wiley, 2014. ISBN 9780470908747. Citado na página 24.

LAZIĆ, L.; PEJANOVIĆ, G.; ŽIVKOVIĆ, M. Wind forecasts for wind power generation using the Eta model. *Renew. Energy*, v. 35, n. 6, p. 1236–1243, jun 2010. ISSN 09601481. Citado na página 26.

LEI, M. et al. A review on the forecasting of wind speed and generated power. *Renew. Sustain. Energy Rev.*, v. 13, n. 4, p. 915–920, may 2009. ISSN 13640321. Citado na página 26.

LIU, D. et al. Short-term wind speed forecasting using wavelet transform and support vector machines optimized by genetic algorithm. *Renew. Energy*, Elsevier Ltd, v. 62, p. 592–597, 2014. ISSN 09601481. Disponível em: <http://dx.doi.org/10.1016/j.renene.2013.08.011>. Citado 2 vezes nas páginas 26 e 62.

LIU, X.; KONG, X.; LEE, K. Y. Wind Speed Prediction with high efficiency convex optimization Support Vector Machine. In: *Proceeding 11th World Congr. Intell. Control Autom.* Shenyang, China: IEEE, 2014. p. 908–915. ISBN 978-1-4799-5825-2. Citado na página 26.

LUO, Q. Advancing Knowledge Discovery and Data Mining. In: *First Int. Work. Knowl. Discov. Data Min. (WKDD 2008)*. Adelaide, SA, Australia: IEEE, 2008. p. 3–5. ISBN 0-7695-3090-7. Citado na página 25.

MAÇAIRA, P. M.; SOUZA, R. C.; OLIVEIRA, F. L. C. Forecasting Brazil's electricity consumption with Pegels Exponential Smoothing Techniques. *IEEE Lat. Am. Trans.*, v. 14, n. 3, p. 1252–1258, mar 2016. ISSN 1548-0992. Citado na página 20.

MALIK, H.; SAVITA. Application of artificial neural network for long term wind speed prediction. In: *2016 Conf. Adv. Signal Process.* Pune, India: IEEE, 2016. p. 217–222. ISBN 978-1-5090-0849-0. Citado 2 vezes nas páginas 26 e 62.

MME. *Monthly Energy Bulletin - Brazil*. Brazil, 2017. v. 6, 1–2 p. Citado na página 19.

MOHANDES, M.; REHMAN, S.; RAHMAN, S. M. Estimation of wind speed profile using adaptive neuro-fuzzy inference system (ANFIS). *Appl. Energy*, Elsevier Ltd, v. 88, n. 11, p. 4024–4032, 2011. ISSN 03062619. Disponível em: <http://dx.doi.org/10.1016/j.apenergy.2011.04.015>. Citado na página 26.

MOHANDES, M. A.; REHMAN, S.; HALAWANI, T. O. A neural networks approach for wind speed prediction. *Renew. Energy*, v. 13, n. 3, p. 345–354, mar 1998. ISSN 09601481. Citado 2 vezes nas páginas 26 e 62.

NOAA. *National Oceanic and Atmospheric Administration Weather*. 2017. 1 p. Disponível em: <http://www.noaa.gov/weather>. Citado na página 15.

NREL. *Wind Power Today 2010*. Oak Ridge, TN, EUA, 2010. 1–32 p. Disponível em: <https://www.nrel.gov/docs/fy10osti/47531.pdf>. Citado na página 20.

OOI, H.-L.; NG, S.-C.; LIM, E. ANO Detection with K-Nearest Neighbor Using Minkowski Distance. *Int. J. Signal Process. Syst.*, v. 1, n. 2, p. 208–211, 2013. ISSN 23154535. Citado na página 40.

PACHECO, F. Energias Renováveis : Breves Conceitos. *Conjunt. e Planej.*, Salvador, Brazil, SEI, n. 149, p. 4–11, oct 2006. Disponível em: <http://docplayer.com.br/936890-Energias-renovaveis-breves-conceitos.html>. Citado na página 17.

PINTO, T. et al. Short-term wind speed forecasting using Support Vector Machines. *IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments*, IEEE, Orlando, FL, USA, v. 318912, n. 318912, p. 40–46, dec 2015. Citado 3 vezes nas páginas 26, 27 e 62.

PLATT, J. C. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. US, 1998. 1–21 p. Disponível em: <https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/>. Citado na página 38.

PLATT, J. C. Fast training Support Vector Machines using parallel sequential minimal optimization. In: *Adv. kernel methods - Support Vector Learn.* Cambridge, MA, US: MIT Press, 1999. cap. 12, p. 40–65. Disponível em: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/smo-book.pdf>. Citado na página 38.

QUINLAN, J. R. Learning With Continuous Classes. In: *Proc. Aust. Jt. Conf. Artif. Intell.* Hobart, Tasmania: World Scientific, 1992. v. 92, p. 343–348. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.885>. Citado na página 39.

RAMOS, C. M. C. et al. Modelagem da variação horária da temperatura do ar em Petrolina, PE, e Botucatu, SP. *Rev. Bras. Eng. Agrícola e Ambient.*, Campina Grande, PB, Brazil, v. 15, n. 9, p. 959–965, sep 2011. ISSN 1415-4366. Citado na página 43.

RAMOS, S. et al. A data-mining based methodology for wind forecasting. In: *2011 16th Int. Conf. Intell. Syst. Appl. to Power Syst.* Hersonissos, Greece: IEEE, 2011. p. 1–6. ISBN 978-1-4577-0809-1. Citado 2 vezes nas páginas 28 e 62.

REBOITA, M. S. et al. Entendendo o Tempo e o Clima na América do Sul. *Terra e Didat.*, v. 8, n. 1, p. 34–50, 2012. Citado na página 14.

SALCEDO-SANZ, S. et al. Short term wind speed prediction based on evolutionary support vector regression algorithms. *Expert Syst. Appl.*, v. 38, n. 4, p. 4052–4057, apr 2011. ISSN 09574174. Citado na página 62.

SHAO, H.; CUI, F.; DENG, X. Short-term wind speed forecasting using the wavelet decomposition and AdaBoost technique in wind farm of East China. *IET Gener. Transm. Distrib.*, v. 10, n. 11, p. 2585–2592, aug 2016. ISSN 1751-8687. Citado na página 26.

SHEVADE, S. et al. Improvements to the SMO algorithm for SVM regression. *IEEE Trans. Neural Networks*, v. 11, n. 5, p. 1188–1193, sep 2000. ISSN 10459227. Citado na página 38.

SINGH, A.; YADAV, A.; RANA, A. K-means with Three different Distance Metrics. *Int. J. Comput. Appl.*, v. 67, n. 10, p. 13–17, 2013. ISSN 09758887. Disponível em: <http://research.ijcaonline.org/volume67/number10/pxc3886785.pdf>. Citado na página 40.

SOMAN, S. S. et al. A review of wind power and wind speed forecasting methods with different time horizons. In: *North Am. Power Symp. 2010.* Arlington, TX, USA: IEEE, 2010. p. 1–8. ISBN 978-1-4244-8046-3. Citado na página 26.

STAVISS, B. Usina Eólica de Alegria I. *Infraesturura urbana Proj. custos e construção*, publicado na Web, n. 6, p. 3, aug 2011. Disponível em: <http://infraestruturaurbana17.pini.com.br/solucoes-tecnicas/6/artigo227165-2.aspx>. Citado na página 19.

TAYLOR, R. Interpretation of the Correlation Coefficient: A Basic Review. *J. Diagnostic Med. Sonogr.*, v. 6, n. 1, p. 35–39, jan 1990. ISSN 8756-4793. Citado na página 33.

TAYMAN, J.; SWANSON, D. A. On the validity of MAPE as a measure of population forecast accuracy. *Popul. Res. Policy Rev.*, v. 18, n. 4, p. 299–322, 1999. ISSN 01675923. Citado 2 vezes nas páginas 33 e 34.

TEXEIRA, R. F. B. *Satélites Meteorológicos: Imagens, aplicações e curiosidades.* 1. ed. Fortaleza, Brasil.: Típrografia Íris, 2016. 192 p. ISBN 978-85-64314-30-6. Citado na página 15.

TOLMASQUIM, M. As origens da crise energética brasileira. *Ambient. Soc.*, Campinas, Sao Paulo, Brazil, v. 2, n. 6-7, p. 179–183, jun 2000. ISSN 1414-753X. Disponível em: <http://dx.doi.org/10.1590/S1414-753X2000000100012>. Citado na página 17.

TOLMASQUIM, M. T. *Energia Renovável: Hidráulica, Biomassa, Eólica, Solar, Oceânica*. Rio de Janeiro, Brasil: Empresa de Pesquisa Energética (EPE), 2016. 452p p. ISBN 978-85-60025-06-0. Citado 2 vezes nas páginas 20 e 21.

UCZAI, P. *Energias Renováveis - Riqueza Sustentável ao Alcance da Sociedade*. Brasilia, Brasil: Edições Câmara, 2012. 273 p. ISBN 9788573659740. Citado na página 19.

ÜSTÜN, B.; MELSSEN, W. J.; BUYDENS, L. M. C. Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel. *Chemom. Intell. Lab. Syst.*, v. 81, n. 1, p. 29–40, oct 2006. ISSN 01697439. Citado na página 48.

VELO, R.; LÓPEZ, P.; MASEDA, F. Wind speed estimation using multilayer perceptron. *Energy Convers. Manag.*, v. 81, p. 1–9, 2014. ISSN 01968904. Citado na página 26.

WANG, J. et al. Medium-term wind speeds forecasting utilizing hybrid models for three different sites in Xinjiang, China. *Renew. Energy*, Elsevier, v. 76, p. 91–101, apr 2015. ISSN 09601481. Citado na página 26.

WANG, J. et al. A novel hybrid approach for wind speed prediction. *Inf. Sci. (Ny).*, Elsevier Inc., v. 273, p. 304–318, 2014. ISSN 00200255. Citado na página 26.

WANG, Y.; WITTEN, I. H. *Inducing Model Trees for Continuous Classes (M5P)*. Hamilton, New Zealand, 1997. 1–10 p. (Computer Science Working Papers). Disponível em: <http://www.cs.waikato.ac.nz/{~}ml/publications/1997/Wang-Witten-Induct.p>. Citado na página 39.

WANG, Z.; BOVIK, A. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Process. Mag.*, v. 26, n. 1, p. 98–117, jan 2009. ISSN 1053-5888. Citado na página 33.

WILLMOTT, C.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.*, v. 30, n. 1, p. 79–82, dec 2005. ISSN 0936-577X. Citado 2 vezes nas páginas 33 e 34.

WITTEN, I. et al. *Data Mining Practical Machine Learning Tools and Techniques*. Fourth. Cambridge, MA, US: Elsevier, 2017. 622 p. ISBN 9780128042915. Disponível em: <http://www.sciencedirect.com/science/book/9780128042915>. Citado 5 vezes nas páginas 22, 25, 33, 39 e 41.

YESILBUDAK, M.; SAGIROGLU, S.; COLAK, I. A new approach to very short term wind speed prediction using k-nearest neighbor classification. *Energy Convers. Manag.*, v. 69, p. 77–86, 2013. ISSN 01968904. Citado 2 vezes nas páginas 27 e 62.

ZAKI, M. J.; MEIRA-JR, W. *Data mining and analysis: Fundamental Concepts and Algorithms*. First. United States of America: Cambridge University Press, 2014. 585 p. ISBN 9780521766333. Citado na página 23.

ZHAO, X.; WANG, S.; LI, T. Review of evaluation criteria and main methods of wind power forecasting. *Energy Procedia*, v. 12, p. 761–769, sep 2011. ISSN 18766102. Citado na página 28.

ZHU, X.; GENTON, M. G. Short-Term Wind Speed Forecasting for Power System Operations. *Int. Stat. Rev.*, v. 80, n. 1, p. 2–23, apr 2012. ISSN 03067734. Citado na página 20.

ZUO, Y.; LIU, H. Evaluation on comprehensive benefit of wind power generation and utilization of wind energy. In: *2012 IEEE Int. Conf. Comput. Sci. Autom. Eng.* Beijing, China: IEEE, 2012. p. 635–638. ISBN 978-1-4673-2008-5. Citado na página 17.

APÊNDICE A – CÓDIGO NA LINGUAGEM SQL PARA EXPORTAÇÃO DOS DADOS HORÁRIOS DO PRIMEIRO ESTUDO DE CASO

```
/*COMANDO COPY PARA GERAR UM ARQUIVO CSV */

COPY( SELECT ano, dia_do_ano,

/* GENERALIZACAO PARA HORA DO DIA*/

minuto_do_dia/60 as hora_do_dia,

/* GENERALIZACAO PARA MES INTEIRO */

CASE WHEN dia_do_ano > 0 AND dia_do_ano <= 31 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '1' /*31 dias*/ WHEN dia_do_ano > 31 AND dia_do_ano <= 59 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '2' /*28 dias*/ WHEN dia_do_ano > 59 AND dia_do_ano <= 90 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '3' /*31 dias*/ WHEN dia_do_ano > 90 AND dia_do_ano <= 120 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '4' /*30 dias*/ WHEN dia_do_ano > 120 AND dia_do_ano <= 151 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '5' /*31 dias*/ WHEN dia_do_ano > 151 AND dia_do_ano <= 181 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '6' /*30 dias*/ WHEN dia_do_ano > 181 AND dia_do_ano <= 212 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '7' /*31 dias*/ WHEN dia_do_ano > 212 AND dia_do_ano <= 243 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '8' /*31 dias*/ WHEN dia_do_ano > 243 AND dia_do_ano <= 273 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '9' /*30 dias*/ WHEN dia_do_ano > 273 AND dia_do_ano <= 304 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '10' /*31 dias*/ WHEN dia_do_ano > 304 AND dia_do_ano <= 334 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '11' /*30 dias*/ WHEN dia_do_ano > 334 AND dia_do_ano <= 365 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '12' /*31 dias*/ WHEN dia_do_ano > 0 AND dia_do_ano <= 31 AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN '1' /*31 dias*/ WHEN dia_do_ano > 31 AND dia_do_ano <= 60 AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN '2' /*29 dias*/ WHEN dia_do_ano > 60 AND dia_do_ano <= 91 AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN '3' /*31 dias*/ WHEN dia_do_ano > 91 AND dia_do_ano <= 121 AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN '4' /*30 dias*/ WHEN dia_do_ano > 121 AND dia_do_ano <= 152 AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN '5' /*31 dias*/ WHEN dia_do_ano > 152 AND dia_do_ano <= 182 AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN '6' /*30 dias*/ WHEN dia_do_ano > 182 AND
```

dia_do_ano <= 213 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’7’ /*31 dias*/ WHEN dia_do_ano > 213 AND dia_do_ano <= 244 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’8’ /*31 dias*/ WHEN dia_do_ano > 244 AND dia_do_ano <= 274 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’9’ /*30 dias*/ WHEN dia_do_ano > 274 AND dia_do_ano <= 305 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’10’ /*31 dias*/ WHEN dia_do_ano > 305 AND dia_do_ano <= 335 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’11’ /*30 dias*/ WHEN dia_do_ano > 335 AND dia_do_ano <= 366 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’12’ /*31 dias*/ ELSE ’NULL’ END AS mes,

/* GENERALIZACAO PARA MES NOMINAL */

CASE WHEN dia_do_ano > 0 AND dia_do_ano <= 31 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN ’Jan’ /*31 dias*/ WHEN dia_do_ano > 31 AND dia_do_ano <= 59 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN ’Fev’ /*28 dias*/ WHEN dia_do_ano > 59 AND dia_do_ano <= 90 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN ’Mar’ /*31 dias*/ WHEN dia_do_ano > 90 AND dia_do_ano <= 120 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN ’Abr’ /*30 dias*/ WHEN dia_do_ano > 120 AND dia_do_ano <= 151 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN ’Mai’ /*31 dias*/ WHEN dia_do_ano > 151 AND dia_do_ano <= 181 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN ’Jun’ /*30 dias*/ WHEN dia_do_ano > 181 AND dia_do_ano <= 212 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN ’Jul’ /*31 dias*/ WHEN dia_do_ano > 212 AND dia_do_ano <= 243 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN ’Agt’ /*31 dias*/ WHEN dia_do_ano > 243 AND dia_do_ano <= 273 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN ’Set’ /*30 dias*/ WHEN dia_do_ano > 273 AND dia_do_ano <= 304 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN ’Out’ /*31 dias*/ WHEN dia_do_ano > 304 AND dia_do_ano <= 334 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN ’Nov’ /*30 dias*/ WHEN dia_do_ano > 334 AND dia_do_ano <= 365 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN ’Dez’ /*31 dias*/ WHEN dia_do_ano > 0 AND dia_do_ano <= 31 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’Jan’ /*31 dias*/ WHEN dia_do_ano > 31 AND dia_do_ano <= 60 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’Fev’ /*29 dias*/ WHEN dia_do_ano > 60 AND dia_do_ano <= 91 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’Mar’ /*31 dias*/ WHEN dia_do_ano > 91 AND dia_do_ano <= 121 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’Abr’ /*30 dias*/ WHEN dia_do_ano > 121 AND dia_do_ano <= 152 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’Mai’ /*31 dias*/ WHEN dia_do_ano > 152 AND dia_do_ano <= 182 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’Jun’ /*30 dias*/ WHEN dia_do_ano > 182 AND dia_do_ano <= 213 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’Jul’ /*31 dias*/ WHEN dia_do_ano > 213 AND dia_do_ano <= 244 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’Ago’ /*31 dias*/ WHEN dia_do_ano > 244 AND dia_do_ano <= 274 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN ’Set’

```
/*30 dias*/ WHEN dia_do_ano > 274 AND dia_do_ano <= 305 AND ( ano = 2008 OR ano =
2012 OR ano = 2016) THEN 'Out' /*31 dias*/ WHEN dia_do_ano > 305 AND dia_do_ano
<= 335 AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN 'Nov' /*30 dias*/ WHEN
dia_do_ano > 335 AND dia_do_ano <= 366 AND ( ano = 2008 OR ano = 2012 OR ano = 2016)
THEN 'Dez' /*31 dias*/ ELSE 'NULL' END AS mes_nominal,

/*GENERALIZACAO PARA DIRECAO DO VENTO NOMINAL*/

CASE WHEN AVG(dv) > 337.5 OR AVG(dv) <= 22.5 THEN 'Norte' WHEN AVG(dv)
> 22.5 AND AVG(dv) <= 67.5 THEN 'Nordeste' WHEN AVG(dv) > 67.5 AND AVG(dv) <=
112.5 THEN 'Leste' WHEN AVG(dv) > 112.5 AND AVG(dv) <= 157.5 THEN 'Sudeste' WHEN
AVG(dv) > 157.5 AND AVG(dv) <= 202.5 THEN 'Sul' WHEN AVG(dv) > 202.5 AND AVG(dv)
<= 247.5 THEN 'Sudoeste' WHEN AVG(dv) > 247.5 AND AVG(dv) <= 294.5 THEN 'Oeste'
WHEN AVG(dv) > 294.5 AND AVG(dv) < 337.5 THEN 'Noroeste' ELSE 'NULL' END AS
dv_nominal,

/* TRANFORMACAO PARA HORARIO E NORMALIZACAO EM DUAS CASAS APOS A
VIRGULA*/

ROUND(AVG(ta),2) as TP, ROUND(AVG(ur),2) as UR, ROUND(AVG(pa),2) as PA,
ROUND(AVG(vv),2) as VV, ROUND(AVG(dv),2) AS DV
FROM tb_2009_a_2016

/*IMPLEMENTACAO DO CRITERIO FISICAMENTE IMPOSSIVEL*/

WHERE
ta > 15 AND ta <= 40 AND vv > 0.0 AND vv <= 25 AND dv > 0 AND dv < 360 AND pa >=
950 AND pa <= 980 AND ur > 0 AND ur < 100

/*IMPLEMENTACAO DO SEGUNDO E DO TERCEIRO CRITERIO DA SONDA*/

AND (ta_vl = 999) AND (vv_vl = 999) AND dv_vl = 999 AND (ur_vl = 9 OR ur_vl = 99) AND
(pa_vl = 99 OR pa_vl = 999) GROUP BY hora_do_dia, dia_do_ano, mes, ano ORDER BY ano,
dia_do_ano

/*FINALIZANDO O COMANDO COPY E E GERANDO UM ARQUIVO CSV */

)TO 'C:/arquivo.csv' DELIMITER ',' CSV HEADER
```

APÊNDICE B – CÓDIGO NA LINGUAGEM SQL PARA EXPORTAÇÃO DOS DADOS DIÁRIOS DO PRIMEIRO ESTUDO DE CASO

/*COMANDO COPY PARA GERAR UM ARQUIVO CSV */

COPY(SELECT ano, dia_do_ano,

/*GENERALIZACAO DO ATRIBUTO MES NA FORMA NUMERICA */

CASE WHEN dia_do_ano > 0 AND dia_do_ano <= 31 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '1' /*31 dias*/ WHEN dia_do_ano > 31 AND dia_do_ano <= 59 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '2' /*28 dias*/ WHEN dia_do_ano > 59 AND dia_do_ano <= 90 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '3' /*31 dias*/ WHEN dia_do_ano > 90 AND dia_do_ano <= 120 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '4' /*30 dias*/ WHEN dia_do_ano > 120 AND dia_do_ano <= 151 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '5' /*31 dias*/ WHEN dia_do_ano > 151 AND dia_do_ano <= 181 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '6' /*30 dias*/ WHEN dia_do_ano > 181 AND dia_do_ano <= 212 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '7' /*31 dias*/ WHEN dia_do_ano > 212 AND dia_do_ano <= 243 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '8' /*31 dias*/ WHEN dia_do_ano > 243 AND dia_do_ano <= 273 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '9' /*30 dias*/ WHEN dia_do_ano > 273 AND dia_do_ano <= 304 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '10' /*31 dias*/ WHEN dia_do_ano > 304 AND dia_do_ano <= 334 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '11' /*30 dias*/ WHEN dia_do_ano > 334 AND dia_do_ano <= 365 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN '12' /*31 dias*/ WHEN dia_do_ano > 0 AND dia_do_ano <= 31 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN '1' /*31 dias*/ WHEN dia_do_ano > 31 AND dia_do_ano <= 60 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN '2' /*29 dias*/ WHEN dia_do_ano > 60 AND dia_do_ano <= 91 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN '3' /*31 dias*/ WHEN dia_do_ano > 91 AND dia_do_ano <= 121 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN '4' /*30 dias*/ WHEN dia_do_ano > 121 AND dia_do_ano <= 152 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN '5' /*31 dias*/ WHEN dia_do_ano > 152 AND dia_do_ano <= 182 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN '6' /*30 dias*/ WHEN dia_do_ano > 182 AND dia_do_ano <= 213 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN '7' /*31 dias*/ WHEN dia_do_ano > 213 AND dia_do_ano <= 244 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN '8' /*31 dias*/ WHEN dia_do_ano > 244 AND dia_do_ano <= 274 AND (ano = 2008 OR ano = 2012 OR ano = 2016) THEN '9' /*30 dias*/ WHEN dia_do_ano > 274 AND

```
dia_do_ano <= 305 AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN '10' /*31 dias*/
WHEN dia_do_ano > 305 AND dia_do_ano <= 335 AND ( ano = 2008 OR ano = 2012 OR ano
= 2016) THEN '11' /*30 dias*/ WHEN dia_do_ano > 335 AND dia_do_ano <= 366 AND ( ano
= 2008 OR ano = 2012 OR ano = 2016) THEN '12' /*31 dias*/ ELSE 'NULL' END AS mes,

/*GENERALIZACAO DO ATRIBUTO MES NA FORMA NOMINAL*/

CASE WHEN dia_do_ano > 0 AND dia_do_ano <= 31 AND ano <> 2012 AND ano <> 2016
AND ano <> 2008 THEN 'Jan' /*31 dias*/ WHEN dia_do_ano > 31 AND dia_do_ano <= 59
AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN 'Fev' /*28 dias*/ WHEN
dia_do_ano > 59 AND dia_do_ano <= 90 AND ano <> 2012 AND ano <> 2016 AND ano <>
2008 THEN 'Mar' /*31 dias*/ WHEN dia_do_ano > 90 AND dia_do_ano <= 120 AND ano <>
2012 AND ano <> 2016 AND ano <> 2008 THEN 'Abr' /*30 dias*/ WHEN dia_do_ano > 120
AND dia_do_ano <= 151 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN 'Mai'
/*31 dias*/ WHEN dia_do_ano > 151 AND dia_do_ano <= 181 AND ano <> 2012 AND ano <>
2016 AND ano <> 2008 THEN 'Jun' /*30 dias*/ WHEN dia_do_ano > 181 AND dia_do_ano
<= 212 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN 'Jul' /*31 dias*/ WHEN
dia_do_ano > 212 AND dia_do_ano <= 243 AND ano <> 2012 AND ano <> 2016 AND ano <>
2008 THEN 'Agt' /*31 dias*/ WHEN dia_do_ano > 243 AND dia_do_ano <= 273 AND ano <>
2012 AND ano <> 2016 AND ano <> 2008 THEN 'Set' /*30 dias*/ WHEN dia_do_ano > 273
AND dia_do_ano <= 304 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN 'Out'
/*31 dias*/ WHEN dia_do_ano > 304 AND dia_do_ano <= 334 AND ano <> 2012 AND ano <>
2016 AND ano <> 2008 THEN 'Nov' /*30 dias*/ WHEN dia_do_ano > 334 AND dia_do_ano
<= 365 AND ano <> 2012 AND ano <> 2016 AND ano <> 2008 THEN 'Dez' /*31 dias*/
WHEN dia_do_ano > 0 AND dia_do_ano <= 31 AND ( ano = 2008 OR ano = 2012 OR ano =
2016) THEN 'Jan' /*31 dias*/ WHEN dia_do_ano > 31 AND dia_do_ano <= 60 AND ( ano =
2008 OR ano = 2012 OR ano = 2016) THEN 'Fev' /*29 dias*/ WHEN dia_do_ano > 60 AND
dia_do_ano <= 91 AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN 'Mar' /*31 dias*/
WHEN dia_do_ano > 91 AND dia_do_ano <= 121 AND ( ano = 2008 OR ano = 2012 OR ano
= 2016) THEN 'Abr' /*30 dias*/ WHEN dia_do_ano > 121 AND dia_do_ano <= 152 AND (
ano = 2008 OR ano = 2012 OR ano = 2016) THEN 'Mai' /*31 dias*/ WHEN dia_do_ano > 152
AND dia_do_ano <= 182 AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN 'Jun' /*30
dias*/ WHEN dia_do_ano > 182 AND dia_do_ano <= 213 AND ( ano = 2008 OR ano = 2012
OR ano = 2016) THEN 'Jul' /*31 dias*/ WHEN dia_do_ano > 213 AND dia_do_ano <= 244
AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN 'Ago' /*31 dias*/ WHEN dia_do_ano
> 244 AND dia_do_ano <= 274 AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN 'Set'
/*30 dias*/ WHEN dia_do_ano > 274 AND dia_do_ano <= 305 AND ( ano = 2008 OR ano =
2012 OR ano = 2016) THEN 'Out' /*31 dias*/ WHEN dia_do_ano > 305 AND dia_do_ano
<= 335 AND ( ano = 2008 OR ano = 2012 OR ano = 2016) THEN 'Nov' /*30 dias*/ WHEN
```

```
dia_do_ano > 335 AND dia_do_ano <= 366 AND ( ano = 2008 OR ano = 2012 OR ano = 2016)
THEN 'Dez' /*31 dias*/ ELSE 'NULL' END AS mes_nominal,

/*GENERALIZACAO DO ATRIBUTO DIRECAO DO VENTO NA FORMA NOMINAL */

CASE WHEN AVG(dv) > 337.5 OR AVG(dv) <= 22.5 THEN 'Norte' WHEN AVG(dv) > 22.5
AND AVG(dv) <= 67.5 THEN 'Nordeste' WHEN AVG(dv) > 67.5 AND AVG(dv) <= 112.5
THEN 'Leste' WHEN AVG(dv) > 112.5 AND AVG(dv) <= 157.5 THEN 'Sudeste' WHEN
AVG(dv) > 157.5 AND AVG(dv) <= 202.5 THEN 'Sul' WHEN AVG(dv) > 202.5 AND AVG(dv)
<= 247.5 THEN 'Sudoeste' WHEN AVG(dv) > 247.5 AND AVG(dv) <= 294.5 THEN 'Oeste'
WHEN AVG(dv) > 294.5 AND AVG(dv) < 337.5 THEN 'Noroeste' ELSE 'NULL' END AS
dv_nominal,

/* TRANFORMACAO PARA DIARIO E NORMALIZACAO EM DUAS CASAS APOS A
VIRGULA*/

ROUND(AVG(ta),2) as TP, ROUND(AVG(ur),2) as UR, ROUND(AVG(pa),2) as PA,
ROUND(AVG(vv),2) as VV, ROUND(AVG(dv),2) AS DV FROM tb_2009_a_2016 WHERE

/*IMPLEMENTACAO DO CRITERIO FISICAMENTE IMPOSSIVEL*/

ta > 15 AND ta <= 40 AND vv > 0.0 AND vv <= 25 AND dv > 0 AND dv < 360 AND pa >=
950 AND pa <= 980 AND ur > 0 AND ur < 100

/*IMPLEMENTACAO DO SEGUNDO E TERCEIRO CRITERIO DA SONDA*/

AND (ta_vl = 999) AND (vv_vl = 999) AND dv_vl = 999 AND (ur_vl = 9 OR ur_vl = 99) AND
(pa_vl = 99 OR pa_vl = 999)

GROUP BY dia_do_ano, mes, ano ORDER BY ano, dia_do_ano

/*FINALIZANDO O COMANDO COPY E E GERANDO UM ARQUIVO CSV */

)TO 'C:/arquivo.csv' DELIMITER ',' CSV HEADER
```

APÊNDICE C – CÓDIGO NA LINGUAGEM SQL PARA EXPORTAÇÃO DOS DADOS HORÁRIOS DO SEGUNDO ESTUDO DE CASO

```
/*COMANDO COPY PARA GERAR UM ARQUIVO CSV */

COPY(

/*GENERALIZACAO DO ANO NA FORMA NUMERICA */

SELECT CAST(substring(ano_mes_dia from 1 for 4) AS int) as ano,

/*GENERALIZACAO DO ATRIBUTO MES NA FORMA NUMERICA */

CAST(substring(ano_mes_dia from 5 for 2)AS INT) as mes,

/*GENERALIZACAO DO ATRIBUTO MES NA FORMA NOMINAL */

CASE WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 1 THEN 'Janeiro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 2 THEN 'Fevereiro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 3 THEN 'Marco' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 4 THEN 'Abril' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 5 THEN 'Maio' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 6 THEN 'Junho' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 7 THEN 'Julho' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 8 THEN 'Agosto' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 9 THEN 'Setembro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 10 THEN 'Outubro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 11 THEN 'Novembro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 12 THEN 'Dezembro' /*31 dias*/ ELSE 'NULL' END AS mes_nominal,

/*GENERALIZACAO DA SEMANA DO MES */

CASE WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) < 8 THEN '1' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) < 16 THEN '2' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) < 24 THEN '3' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 31 THEN '4' ELSE 'null' END AS n_semana_mes,
```

```
/*GENERALIZACAO DOS TRES DIAS DO MES NUMERICO */

CASE WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 0 and
CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 3 THEN '1' WHEN
CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 3 and CAST(substring(ano_mes_dia
from 7 for 2)AS INT) <= 6 THEN '2' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS
INT) > 6 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 9 THEN '3' WHEN
CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 9 and CAST(substring(ano_mes_dia
from 7 for 2)AS INT) <= 12 THEN '4' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS
INT) > 12 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 15 THEN '5' WHEN
CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 15 and CAST(substring(ano_mes_dia
from 7 for 2)AS INT) <= 18 THEN '6' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS
INT) > 18 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 21 THEN '7' WHEN
CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 21 and CAST(substring(ano_mes_dia
from 7 for 2)AS INT) <= 24 THEN '8' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS
INT) > 24 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 27 THEN '9' WHEN
CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 27 and CAST(substring(ano_mes_dia
from 7 for 2)AS INT) <= 31 THEN '10' ELSE 'null' END AS tres_dias,

/*GENERALIZACAO DO DIA DO MES*/

CAST(substring(ano_mes_dia from 7 for 2)AS INT) as dia,

/*GENERALIZACAO DA HORA DO DIA NUMERICO */

CAST(substring(hora_minuto from 1 for 2) AS int) AS hora,

/*NORMALIZACAO DE DUAS CASAS DECIMAIS*/

ROUND(AVG(temperatura),2) as temperatura, ROUND(AVG(umidade_relativa),2)
as umidade_relativa, ROUND(AVG(pressao_atmosferica),2) as pressao_atmosferica,
ROUND(AVG(vv_50m),2) as vv_50m, ROUND(AVG(dv_50m),2) as dv_50,

/*GENERALIZACAO DO ATRIBUTO DIRECAO DO VENTO NOMINAL */

CASE WHEN AVG(dv_50m) > 337.5 OR AVG(dv_50m) <= 22.5 THEN 'Norte' WHEN
AVG(dv_50m) > 22.5 AND AVG(dv_50m) <= 67.5 THEN 'Nordeste' WHEN AVG(dv_50m)
> 67.5 AND AVG(dv_50m) <= 112.5 THEN 'Leste' WHEN AVG(dv_50m) > 112.5 AND
AVG(dv_50m) <= 157.5 THEN 'Sudeste' WHEN AVG(dv_50m) > 157.5 AND AVG(dv_50m)
<= 202.5 THEN 'Sul' WHEN AVG(dv_50m) > 202.5 AND AVG(dv_50m) <= 247.5 THEN
```

```
'Sudoeste' WHEN AVG(dv_50m) > 247.5 AND AVG(dv_50m) <= 294.5 THEN 'Oeste' WHEN
AVG(dv_50m) > 294.5 AND AVG(dv_50m) < 337.5 THEN 'Noroeste' ELSE 'NULL' END AS
dv_50m_nominal FROM parque_ceara

/*PROCESSAMENTO NOS DADOS */

WHERE cod_erro = '0' GROUP BY hora, dia, tres_dias,n_semana_mes, mes, ano ORDER BY
ano,mes,n_semana_mes, tres_dias, dia, hora

    ) TO 'C:/arquivo.csv' DELIMITER ',' CSV HEADER
```

APÊNDICE D – CÓDIGO NA LINGUAGEM SQL PARA EXPORTAÇÃO DOS DADOS DIÁRIOS DO SEGUNDO ESTUDO DE CASO

```
/*COMANDO COPY PARA GERAR UM ARQUIVO CSV */

COPY(

/*GENERALIZACAO DO ANO NA FORMA NUMERICA */

SELECT CAST(substring(ano_mes_dia from 1 for 4) AS int) as ano,

/*GENERALIZACAO DO MES NA FORMA NUMERICA */

CAST(substring(ano_mes_dia from 5 for 2)AS INT) as mes,

/*GENERALIZACAO DO MES NA FORMA NOMINAL */

CASE WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 1 THEN 'Janeiro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 2 THEN 'Fevereiro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 3 THEN 'Marco' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 4 THEN 'Abril' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 5 THEN 'Maio' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 6 THEN 'Junho' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 7 THEN 'Julho' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 8 THEN 'Agosto' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 9 THEN 'Setembro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 10 THEN 'Outubro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 11 THEN 'Novembro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 12 THEN 'Dezembro' /*31 dias*/ ELSE 'NULL' END AS mes_nominal,

/*GENERALIZACAO DA SEMANA DO MES NUMERICA */

CASE WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) < 8 THEN '1' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) < 16 THEN '2' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) < 24 THEN '3' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 31 THEN '4' ELSE 'null' END AS n_semana_mes,

/*GENERALIZACAO DO ATRIBUTO NUMERICO TRES DIAS */
```

```
CASE WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 0 and
CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 3 THEN '1' WHEN
CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 3 and CAST(substring(ano_mes_dia
from 7 for 2)AS INT) <= 6 THEN '2' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS
INT) > 6 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 9 THEN '3' WHEN
CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 9 and CAST(substring(ano_mes_dia
from 7 for 2)AS INT) <= 12 THEN '4' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS
INT) > 12 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 15 THEN '5' WHEN
CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 15 and CAST(substring(ano_mes_dia
from 7 for 2)AS INT) <= 18 THEN '6' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS
INT) > 18 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 21 THEN '7' WHEN
CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 21 and CAST(substring(ano_mes_dia
from 7 for 2)AS INT) <= 24 THEN '8' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS
INT) > 24 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 27 THEN '9' WHEN
CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 27 and CAST(substring(ano_mes_dia
from 7 for 2)AS INT) <= 31 THEN '10' ELSE 'null' END AS tres_dias,

/*NORMALIZACAO PARA DUAS CASAS DECIMAS POS VIRGULA*/

ROUND(AVG(temperatura),2) as temperatura, ROUND(AVG(umidade_relativa),2)
as umidade_relativa, ROUND(AVG(pressao_atmosferica),2) as pressao_atmosferica,
ROUND(AVG(vv_50m),2) as vv_50m, ROUND(AVG(dv_50m),2) as dv_50,

/*GENERALIZACAO PARA DIRECAO DO VENTO NOMINAL*/

CASE WHEN AVG(dv_50m) > 337.5 OR AVG(dv_50m) <= 22.5 THEN 'Norte' WHEN
AVG(dv_50m) > 22.5 AND AVG(dv_50m) <= 67.5 THEN 'Nordeste' WHEN AVG(dv_50m)
> 67.5 AND AVG(dv_50m) <= 112.5 THEN 'Leste' WHEN AVG(dv_50m) > 112.5 AND
AVG(dv_50m) <= 157.5 THEN 'Sudeste' WHEN AVG(dv_50m) > 157.5 AND AVG(dv_50m)
<= 202.5 THEN 'Sul' WHEN AVG(dv_50m) > 202.5 AND AVG(dv_50m) <= 247.5 THEN
'Sudoeste' WHEN AVG(dv_50m) > 247.5 AND AVG(dv_50m) <= 294.5 THEN 'Oeste' WHEN
AVG(dv_50m) > 294.5 AND AVG(dv_50m) < 337.5 THEN 'Noroeste' ELSE 'NULL' END AS
dv_50m_nominal FROM parque_ceara

/*PROCESSAMENTO NOS DADOS */

WHERE cod_erro = '0' GROUP BY tres_dias,n_semana_mes, mes, ano ORDER BY
ano,mes,n_semana_mes, tres_dias ) TO 'C:/arquivo.csv' DELIMITER ',' CSV HEADER
```

APÊNDICE E – CÓDIGO NA LINGUAGEM SQL PARA EXPORTAÇÃO DOS DADOS DE 3 DIAS DO SEGUNDO ESTUDO DE CASO

```
/*COMANDO COPY PARA GERAR UM ARQUIVO CSV */

COPY(

/*GENERALIZACAO DO ANO NA FORMA NUMERICA */

SELECT CAST(substring(ano_mes_dia from 1 for 4) AS int) as ano,

/*GENERALIZACAO DO MES NA FORMA NUMERICA */

CAST(substring(ano_mes_dia from 5 for 2)AS INT) as mes,

/*GENERALIZACAO DO MES NA FORMA NOMINAL */

CASE WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 1 THEN 'Janeiro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 2 THEN 'Fevereiro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 3 THEN 'Marco' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 4 THEN 'Abril' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 5 THEN 'Maio' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 6 THEN 'Junho' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 7 THEN 'Julho' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 8 THEN 'Agosto' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 9 THEN 'Setembro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 10 THEN 'Outubro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 11 THEN 'Novembro' /*31 dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 12 THEN 'Dezembro' /*31 dias*/ ELSE 'NULL' END AS mes_nominal,

/*GENERALIZACAO DA SEMANA DO MES NA FORMA NUMERICA */

CASE WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) < 8 THEN '1' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) < 16 THEN '2' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) < 24 THEN '3' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 31 THEN '4' ELSE 'null' END AS n_semana_mes,

/*GENERALIZACAO DOS TRES DIAS DO MES NA FORMA NUMERICA */
```

CASE WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 0 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 3 THEN '1' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 3 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 6 THEN '2' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 6 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 9 THEN '3' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 9 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 12 THEN '4' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 12 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 15 THEN '5' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 15 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 18 THEN '6' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 18 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 21 THEN '7' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 21 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 24 THEN '8' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 24 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 27 THEN '9' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) > 27 and CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 31 THEN '10' ELSE 'null' END AS tres_dias,

/*NORMALIZACAO PARA DUAS CASAS APOS A VIRGULA */

ROUND(AVG(temperatura),2) as temperatura, ROUND(AVG(umidade_relativa),2) as umidade_relativa, ROUND(AVG(pressao_atmosferica),2) as pressao_atmosferica, ROUND(AVG(vv_50m),2) as vv_50m, ROUND(AVG(dv_50m),2) as dv_50,

/*GENERALIZACAO DA DIRECAO DO VENTO NA FORMA NUMERICA */

CASE WHEN AVG(dv_50m) > 337.5 OR AVG(dv_50m) <= 22.5 THEN 'Norte' WHEN AVG(dv_50m) > 22.5 AND AVG(dv_50m) <= 67.5 THEN 'Nordeste' WHEN AVG(dv_50m) > 67.5 AND AVG(dv_50m) <= 112.5 THEN 'Leste' WHEN AVG(dv_50m) > 112.5 AND AVG(dv_50m) <= 157.5 THEN 'Sudeste' WHEN AVG(dv_50m) > 157.5 AND AVG(dv_50m) <= 202.5 THEN 'Sul' WHEN AVG(dv_50m) > 202.5 AND AVG(dv_50m) <= 247.5 THEN 'Sudoeste' WHEN AVG(dv_50m) > 247.5 AND AVG(dv_50m) <= 294.5 THEN 'Oeste' WHEN AVG(dv_50m) > 294.5 AND AVG(dv_50m) < 337.5 THEN 'Noroeste' ELSE 'NULL' END AS dv_50m_nominal FROM parque_ceara

/*PROCESSAMENTO NOS DADOS */

WHERE cod_erro = '0' GROUP BY tres_dias,n_semana_mes, mes, ano ORDER BY ano,mes,n_semana_mes, tres_dias) TO 'C:/arquivo.csv' DELIMITER ',' CSV HEADER

APÊNDICE F – CÓDIGO NA LINGUAGEM SQL PARA EXPORTAÇÃO DOS DADOS SEMANAIS DO SEGUNDO ESTUDO DE CASO

```
/*COMANDO COPY PARA GERAR UM ARQUIVO CSV */

COPY(

/*GENERALIZACAO DO ANO NA FORMA NUMERICA */

SELECT CAST(substring(ano_mes_dia from 1 for 4) AS int) as ano,

/*GENERALIZACAO DO MES NA FORMA NUMERICA */

CAST(substring(ano_mes_dia from 5 for 2)AS INT) as mes,

/*GENERALIZACAO DO MES NA FORMA NOMINAL */

CASE WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 1 THEN 'Janeiro' /*31
dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 2 THEN 'Fevereiro' /*31
dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 3 THEN 'Marco' /*31
dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 4 THEN 'Abril' /*31
dias*/ WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 5 THEN 'Maio' /*31 dias*/
WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 6 THEN 'Junho' /*31 dias*/
WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 7 THEN 'Julho' /*31 dias*/
WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 8 THEN 'Agosto' /*31 dias*/
WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 9 THEN 'Setembro' /*31 dias*/
WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 10 THEN 'Outubro' /*31 dias*/
WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 11 THEN 'Novembro' /*31 dias*/
WHEN CAST(substring(ano_mes_dia from 5 for 2) AS int) = 12 THEN 'Dezembro' /*31 dias*/
ELSE 'NULL' END AS mes_nominal,

/*GENERALIZACAO DA SEMANA DO MES NA FORMA NUMERICA */

CASE WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) < 8 THEN
'1' WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) < 16 THEN '2'
WHEN CAST(substring(ano_mes_dia from 7 for 2)AS INT) < 24 THEN '3' WHEN
CAST(substring(ano_mes_dia from 7 for 2)AS INT) <= 31 THEN '4' ELSE 'null' END AS
n_semana_mes,

/*NORMALIZACAO PARA DUAS CASAS APOS A VIRGULA*/
```

```
ROUND(AVG(temperatura),2) as temperatura, ROUND(AVG(umidade_relativa),2) as umidade_relativa, ROUND(AVG(pressao_atmosferica),2) as pressao_atmosferica, ROUND(AVG(vv_50m),2) as vv_50m, ROUND(AVG(dv_50m),2) as dv_50,

/*GENERALIZACAO DA DIRECAO DO VENTO NA FORMA NOMINAL*/

CASE WHEN AVG(dv_50m) > 337.5 OR AVG(dv_50m) <= 22.5 THEN 'Norte' WHEN AVG(dv_50m) > 22.5 AND AVG(dv_50m) <= 67.5 THEN 'Nordeste' WHEN AVG(dv_50m) > 67.5 AND AVG(dv_50m) <= 112.5 THEN 'Leste' WHEN AVG(dv_50m) > 112.5 AND AVG(dv_50m) <= 157.5 THEN 'Sudeste' WHEN AVG(dv_50m) > 157.5 AND AVG(dv_50m) <= 202.5 THEN 'Sul' WHEN AVG(dv_50m) > 202.5 AND AVG(dv_50m) <= 247.5 THEN 'Sudoeste' WHEN AVG(dv_50m) > 247.5 AND AVG(dv_50m) <= 294.5 THEN 'Oeste' WHEN AVG(dv_50m) > 294.5 AND AVG(dv_50m) < 337.5 THEN 'Noroeste' ELSE 'NULL' END AS dv_50m_nominal FROM parque_ceara

/*PROCESSAMENTO NOS DADOS */

WHERE cod_erro = '0' GROUP BY n_semana_mes, mes, ano ORDER BY ano,mes,n_semana_mes ) TO 'C:/arquivo.csv' DELIMITER ',' CSV HEADER
```