

UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE – UERN

FACULDADE DE CIÊNCIAS EXATAS E NATURAIS – FANAT

DEPARTAMENTO DE INFORMÁTICA – DI

ÁTILLA NEGREIROS MAIA

**MINERAÇÃO DE DADOS PARA ESTIMATIVA DA RADIAÇÃO SOLAR
GLOBAL**

MOSSORÓ - RN

2017

ÁTILLA NEGREIROS MAIA

**MINERAÇÃO DE DADOS PARA ESTIMATIVA DA RADIAÇÃO SOLAR
GLOBAL**

Monografia apresentada à Universidade do Estado do Rio Grande do Norte como um dos pré-requisitos para obtenção do grau de bacharel em Ciência da Computação, sob orientação do Prof. Dr. Marcelino Pereira dos Santos Silva.

MOSSORÓ - RN

2017

Ficha catalográfica gerada pelo Sistema Integrado de Bibliotecas
e Diretoria de Informatização (DINF) - UERN,
com os dados fornecidos pelo(a) autor(a)

M217m Maia, Átilla Negreiros.
 MINERAÇÃO DE DADOS PARA ESTIMATIVA DA RADIAÇÃO SOLAR
 GLOBAL / Átilla Negreiros Maia - 2017.
 54 p.

 Orientador: Marcelino Pereira dos Santos Silva.
 Coorientador: .
 Monografia (Graduação) - Universidade do Estado do Rio Grande do
 Norte, Ciência da Computação, 2017.

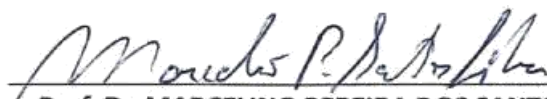
 1. Banco de dados . 2. Mineração de dados. 3. Radiação solar global. 4.
 Aprendizado de máquina. I. Silva, Marcelino Pereira dos Santos ,
 orient. II. Título.

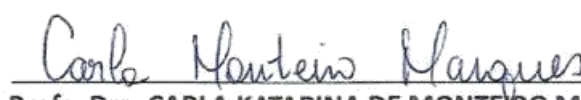
MINERAÇÃO DE DADOS PARA ESTIMATIVA DA RADIAÇÃO SOLAR GLOBAL

Monografia apresentada como pré-requisito para a obtenção do título de Bacharel em Ciência da Computação da Universidade do Estado do Rio Grande do Norte – UERN, submetida à aprovação da banca examinadora composta pelos seguintes membros:

Aprovada em: 26/04/2017.

Banca Examinadora


Prof. Dr. MARCELINO PEREIRA DOS SANTOS SILVA
Universidade do Estado do Rio Grande do Norte - UERN


Profa. Dra. CARLA KATARINA DE MONTEIRO MARQUES
Universidade do Estado do Rio Grande do Norte - UERN


Profa. Ma. KARLA HARYANNA SANTOS MOURA
Instituto Federal do Rio Grande do Norte - IFRN

À minha família.

AGRADECIMENTOS

Primeiramente a Deus por ter me dado força para superar as dificuldades e permitir que atingisse este objetivo.

Aos meus pais, Rui Charles Maia Mendes e Liduina Costa Negreiros, e minha tia, Laurenny Costa Negreiros, pelo amor, incentivo e por estarem ao meu lado em todos os momentos.

Aos meus avôs, Manoel Mendes e Lauro Mariano, e minhas avós, Maria Clivanilde e Maria Nilda (*in memoriam*), pelos ensinamentos constantes, apoio e sonho de ver-me concluir o ensino superior.

À minha namorada, Ingrid Queiroz de Miranda, pelo amor, compreensão, companheirismo, carinho, motivação e por sua presença nos momentos em que mais necessitei.

Ao professor orientador, Marcelino Pereira dos Santos Silva, pela dedicação, compreensão, disponibilidade e orientação, fundamental para a conclusão deste trabalho. Agradeço por sua amizade, incentivo e conselhos, os quais encorajaram-me a participar do Ciência Sem Fronteiras.

Ao mestrando e amigo, Nicksson Ckayo Arrais de Freitas, pelo apoio e colaboração ao longo desses meses.

A todos os professores e integrantes do Departamento de Informática, pela dedicação e por proporcionar-me o conhecimento que tenho hoje. Agradeço, em especial, a professora Cicilia Raquel Maia Leite pela amizade, disponibilidade e orientação ao longo da minha vida acadêmica.

A todos da Universidade do Estado do Rio Grande do Norte que colaboraram para a minha formação acadêmica. Agradeço, em especial, a DAINTE pelo esforço e apoio durante o processo de seleção do Ciência Sem Fronteiras.

Aos meus amigos e colegas de turma, pela colaboração, incentivo e contribuição para a minha formação acadêmica.

Agradeço a todos que me deram caronas ao longo desses anos, proporcionando minha presença nas aulas.

“As grandes ideias surgem da observação dos pequenos detalhes.”

(Augusto Cury)

RESUMO

Informações provenientes da estimativa da radiação solar são cruciais para o aproveitamento de sistemas solares, que são alternativas para complementar a matriz energética brasileira, fortemente concentrada em recursos hídricos. Apesar da disponibilidade de imensos volumes de bancos de dados meteorológicos e radiométricos, a análise manual desses dados ultrapassa os limites da capacidade humana de interpretação, tornando-a uma tarefa inviável e ineficiente. A tecnologia de mineração de dados constitui-se numa solução para analisar o grande volume de informações, através de técnicas e métodos que viabilizam a extração de padrões relevantes de grandes repositórios de dados. Estes padrões podem auxiliar na tomada de decisão e compreensão do domínio do problema. O objetivo deste trabalho é minerar dados meteorológicos e radiométricos para estimar a radiação solar global. Ferramentas de aprendizado de máquina são empregadas para classificar e validar os conjuntos de dados. Os resultados mostraram-se relevantes, apontando para trabalhos futuros promissores neste domínio tão estratégico.

Palavras-chave: Banco de dados. Mineração de dados. Radiação solar global. Aprendizado de máquina.

ABSTRACT

Information from estimation of solar radiation is crucial for the use of solar systems, which are alternatives to complement the Brazilian energy matrix, strongly focused on water resources. Besides the availability of huge volumes of meteorological and radiometric databases, the manual analysis these data surpasses the limits of the human ability of interpretation, making it an infeasible and inefficient task. Data mining technology consists in a solution to analyze the large volume of information through techniques and methods that provide the extraction of relevant patterns from large data repositories. These patterns can aid in decision making and understanding of the problem domain. The aim of this work is to mine meteorological and radiometric data to estimate global solar radiation. Machine learning tools are applied to classify and validate the data sets. The results were relevant, pointing towards promising future works in this strategic area.

Keywords: Database. Data mining. Global solar radiation. Machine learning.

LISTA DE FIGURAS

Figura 2.1 – Etapas do processo de KDD	18
Figura 2.2 – Técnicas de diferentes áreas utilizadas na mineração de dados	20
Figura 2.3 – Modelos básicos de mineração de dados	20
Figura 2.4 – Exemplo geral de uma árvore de decisão	22
Figura 2.5 – Exemplo de um conjunto de dados de treinamento	25
Figura 2.6 – Árvores geradas para o exemplo (a) apresenta a árvore momentânea (b) representa a árvore de decisão do exemplo extraída pelo WEKA.....	28
Figura 2.7 – Processos de interação entre a radiação solar e os constituintes atmosféricos	30
Figura 2.8 – Média anual do total diário de radiação solar global no Brasil.....	31
Figura 3.1 – Metodologia do trabalho	34
Figura 3.2 – Distribuição das estações da rede SONDA.....	35
Figura 3.3 – Classificação das variáveis medidas pelas estações da rede SONDA .	36
Figura 3.4 – Resultado da validação dos dados radiométricos e meteorológicos da estação de Petrolina para o ano de 2015.....	37
Figura 3.5 – Representação da radiação solar global em classes para os dados de 1 minuto gerada pela ferramenta WEKA.....	39
Figura 3.6 – Relação entre a radiação solar global e as variáveis candidatas a preditores para a base de dados de 1 minuto gerada pela ferramenta WEKA.....	41
Figura 3.7 – Relação entre a radiação solar global e as variáveis candidatas a preditores para a base de dados de 30 minutos gerada pela ferramenta WEKA	42

LISTA DE TABELAS

Tabela 4.1 – Resultados dos modelos para estimativa da radiação solar global para 1 minuto com uma variável de entrada.....	44
Tabela 4.2 – Resultados dos modelos para estimativa da radiação solar global para 30 minutos com uma variável de entrada.....	44
Tabela 4.3 – Resultados dos modelos para estimativa da radiação solar global para 1 minuto com algumas variáveis de entrada	45
Tabela 4.4 – Resultados dos modelos para estimativa da radiação solar global para 30 minutos com algumas variáveis de entrada	46
Tabela 4.5 – Resultados dos modelos para estimativa da radiação solar global com mês e iluminância como variáveis de entrada.....	47

LISTA DE ABREVIATURAS E SIGLAS

ARMA	<i>Autoregressive Moving Average</i>
BSRN	<i>Baseline Surface Radiation Network</i>
CART	<i>Classification and Regression Trees</i>
CSP	<i>Concentrating Solar Power</i>
INPE	Instituto Nacional de Pesquisas Espaciais
EM	<i>Expectation Maximization</i>
FUNCEME	Fundação Cearense de Meteorologia e Recursos Hídricos
KDD	<i>Knowledge Discovery in Databases</i>
kNN	<i>k Nearest Neighbor</i>
MDL	<i>Minimum Description Length</i>
PCD	Plataforma de Coleta de Dados
SONDA	Sistema de Organização Nacional de Dados Ambientais
SVM	<i>Support Vector Machine</i>
TDIDT	<i>Top-Down Induction of Decision Tree</i>
WEKA	<i>Waikato Environment for Knowledge Analysis</i>

SUMÁRIO

1 INTRODUÇÃO	14
1.1 MOTIVAÇÃO	14
1.2 OBJETIVOS	15
1.2.1 Objetivo geral	15
1.2.2 Objetivos específicos	15
1.3 ORGANIZAÇÃO DO TRABALHO	15
2 FUNDAMENTAÇÃO TEÓRICA	17
2.1 DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS	17
2.2 MINERAÇÃO DE DADOS.....	19
2.2.1 Modelos descritivos	21
2.2.2 Modelos preditivos	22
2.3 WEKA	23
2.3.1 Algoritmo C4.5	24
2.4 RADIAÇÃO SOLAR	28
2.4.1 Radiação solar global	30
2.5 TRABALHOS RELACIONADOS	32
3 METODOLOGIA DE MINERAÇÃO DOS DADOS	34
3.1 LEVANTAMENTO DOS DADOS	34
3.2 PRÉ-PROCESSAMENTO E TRANSFORMAÇÃO DOS DADOS	38
3.3 DEFINIÇÃO DOS ATRIBUTOS	40
3.4 CLASSIFICAÇÃO	40
3.5 AVALIAÇÃO DO MODELO	43
4 RESULTADOS E DISCUSSÕES	44
5 CONSIDERAÇÕES FINAIS	48
REFERÊNCIAS	49
APÊNDICE A – SELEÇÃO DOS DADOS VÁLIDOS	52
ANEXO A – FLUXOGRAMA DO PROCESSO DE VALIDAÇÃO PARA DADOS ANEMOMÉTRICOS E METEOROLÓGICOS	53

ANEXO B – FLUXOGRAMA DO PROCESSO DE VALIDAÇÃO PARA DADOS SOLARIMÉTRICOS	54
--	-----------

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

A demanda de energia elétrica e a necessidade de independência e diversificação energética, devido ao crescimento populacional e os impactos ambientais consequentes do uso de combustíveis fósseis, respectivamente, têm fomentado o desenvolvimento de estratégias para produção de energia limpa proveniente de fontes renováveis, como a energia solar.

O conhecimento a respeito da previsão da radiação solar é imprescindível para o setor energético, pois viabiliza orientações para a instalação, o planejamento e o gerenciamento dos recursos e sistemas solares, como o fotovoltaico e o heliotérmico.

No entanto, a quantidade insuficiente de informações em relação à disponibilidade e variabilidade da radiação solar é uma das grandes barreiras para a exploração do potencial elétrico da energia solar. Diversas bases de dados armazenam e disponibilizam grandes volumes de dados meteorológicos e radiométricos, porém a análise desses dados sem a utilização de técnicas computacionais avançadas é uma atividade praticamente inexecutável para o ser humano.

Nesse contexto, técnicas e métodos de mineração de dados possibilitam extrair padrões relevantes de imensos repositórios de dados com o objetivo de auxiliar especialistas e profissionais da área na tomada de decisão e compreensão do domínio. A mineração de dados consiste de etapas interrelacionadas e recorrentes, com aplicações de técnicas, algoritmos e métodos de diversas áreas do conhecimento (estatística, inteligência artificial, aprendizado de máquina, entre outras) para execução do pré-processamento, transformação, previsão e/ou descrição dos dados, e validação e avaliação dos resultados obtidos.

A previsão da radiação solar global é um tema ascendente e promissor no cenário nacional, tendo em vista a abundância da energia solar (em forma de luz e calor) e as vulnerabilidades apresentadas pela dependência de recursos hídricos. A matriz energética brasileira é contemplada, em sua grande maioria, por fontes de energias renováveis, contudo, é predominantemente hidráulica. As estiagens que o Brasil vem sofrendo ao longo dos anos têm alertado o setor energético, tais como a

crise hídrica na região Sudeste em 2014 que motivou especulações de um novo racionamento de energia (treze anos após a maior crise energética do país), e a seca prolongada no Nordeste que ameaça a produção de energia elétrica na região.

1.2 OBJETIVOS

1.2.1 Objetivo geral

Este trabalho tem a finalidade de desenvolver modelos de estimativa da radiação solar global a partir da mineração de dados meteorológicos e radiométricos.

1.2.2 Objetivos específicos

- Identificar bases de dados pertinentes e consistentes para o domínio do problema;
- Contatar órgãos do setor meteorológico;
- Efetuar o pré-processamento e a transformação dos dados;
- Realizar a análise do conjunto de variáveis de entrada;
- Validar e avaliar os resultados obtidos;
- Viabilizar informações de cunho científico para o setor meteorológico e/ou energético brasileiro.

1.3 ORGANIZAÇÃO DO TRABALHO

O trabalho está estruturado em 5 capítulos: Introdução, Fundamentação Teórica, Metodologia, Resultados e Discussões e Considerações Finais.

O capítulo 2 descreve os assuntos pertinentes para o entendimento deste trabalho: Descoberta de Conhecimento em Banco de Dados, Mineração de Dados, WEKA, Radiação Solar e alguns Trabalhos Relacionados.

O capítulo 3 define as etapas utilizadas para o desenvolvimento deste trabalho: Levantamento dos Dados, Pré-processamento e Transformação dos Dados, Definição dos Atributos, Classificação e Avaliação do Modelo.

No capítulo 4 são apresentados os resultados e análises obtidas pela mineração dos dados.

O capítulo 5 apresenta as conclusões e propostas de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta uma breve explicação dos assuntos relevantes para o desenvolvimento deste trabalho, que são: Descoberta de Conhecimento em Banco de Dados, mineração de dados, WEKA, radiação solar e alguns trabalhos relacionados.

2.1 DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS

A informatização da sociedade e o desenvolvimento constante de novas ferramentas de coleta e armazenamento de dados têm proporcionado o crescimento acelerado e a disponibilidade de imensos volumes de dados, de modo a ultrapassar os limites da compreensão humana na interpretação destes dados sem o auxílio de mecanismos computacionais, caracterizando um contexto com muitos dados, mas com pouca informação (do inglês, “*data rich but information poor*”) (HAN; KAMBER; PEI, 2011).

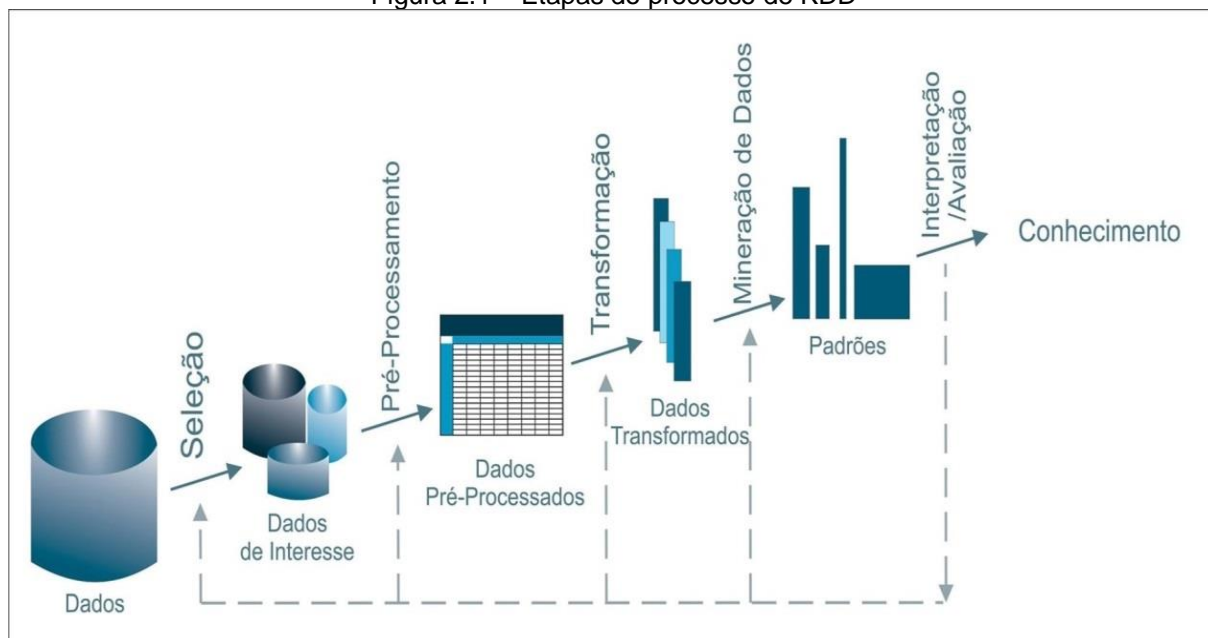
A geração de novas teorias e ferramentas computacionais necessárias para auxiliar os seres humanos no processo de extração de informações (ou conhecimento) úteis dos repositórios de dados é objeto de estudo do *Knowledge Discovery in Databases* (KDD). Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996, p.40), KDD é o processo não trivial de identificar, em dados, padrões que sejam válidos, novos (previamente desconhecidos), potencialmente úteis, e compreensíveis, visando melhorar o entendimento de um problema ou um procedimento de tomada de decisão.

Nesse cenário, a mineração de dados desempenha um papel fundamental devido à capacidade de transformar dados em conhecimento útil, possibilitando assim auxiliar no processo de tomada de decisão em diversas aplicações: análise de mercado, diagnósticos mais precisos de doenças, detecção de fraude, pesquisas biométricas, entre outras.

O termo KDD é usado por alguns autores (por exemplo, Han, Kamber e Pei (2011)) como sinônimo para o termo mineração de dados, entretanto outros autores (Fayyad, Piatetsky-Shapiro e Smyth (1996) e Maimon e Rokach (2010)) consideram que mineração de dados é apenas uma etapa do processo de KDD.

O KDD é um processo com etapas interativas e iterativas, conforme ilustra a Figura 2.1, de modo que pode-se retornar às etapas anteriores à medida que for necessário.

Figura 2.1 – Etapas do processo de KDD



Fonte: Adaptada de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Inicialmente, é desenvolvida uma compreensão do domínio da aplicação, definindo o conhecimento a ser descoberto e as decisões a serem tomadas nas etapas posteriores.

A etapa de Seleção consiste em selecionar os dados de interesse (ou dados relevantes) disponíveis no repositório de dados. Além da grande variedade, os dados podem ter vários formatos, como texto, planilha, imagem, por exemplo.

A etapa de Pré-processamento dos Dados trata da qualidade destes dados, garantindo fatores como exatidão, integridade, consistência, prontidão, credibilidade e interpretabilidade. A limpeza dos dados é realizada, de forma que inconsistências e ruídos sejam removidos ou corrigidos e dados ausentes ignorados ou preenchidos manualmente (por meio de uma constante global ou pela média do atributo, por exemplo). Outras fontes de dados podem ser integradas – observando as redundâncias e dependências das variáveis – mantendo um repositório único e consistente (HAN; KAMBER; PEI, 2011).

Após o pré-processamento, os dados são transformados em formatos apropriados para a mineração. A transformação dos dados pode ser executada, por

exemplo, por normalização e discretização. Na normalização, os dados de um atributo são dimensionados na mesma escala, por exemplo, com valores entre 0 e 1 ou -1 e 1. Já na discretização, os dados correspondentes a atributos numéricos, como temperatura, por exemplo, são substituídos por valores nominais, como a identificação de intervalos (10–20, 20–30, 30–40) ou conceitos (baixa, média, alta).

As etapas de Pré-processamento e Transformação – consideradas para alguns autores como uma única etapa – demandam a maior parte do tempo do processo de KDD, cerca de 70% (SILVA, 2004).

A Mineração de Dados é o núcleo do processo de KDD, onde definem-se os métodos para explorar os dados, desenvolve-se o modelo para compreensão de fenômenos a partir dos dados, análise e previsão, e descobrem-se padrões previamente desconhecidos (MAIMON; ROKACH, 2010).

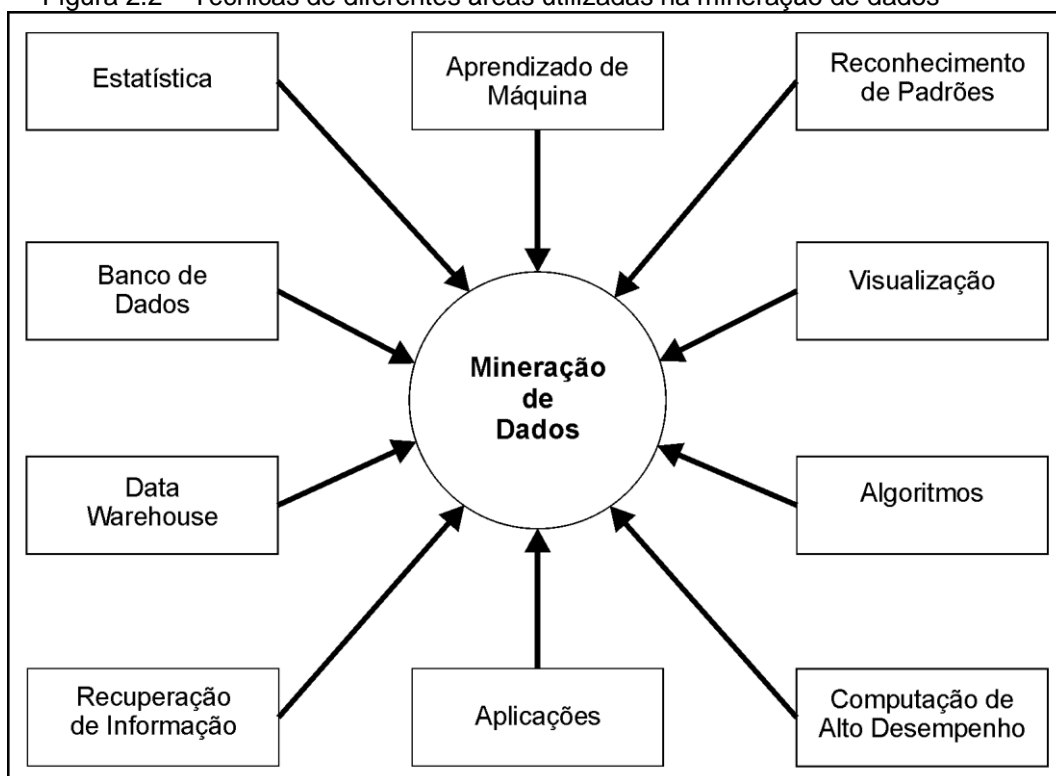
Com a aquisição dos padrões, a Interpretação/Avaliação identifica aqueles interessantes que representam o conhecimento útil. Para Han, Kamber e Pei (2011), um padrão é considerado interessante caso seja: compreendido por humanos, válido com algum grau de certeza, potencialmente útil, e novo (pelo menos para o sistema) – ao encontro da definição utilizada por Fayyad, Piatetsky-Shapiro e Smyth (1996).

O conhecimento descoberto pode ser incorporado a outro sistema para ações adicionais (tomada de decisão, previsão, diagnóstico, etc), ou documentado e reportado para as partes interessadas.

2.2 MINERAÇÃO DE DADOS

A mineração de dados é um processo de natureza interdisciplinar que tem utilizado e incorporado técnicas e algoritmos de diferentes domínios do conhecimento, tais como estatística, aprendizado de máquina, reconhecimento de padrões, sistemas de banco de dados, *data warehouse*, recuperação de informação, visualização, algoritmos e computação de alto desempenho, conforme ilustra a Figura 2.2 (HAN; KAMBER; PEI, 2011).

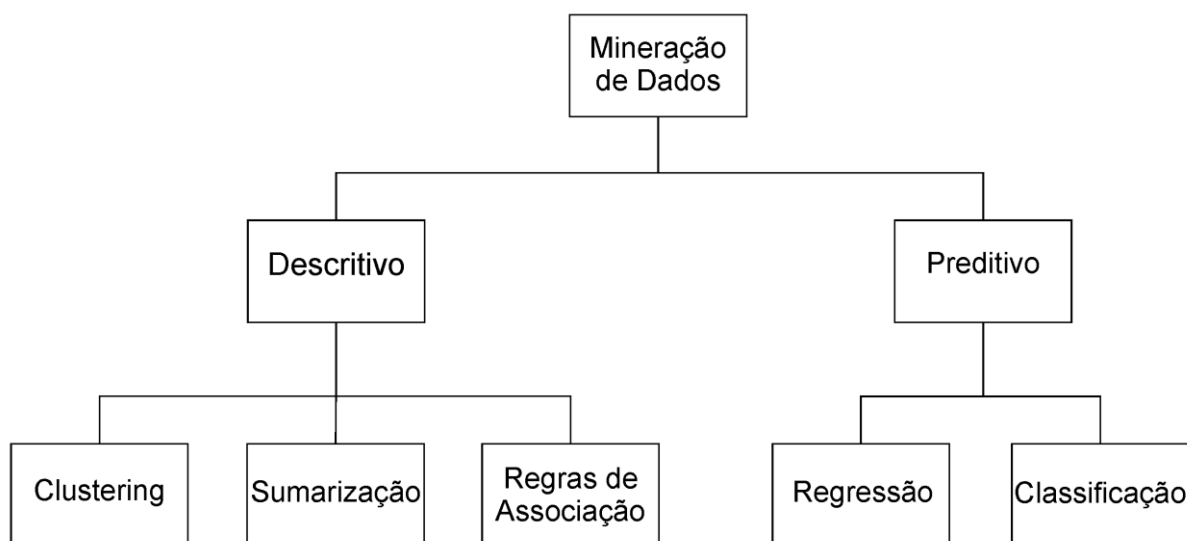
Figura 2.2 – Técnicas de diferentes áreas utilizadas na mineração de dados



Fonte: Adaptada de Han, Kamber e Pei (2011).

Os métodos de mineração de dados apresentam diferentes propósitos e objetivos, e são, tradicionalmente, divididos em modelos descritivos e modelos preditivos, demonstrados na Figura 2.3 (MAIMON; ROKACH, 2010).

Figura 2.3 – Modelos básicos de mineração de dados



Fonte: Adaptado de Maimon e Rokach (2010)

2.2.1 Modelos descritivos

Os modelos descritivos, também conhecidos como aprendizagem não-supervisionada, tais como *clustering*, sumarização e regras de associação, identificam os padrões ou relações nos dados analisando suas propriedades, pois o rótulo da classe de cada amostra de treinamento não é conhecida (SILVA, 2004).

Clustering é uma técnica usada para agrupar dados que apresentam similaridades. Os *clusters* devem ser homogêneos em si e heterogêneos entre si. Em outras palavras, um conjunto de registros que são semelhantes entre si (localidade, cor, tamanho, idade, etc) e diferentes dos registros em outros *clusters* (LAROSE, 2014). Por exemplo, uma rede de supermercados pode realizar uma análise de mercado através de informações geográficas e de estilo de vida dos seus clientes e formar grupos com intuito de oferecer serviços exclusivos.

A sumarização consiste em descrições compactas de um subconjunto de dados. Segundo Maimon e Rokach (2010), a sumarização é um tipo especial de *clustering* que fornece uma visão de alto nível dos dados a partir de subconjuntos com simples descrições associadas (textual ou gráfica). O levantamento do perfil dos clientes (entre 30 e 40 anos, empregado, renda superior a dois salários mínimos, etc) que frequentam diariamente um determinado supermercado é um exemplo da aplicação desse método.

As regras de associação buscam encontrar os atributos que estão frequentemente relacionados. Elas são baseadas em inferências como “SE *antecedente*, ENTÃO *consequente*”, juntamente com uma medida de confiança e suporte associada à regra. A medida de confiança refere-se à porcentagem das transações em um conjunto de dados contendo o *antecedente* que também contêm o *consequente*, em outras palavras, a probabilidade condicional do *consequente* em relação ao *antecedente*. E a medida de suporte refere-se à porcentagem das transações em um conjunto de dados para os quais a regra é encontrada. Supondo que o gerente de um supermercado deseja saber os itens que são frequentemente comprados juntos pelos clientes e a seguinte regra de associação, por exemplo, é expressa: “SE compra leite, ENTÃO compra pão” [suporte=3%, confiança=75%]. Essa regra indica que 75% dos clientes que compram leite também compram pão e 3% das transações em análise mostram que leite e pão são comprados juntos (LAROSE, 2014; WITTEN et al., 2016).

2.2.2 Modelos preditivos

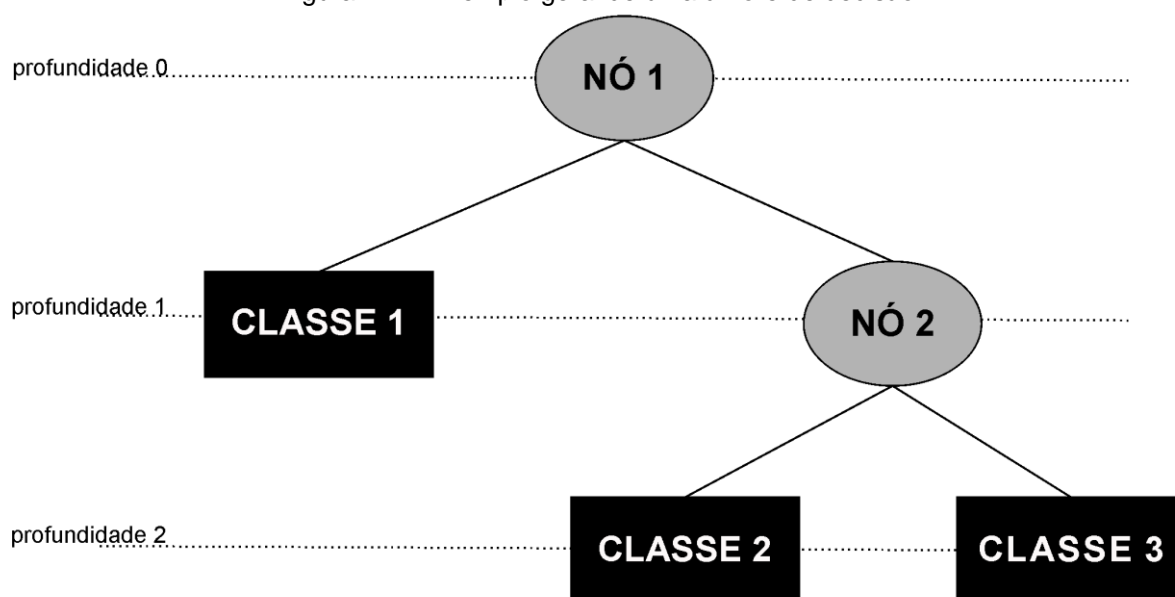
Os modelos preditivos ou aprendizagem supervisionada, como são conhecidos na comunidade de aprendizado de máquina, como regressão e classificação, são utilizados para realizar previsões a respeito de dados desconhecidos utilizando valores conhecidos. Esses modelos desempenham essa tarefa a partir da análise de um conjunto de dados de treinamento.

Regressão é um modelo de predição que identifica relações entre atributos com valores de dados numéricos. Quando essa relação é expressa por uma função linear o modelo de regressão é definido como linear, caso contrário, como não linear.

A classificação é similar ao modelo de regressão, exceto que a previsão é realizada para valores nominais ao invés de numéricos. Um modelo bastante utilizado na classificação é a árvore de decisão.

Árvore de decisão é um conjunto de nós de decisão conectados por arestas, de modo que os nós internos correspondam a testes (particionando os dados em subconjuntos). As arestas representam os resultados dos testes, e os nós terminais (ou nós “folhas”) denotam as classes dos dados. A “raiz” consiste no nó mais alto da árvore (Figura 2.4).

Figura 2.4 – Exemplo geral de uma árvore de decisão



Fonte: Adaptado de Barros et al. (2013)

A visualização gráfica e a possibilidade de conversão em regras de associação contribuem para que as árvores de decisão sejam facilmente compreendidas. Essa facilidade de compreensão e interpretação tornam as árvores de decisão uma alternativa natural para outros modelos bem conhecidos, tais como redes neurais e *Support Vector Machines* (SVM). Além disso, a capacidade de lidar com dados multidimensionais e redundantes, robustez na presença de ruído e algoritmos de indução (aprendizagem) com baixo custo computacional são características que permitem a disseminação das árvores, tornado uma das técnicas mais utilizadas na mineração de dados (BARROS et al., 2013).

Na literatura, os algoritmos de aprendizagem de árvores de decisão ID3 e C4.5, desenvolvidos por Quinlan (1986, 1996), e o *Classification and Regression Trees* (CART), desenvolvido por um grupo de estatísticos (L. Breiman, J. Friedman, R. Olshen e C. Stone), são considerados referências para diversos trabalhos relacionados a indução de árvores de decisão. (LAROSE, 2014; HAN; KAMBER; PEI, 2011; WITTEN et al., 2016).

2.3 WEKA

A ferramenta, ou *toolkit*, de mineração de dados intitulada WEKA (do inglês *Waikato Environment for Knowledge Analysis*) foi desenvolvida na linguagem Java na Universidade de Waikato, Nova Zelândia, e contém uma série de algoritmos de mineração e análise de dados. WEKA possui código aberto sob a licença GNU *General Public License* e está disponível livremente em (WEKA, 2017).

O desenvolvimento do projeto teve início em 1993 com a finalidade de investigar as técnicas de aprendizado de máquina em áreas-chaves da economia neozelandesa, especificamente a agricultura, e teve sua primeira versão (v 2.1) publicada em 1996. Os componentes de software do WEKA resultam, em grande parte, de teses e dissertações de grupos de pesquisa da Universidade de Waikato (SILVA, 2004).

WEKA dispõe de uma coleção de algoritmos e ferramentas para pré-processamento e transformação dos dados, classificação, *clustering*, regras de associação, regressão e visualização, a qual também fornece algoritmos de validação de resultados.

A ferramenta possui características como: interface gráfica amigável e de fácil manuseio nas aplicações, inclusive por pessoas que não são especialistas em mineração de dados; ampla portabilidade, ou seja, está disponível em quase todas as plataformas de computação, e mantém-se como um sistema atualizado, onde novos algoritmos são adicionados à medida que surgem na literatura. Esses elementos e a sua gratuidade tornaram WEKA uma ferramenta robusta, sendo uma das mais utilizadas para o processo de mineração de dados.

Dentre os algoritmos de classificação disponíveis no WEKA, o presente trabalho emprega o J48, uma implementação do algoritmo C4.5, referência em abordagens de classificação conforme citado na seção 2.2.2. O J48 implementa a versão C4.5 lançamento 8, última versão antes da versão C5.0 (comercial) (WITTEN et al., 2016).

2.3.1 Algoritmo C4.5

O algoritmo C4.5 usa o método divisão e conquista (do inglês *divide and conquer*) em uma abordagem *top-down* de forma recursiva para construir uma árvore de decisão a partir de um conjunto de treinamento.

De acordo com Basgalupp (2010), o algoritmo adotado no C4.5, mais conhecido como *Top-Down Induction of Decision Tree* (TDIDT), é recursivo de busca gulosa e tenta encontrar a “melhor” forma de particionar o conjunto de treinamento. Esse conjunto passa a ser representado pela raiz da árvore; em seguida, é selecionado um atributo preditivo como teste desse nó, o qual dividirá os casos em subconjuntos. A busca permanece até classificar todos os casos ou utilizar todos os atributos preditivos.

O C4.5 apresenta três tipos de testes que podem ser aplicados aos atributos. Eles são utilizados conforme o tipo do atributo (discreto ou numérico):

- “ $A=?$ ”, onde A é discreto e com valores $\{a_1, a_2, \dots, a_n\}$ – o resultado corresponde aos valores de A , ou seja, uma aresta para cada $\{a_1, a_2, \dots, a_n\}$.
- “ $A \in D_A?$ ”, A é discreto e D_A é um subconjunto de A – o resultado consiste em subconjuntos dos valores de A . Uma aresta identificada como “verdadeiro” (por convenção, aresta esquerda), que corresponde

à partição de A que satisfaz o critério, e outra aresta como “falso” (aresta direita), correspondendo à partição de A que não satisfaz.

- “ $A \leq \theta$?”, A é um valor numérico e θ é um limite constante – o resultado é verdadeiro ou falso, ou seja, uma aresta para cada valor. Ao ordenar os valores do atributo, identifica-se um limite para cada par de valores.

Um critério de divisão (do inglês *splitting criterion*) é usado para encontrar o atributo que determina a “melhor” maneira de particionar o conjunto de treinamento. O *Gain Ratio* é o método empregado no C4.5. Esse método é uma evolução do critério de divisão Ganho de Informação (do inglês *Gain Information*) aplicado ao ID3, seu predecessor. Ambos baseiam-se no conceito de entropia da informação formulada por Claude Shannon.

A entropia, ou melhor, a menor quantidade de bits necessários para representar uma classe de um caso em D é definida pela Equação 2.1. Essa equação $p(D,i)$ refere-se à proporção de casos em D que pertence à classe i , e k corresponde ao número de classes.

$$\text{entropia}(D) = - \sum_{i=1}^k p(D,i) \log_2(p(D,i)) \quad (2.1)$$

Figura 2.5 – Exemplo de um conjunto de dados de treinamento

Dia	Aspecto	Temperatura	Umidade	Vento	Jogar Tênis
1	Ensolarado	Quente	Elevada	Falso	Não
2	Ensolarado	Quente	Elevada	Verdadeiro	Não
3	Ensolarado	Ameno	Elevada	Falso	Não
4	Ensolarado	Fresco	Normal	Falso	Sim
5	Ensolarado	Ameno	Normal	Falso	Sim
6	Nublado	Quente	Elevada	Falso	Sim
7	Nublado	Fresco	Normal	Verdadeiro	Sim
8	Nublado	Ameno	Elevada	Verdadeiro	Sim
9	Nublado	Quente	Normal	Falso	Sim
10	Chuvoso	Ameno	Elevada	Falso	Sim
11	Chuvoso	Fresco	Normal	Falso	Sim
12	Chuvoso	Fresco	Normal	Verdadeiro	Não
13	Chuvoso	Ameno	Normal	Falso	Sim
14	Chuvoso	Ameno	Elevada	Verdadeiro	Não

Fonte: Adaptado de Quinlan (1986)

Aplicando a Equação 2.1 no conjunto de dados de treinamento representado na Figura 2.5, a entropia da classe “Jogar Tênis”, cujos valores possíveis são “Sim” (9) e “Não” (5), totalizando 14 instâncias:

$$\text{entropia}(\text{JogarTenis}) = -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits}$$

Supondo um atributo preditivo T que ao particionar um conjunto D em $\{D_1, D_2, \dots, D_n\}$ demande uma quantidade de informação (calculada pela soma ponderada das entropias dos subconjuntos individuais), o Ganho de Informação (do inglês *Information Gain*) é definido como a quantidade de informação original subtraída da quantidade de informação exigida no particionamento, e é expresso na Equação 2.2.

$$\text{GanhoInfo}(D, T) = \text{entropia}(D) - \sum_{i=1}^n \left(\frac{|D_i|}{|D|}\right) \text{entropia}(D_i) \quad (2.2)$$

Dessa forma, aplicando a equação 2.2 para o exemplo “Jogar Tênis”, realiza-se o cálculo do *GanhoInfo* para cada atributo (“Aspecto”, “Temperatura”, “Umidade”, “Vento”). No atributo “Aspecto”, com possíveis valores “Ensolarado” {“Sim” (2), “Não” (3)}, “Nublado” {“Sim” (4), “Não” (0)} e “Chuvoso” {“Sim” (3), “Não” (2)}, tem-se:

$$\text{GanhoInfo}(\text{JogarTenis}, \text{Aspecto})$$

$$= 0.940$$

$$- \left(\frac{5}{14} \times \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right) + \frac{4}{14} \times \left(-\frac{4}{4} \log_2 \frac{4}{4} \right) + \frac{5}{14} \times \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) \right) = 0.246 \text{ bits}$$

$$\text{GanhoInfo}(\text{JogarTenis}, \text{Temperatura}) = 0.940 - 0.911 = 0.029 \text{ bits}$$

$$\text{GanhoInfo}(\text{JogarTenis}, \text{Umidade}) = 0.940 - 0.789 = 0.151 \text{ bits}$$

$$\text{GanhoInfo}(\text{JogarTenis}, \text{Vento}) = 0.048 \text{ bits}$$

Tendo em vista, que o Ganho de Informação favorece atributos com muitos valores (muitas arestas), podendo criar partições inúteis para a classificação. O C4.5

soluciona esse problema através do *split information*, que considera o potencial de informação da partição em si, e é expresso na Equação 2.3 (QUINLAN,1996).

$$SplitInfo(D, T) = - \sum_{i=1}^k \frac{|D_i|}{|D|} \log_2 \left(\frac{|D_i|}{|D|} \right) \quad (2.3)$$

Ou seja, para cada resultado do atributo T , é considerado o número de tuplas para o resultado (valor do atributo) em relação ao número de tuplas de D . Aplicando a Equação 2.3 ao exemplo tem-se:

$$SplitInfo(JogarTennis, Aspecto) = 1.577$$

$$SplitInfo(JogarTennis, Temperatura) = 1.557$$

$$SplitInfo(JogarTennis, Umidade) = 1$$

$$SplitInfo(JogarTennis, Vento) = 0.985$$

Por fim, o cálculo para o *Gain Ratio* é a razão entre o Ganho da Informação e o *split information*, definido pela Equação 2.4. O atributo com maior ganho é eleito como atributo de divisão para o respectivo nó.

$$GainRatio(D) = \frac{Info(D)}{SplitInfo(D, T)} \quad (2.4)$$

O *Gain Ratio* para o nosso caso resulta em:

$$GainRatio(JogarTennis, Aspecto) = 0.246/1.577 = 0.156 \text{ bits}$$

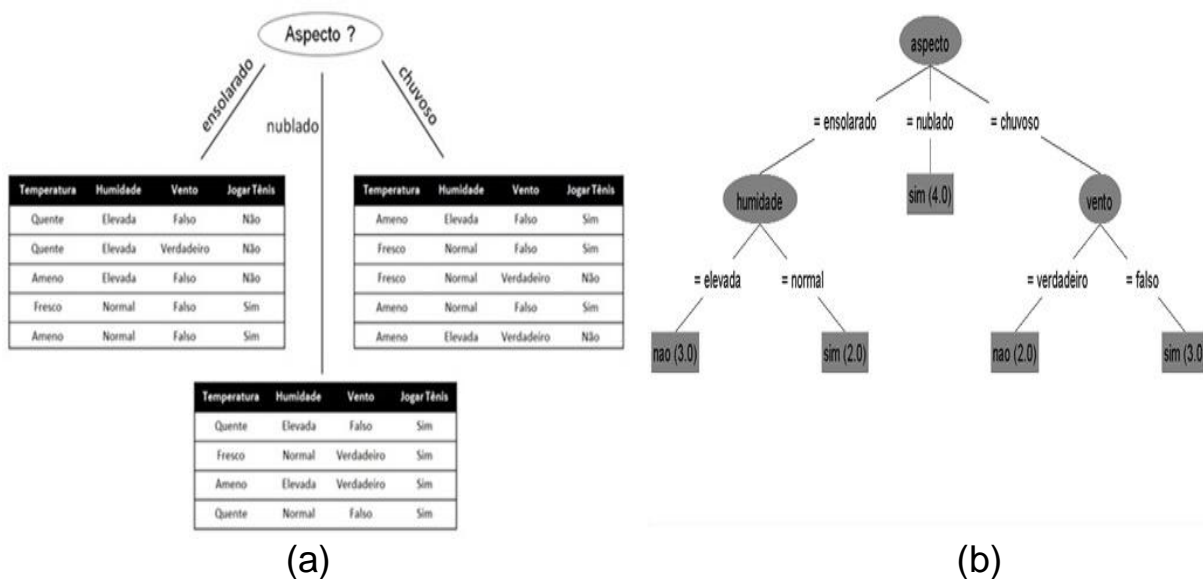
$$GainRatio(JogarTennis, Temperatura) = 0.246/1.557 = 0.02 \text{ bits}$$

$$GainRatio(JogarTennis, Umidade) = 0.151/1 = 0.151 \text{ bits}$$

$$GainRatio(JogarTennis, Vento) = 0.048/0.985 = 0.05 \text{ bits}$$

Portanto, o atributo selecionado é “Aspecto”, pois maximiza o *Gain Ratio*. Nesse ponto, a árvore de decisão, até o momento, é representada pela Figura 2.6a. Dando prosseguimento ao algoritmo, a árvore de decisão gerada é semelhante à árvore ilustrada na Figura 2.6b.

Figura 2.6 – Árvores geradas para o exemplo (a) apresenta a árvore momentânea (b) representa a árvore de decisão do exemplo extraída pelo WEKA



Fonte: Autoria própria

O C4.5 Lançamento 8, apresenta outro sutil ajuste. O Ganho de Informação para atributos numéricos foi modificado baseado no princípio *Minimum Description Length* (MDL) – modificando a Equação 2.2 – e está expresso pela Equação 2.5, em que S corresponde ao número de valores para o atributo T (numérico) e N refere-se ao número de instâncias no nó (WITTEN et al., 2016).

$$GanhoInfo(D, T) = \left(entropia(D) - \sum_{i=1}^n \left(\frac{|D_{i}|}{|D|} \right) entropia(D) \right) - \frac{\log_2(S)}{N} \quad (2.5)$$

2.4 RADIAÇÃO SOLAR

A energia solar é uma fonte de energia renovável promissora e considerada a melhor opção dentre as demais fontes renováveis, pois (1) é o recurso de energia mais abundante que alcança a Terra em forma de luz e calor, (2) é uma fonte inesgotável, (3) a produção de energia não causa danos ao ecossistema terrestre, e (4) um sistema de geração de energia solar pode ser aplicável para moradias e indústrias (KANNAN; VAKEESAN, 2016).

A radiação solar é a fonte primária, direta ou indireta, de quase toda a energia disponível na Terra. Energia hidráulica, biomassa, eólica, e combustíveis fósseis são

formas indiretas de energia solar. Na sua forma direta, a radiação solar pode ser utilizada como fonte de energia térmica e como energia elétrica, por meio do sistema heliotérmico e o fotovoltaico (PACHECO, 2006).

No sistema solar térmico, a energia solar é captada e transformada em calor através de coletores, também conhecidos como painéis solares. No painel, o calor da radiação solar é transferido para um líquido, geralmente água, armazenado no interior do painel, que será utilizado como fonte de calor. A energia térmica é normalmente utilizada nas aplicações para aquecimento de água.

No sistema heliotérmico ou energia solar térmica concentrada, do inglês *Concentrating Solar Power* (CSP), a radiação solar é convertida primeiramente em energia térmica para, em seguida, ser convertida em eletricidade. No caso da CSP, coletores e concentradores solares são usados para a produção do calor empregado no aquecimento do fluido térmico (óleos sintéticos, sal fundido ou vapor d'água), que será aplicado para girar uma turbina conectada a um gerador elétrico. Nesse ponto o processo é semelhante ao de uma usina termelétrica convencional (TOLMASQUIM, 2016).

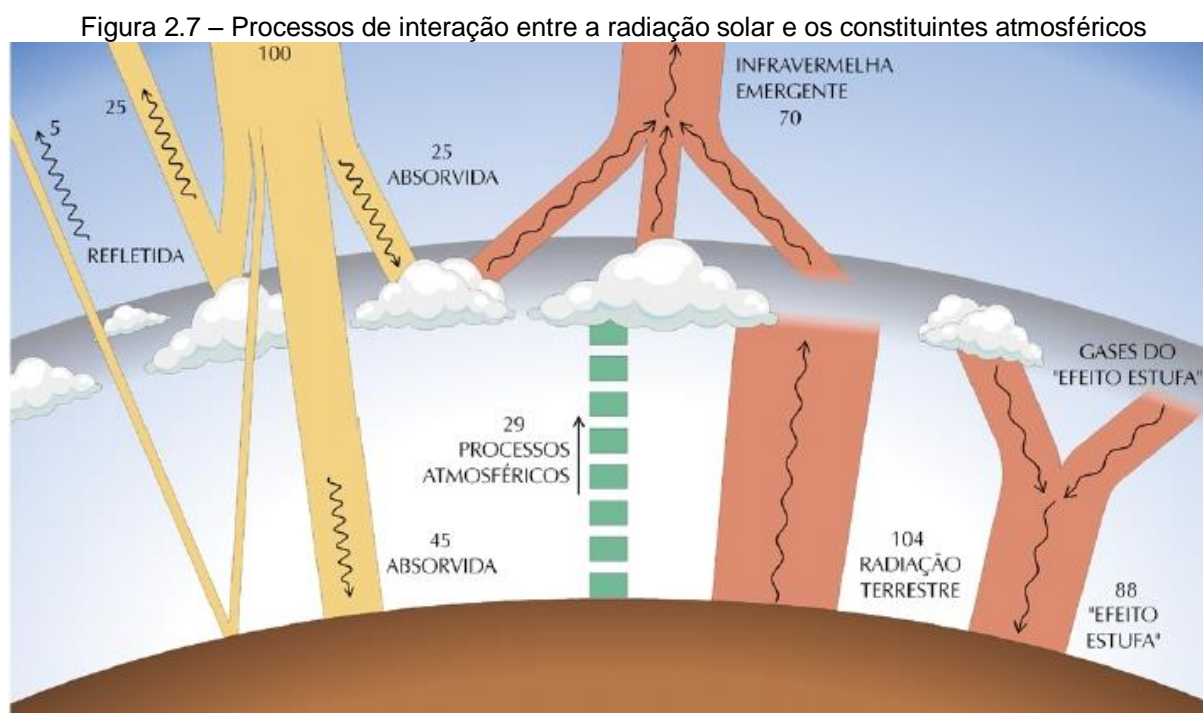
Já no sistema fotovoltaico, a transformação da radiação solar em eletricidade é direta através do efeito fotovoltaico. Um material semicondutor, sendo o silício o mais utilizado na fabricação das células fotovoltaicas, à medida que estimulado pela radiação, permite o fluxo eletrônico dando início ao fluxo de energia na forma de corrente contínua. Quanto maior a intensidade de luz, maior o fluxo de energia elétrica (AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA, 2008).

Considerando o sistema fotovoltaico e a CSP, o primeiro representa 98% da capacidade instalada entre as duas fontes. Apesar de a CSP ter a vantagem de funcionar com armazenamento ou com outra fonte de *back-up*, (permitindo sua operação depois que o Sol se põe) ele ainda permanece como uma das fontes renováveis mais caras, o que dificulta sua expansão (TOLMASQUIM, 2016).

A disponibilidade do recurso solar na superfície é um dos elementos principais para o desempenho economicamente viável dos sistemas de energia solar. Entretanto, o potencial de energia solar sofre variações ao longo do ano, influenciadas por vários fatores, como: movimentos astronômicos de rotação e translação, inclinação e orientação da superfície, atenuação ao atravessar a atmosfera terrestre e dispersão causada por nuvens (ANGELIS-DIMAKIS et al., 2011). Segundo Lima (2015), cerca de 25% da radiação solar incidente no topo da

atmosfera atingem a superfície terrestre sem sofrer alteração dos constituintes atmosféricos, enquanto o restante sofre os efeitos dos processos de absorção, reflexão e espalhamento.

A Figura 2.7 demonstra os processos de interação entre a radiação solar e a atmosfera terrestre.



Fonte: Pereira et al. (2006).

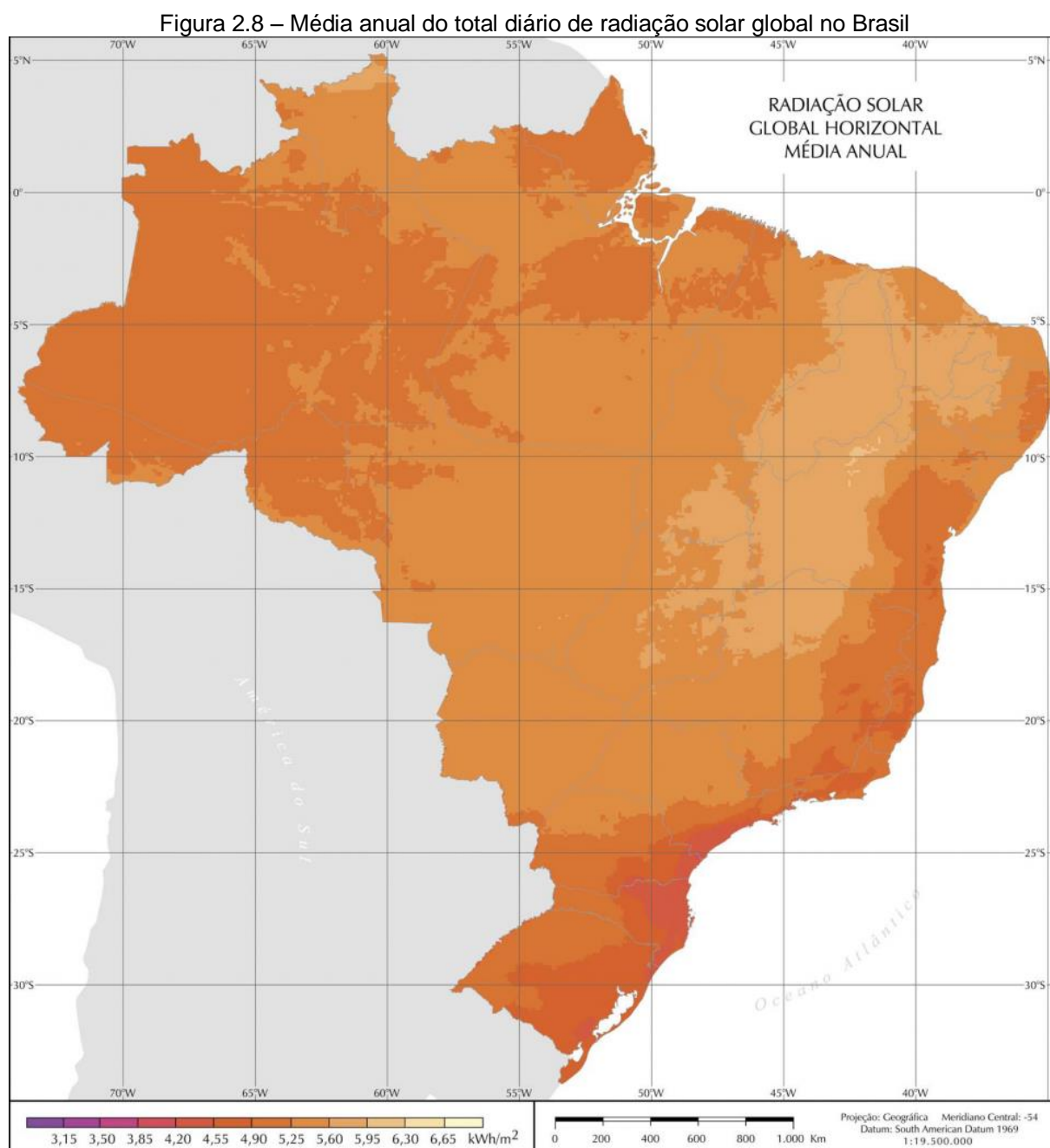
2.4.1 Radiação solar global

A radiação solar incidente na superfície terrestre pode ser decomposta em planos: horizontal e inclinado. Em ambos os planos, a radiação solar global é composta pela soma das componentes difusa e direta, porém no plano inclinado é acrescido o albedo, uma parcela refletida na superfície e nos elementos do entorno (TOLMASQUIM, 2016).

A radiação solar difusa refere-se a componente da radiação solar que sofre os efeitos de absorção, reflexão, dispersão e espalhamento, causados pelos constituintes atmosféricos e nuvens. Já a radiação solar direta caracteriza-se como a que atinge o solo diretamente sem sofrer os efeitos mencionados. Ainda de acordo com Tolmasquim (2016), essas componentes podem ser empregadas de forma distintas pelas tecnologias de sistemas solares: a irradiação solar global é de

bastante interesse para o aproveitamento fotovoltaico, enquanto a irradiação direta solar é relevante para o sistema heliotérmico, ou CSP.

Nesse ponto, as previsões da radiação solar global são informações cruciais para o monitoramento eficiente do potencial elétrico da energia solar, principalmente, da operação dos sistemas fotovoltaicos.



Fonte: Pereira et al. (2006).

Haja vista que a energia elétrica proveniente dos sistemas solares em geral é uma alternativa para complementar a matriz energética brasileira, ainda

predominantemente hidráulica de acordo com Ministério de Minas e Energia (2017), apresenta-se na Figura 2.8 o mapa com a média anual do total diário de irradiação solar global incidente no território brasileiro.

De acordo com Pereira et al. (2006), a variação de irradiação solar global incidente em qualquer região do território brasileiro é na faixa de 1.500 a 2.500 kWh/m², superior ao de países onde o aproveitamento de recursos solares são amplamente disseminados, tais como Alemanha, França e Espanha. Logo, todo o território brasileiro é propício para o aproveitamento da energia solar.

2.5 TRABALHOS RELACIONADOS

Na literatura, vários métodos de mineração de dados têm proporcionado contribuições que auxiliam no processo de estimativa da radiação solar global. Alguns trabalhos com esse propósito foram identificados e relatados nesta seção.

Em Demirtas et al. (2012), um modelo para previsão de radiação solar em um intervalo de 10 minutos é desenvolvido aplicando o algoritmo *k* vizinhos mais próximos, kNN (*k nearest neighbor*). Temperatura, umidade e pressão atmosférica constituem as variáveis de entrada do modelo, cujos dados correspondem somente ao mês de maio de 2012. Foram realizados alguns experimentos com métricas de distância Euclidiana, Manhattan e Minkowski, e diferentes valores para *k* para definir o modelo com melhor desempenho.

Em Mori e Takahashi (2012), foi proposta uma abordagem de mineração de dados para selecionar as variáveis de entrada de modelos de previsão de radiação solar global. Através do algoritmo CART, dentre algumas variáveis meteorológicas e radiométricas de entrada, foram definidas quais variáveis demonstraram maior relevância para uma previsão mais precisa da radiação com relação às estações do ano, verão e inverno.

Wasu, Kariya e Tote (2013) propuseram um modelo de *clustering* que indica os meses e a cidade com temperatura média mensal máxima a partir do mês, temperatura média máxima e temperatura média mínima. As etapas de pré-processamento e mineração de dados foram realizadas por meio da ferramenta WEKA, onde as técnicas *simple k-means* e EM (do inglês, *Expectation Maximization*) foram aplicadas para minerar os dados. Os autores ressaltam que através da temperatura máxima é possível determinar a quantidade de radiação solar

disponível e, conseqüentemente, avaliar a viabilidade da instalação de um sistema de energia solar e prever sua produção de energia.

Outro modelo desenvolvido por Yadav, Malik e Chandel (2015) é baseado na aplicação de redes neurais artificiais para previsão diária da radiação solar global. Eles consideram como variáveis de entrada somente a temperatura média diária – valores são coletados em intervalos de 10 minutos – e temperaturas máxima e mínima diária. As redes neurais correspondem a combinações das variáveis de entrada e foram implementadas no software MATLAB. Os autores argumentam que a temperatura por si só é suficiente para realizar as previsões.

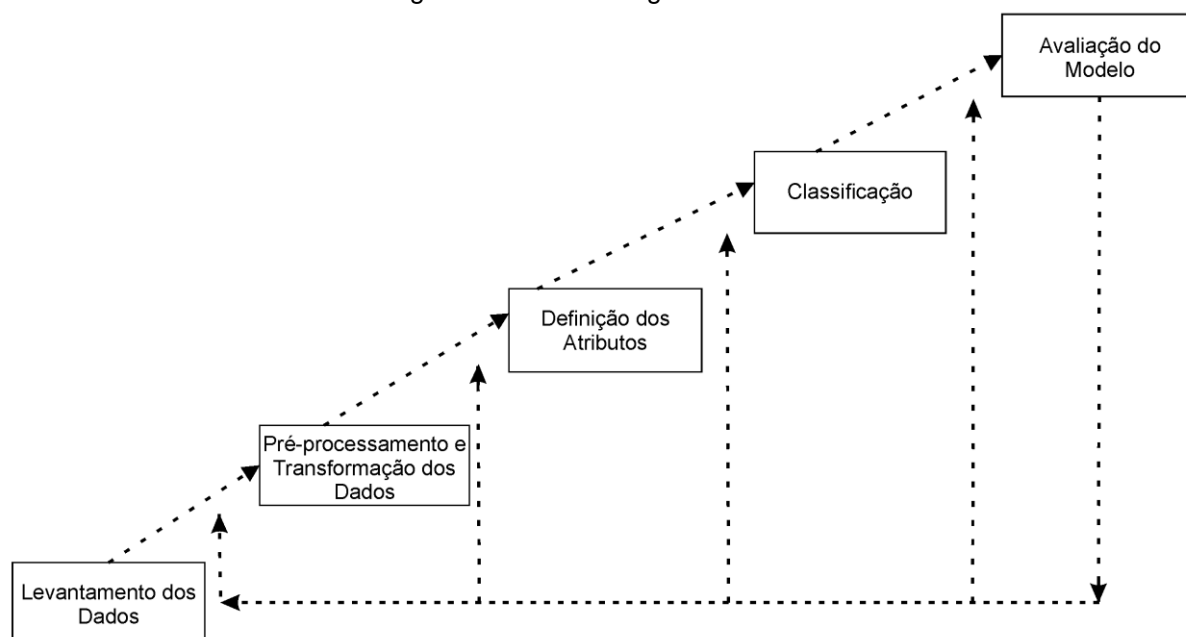
Melzi et al. (2016) apresentaram vários modelos para previsão horária da radiação solar global. Os dados adquiridos referem-se à radiação solar global, expressa em Wm^{-2} , para um período de 11 anos (2004 a 2015) e foram representados em três partes distintas para cada dia: (1) noite antes de o dia nascer, (2) parte do dia entre nascer e pôr-do-sol – na literatura, o termo *sunshine hours* é bastante empregado para descrever essa definição – e (3) noite após o pôr-do-sol. Esses conceitos foram criados para auxiliar na escolha do intervalo anual mais adequado, baseado na variação de *sunshine hours*, para aplicação dos modelos. Os autores usaram os métodos Naive Bayes, *Autoregressive Moving Average* (ARMA), um método de similaridade, rede neural e *support vector machine* com avaliação por meio do *cross validation* e métricas, tais como erro quadrático médio e erro quadrático médio normalizado.

Considerando os trabalhos supracitados, observa-se que, apesar da aplicação de técnicas e métodos distintos de mineração de dados, o conjunto dos dados de entrada varia conforme o modelo empregado, o período do ano e estação de coleta dos dados. Contudo, a definição apropriada desse conjunto, seja por métodos de correlação ou combinações pré-definidas, exerce um papel relevante para o desempenho dos modelos. Além disso, também sugerem a validação do modelo por meio de algumas métricas estatísticas.

3 METODOLOGIA DE MINERAÇÃO DOS DADOS

A metodologia utilizada para o desenvolvimento deste trabalho divide-se em etapas iterativas e interativas e está ilustrada na Figura 3.1.

Figura 3.1 – Metodologia do trabalho



Fonte: Autoria própria.

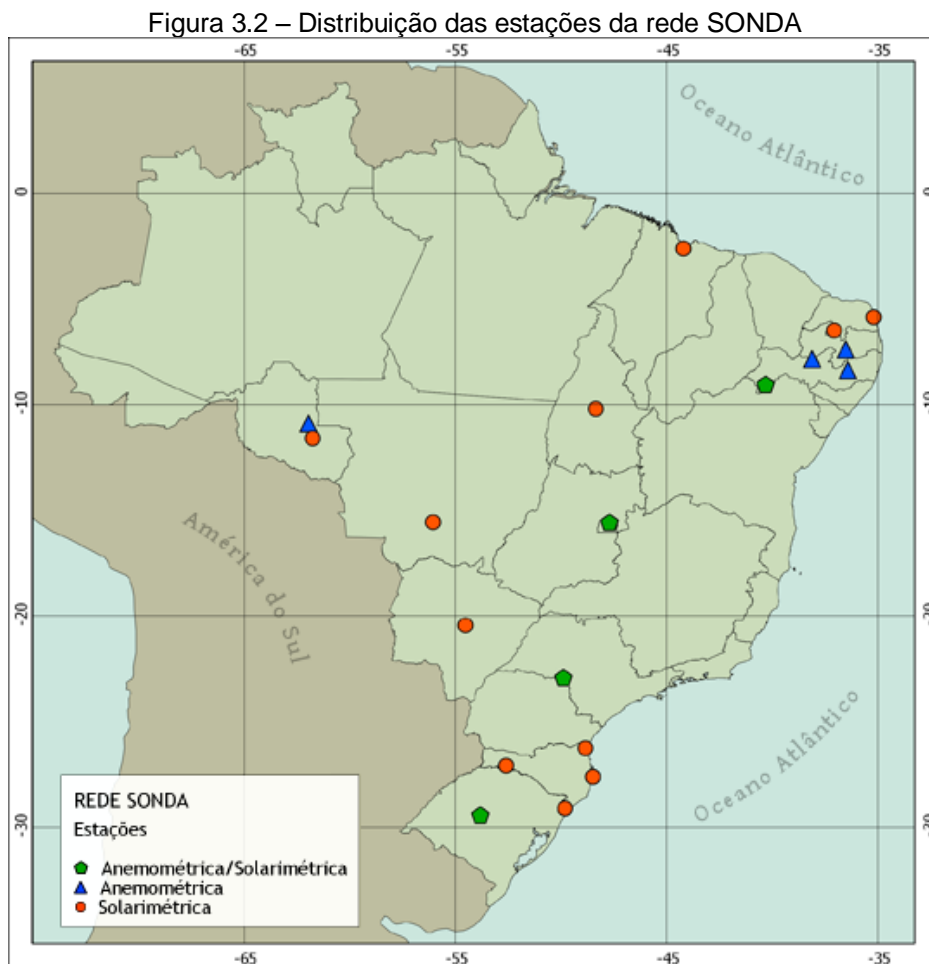
3.1 LEVANTAMENTO DOS DADOS

Os dados utilizados durante o desenvolvimento deste trabalho foram adquiridos através da base de dados do Sistema de Organização Nacional de Dados Ambientais (SONDA), que está disponível em (SONDA, 2017a). A rede SONDA, gerenciada pelo Instituto Nacional de Pesquisas Espaciais (INPE), tem o intuito de melhorar a base de dados dos recursos de energia solar e eólica do Brasil. Os dados da rede SONDA provêm das suas estações, as quais medem variáveis distintas conforme a configuração dos seus sensores.

Vale ressaltar que a Fundação Cearense de Meteorologia e Recursos Hídricos (FUNCEME) foi contatada em busca do apoio de especialistas para maior entendimento do comportamento dinâmico da radiação solar global. A FUNCEME está vinculada à Secretária dos Recursos Hídricos do Estado do Ceará e tem o propósito de estudo da meteorologia, dos recursos hídricos e recursos ambientais

visando propiciar informações necessárias para o desenvolvimento sustentável do Ceará e da região Nordeste (FUNCEME, 2017).

A Figura 3.2 apresenta como as estações da rede SONDA estão distribuídas no território brasileiro.



Fonte: SONDA (2017a).

A Figura 3.3 ilustra como são classificadas as variáveis medidas pelas estações da rede SONDA. Os dados são classificados em Ambientais e Anemométricos. Nos dados Ambientais, apresentam-se dados radiométricos e meteorológicos com periodicidade de coleta de 1 minuto. Os dados Anemométricos têm periodicidade de coleta de 10 minutos e apresenta sensores em alturas de 25 e 50 metros.

Figura 3.3 – Classificação das variáveis medidas pelas estações da rede SONDA

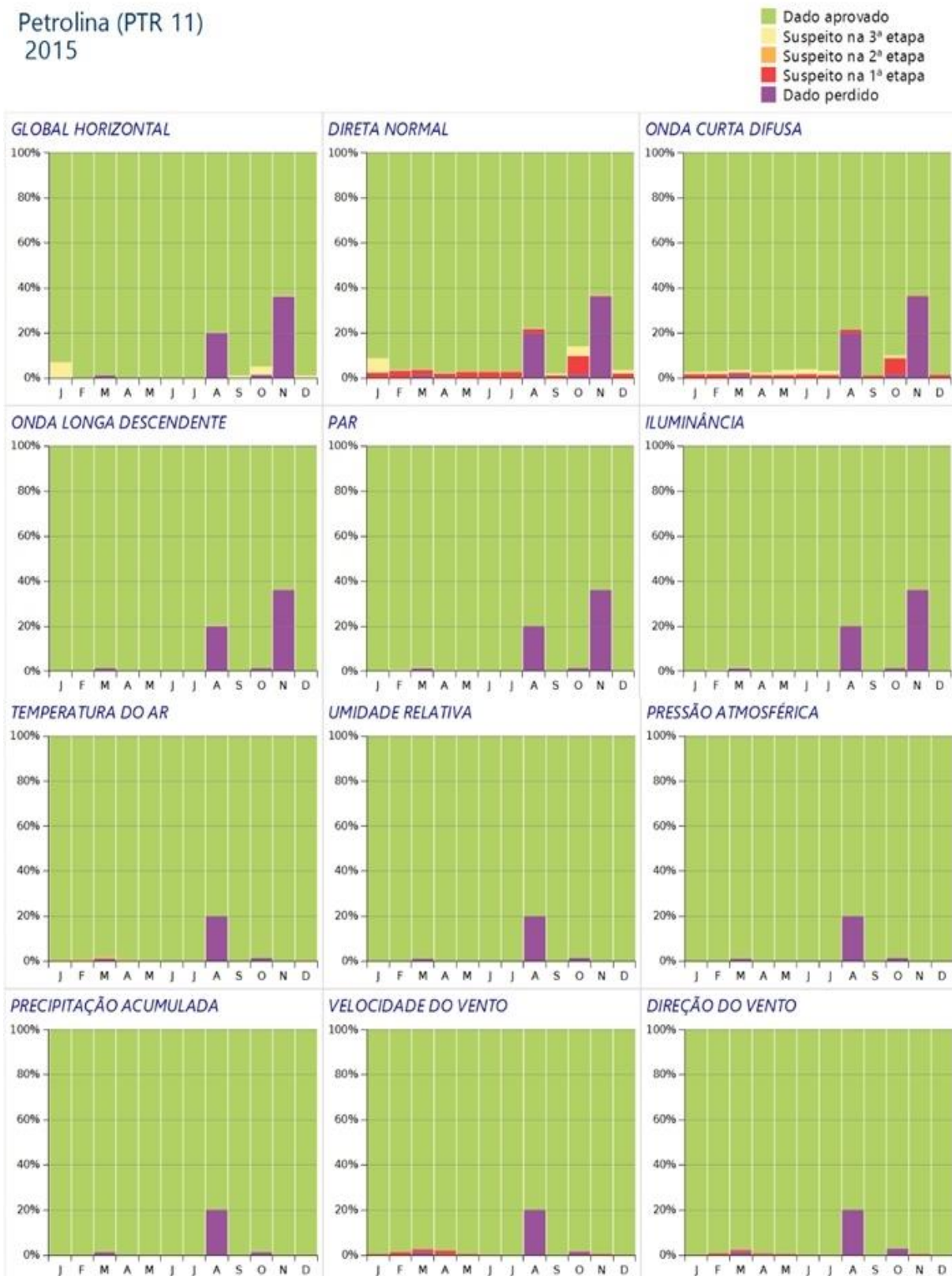


Fonte: SONDA (2017b).

Dentre as estações disponíveis, a base de dados selecionada teve seus dados coletados na estação Anemométrica e Solarimétrica de Petrolina, Pernambuco. O período dos dados é de janeiro de 2015 a dezembro de 2015, com periodicidade de captação de um minuto e estão disponíveis em formato de planilha de acordo com o mês de coleta.

Os dados disponibilizados pelas estações da rede SONDA passam por um processo de validação, cujo objetivo é a identificação de dados inconsistentes. O resultado desse processo também é disponibilizado, conforme o exemplo da Figura 3.4. O gráfico representa o resultado do processo de validação dos dados para cada variável de coleta ao longo do ano. Os dados para cada variável são classificados mensalmente (representado pelo eixo X) como: “Dado Aprovado” (verde), “Suspeito na 3ª etapa” (amarelo), “Suspeito na 2ª etapa” (laranja), “Suspeito na 1ª etapa” (vermelho) e “Dado Perdido” (roxo). A porcentagem de como os dados foram classificados está representada pelo eixo Y do gráfico para cada variável. Assim, através da análise e visualização gráfica dos resultados da validação dos dados, constatou-se que os dados de 2015 da estação de Petrolina apresentam-se mais confiáveis do que os de outras estações.

Figura 3.4 – Resultado da validação dos dados radiométricos e meteorológicos da estação de Petrolina para o ano de 2015



Fonte: SONDA (2017c).

3.2 PRÉ-PROCESSAMENTO E TRANSFORMAÇÃO DOS DADOS

O pré-processamento e a transformação dos dados é a etapa que consome a maior parte do tempo no processo de KDD e é de fundamental importância obter bons resultados na classificação dos dados.

Os dados da rede SONDA são validados, conforme citado na seção 3.1, através de algoritmos baseados nas estratégias de controle de qualidade adotada pela *Baseline Surface Radiation Network* (BSRN), para dados de radiação solar, e Webmet.com, para dados meteorológicos e anemométricos.

As duas primeiras etapas da validação utilizam os mesmos algoritmos. Na primeira etapa um algoritmo considera os dados suspeitos quando estes apresentam-se fisicamente impossíveis como, por exemplo, valores para umidade relativa do ar superiores a 100% ou inferiores a 0%. Já na segunda, outro algoritmo detecta quando um evento é extremamente raro, como a variação da pressão atmosférica menor do que 6 mb (milibar) em um período de três horas consecutivas.

Na terceira etapa, um algoritmo sinaliza como suspeita uma variável anemométrica ou meteorológica que apresente uma evolução temporal não condizente com o esperado, tal como a variação da precipitação menor do que 100 mm (milímetros) para um período de vinte e quatro horas consecutivas. Para as variáveis radiométricas, outro algoritmo identifica dados suspeitos quando estes são inconsistentes em relação a outras variáveis medidas, como a relação entre radiação solar global e a radiação difusa. A última etapa é realizada apenas para dados anemométricos e meteorológicos, e também busca identificar inconsistências com relação a outras variáveis medidas.

O processo de validação realizado pela rede SONDA está ilustrado no Anexo A e Anexo B. Como resultado dessa validação, é elaborada uma planilha com a sinalização dos dados suspeitos. Através de um *script* na linguagem Python (Apêndice A) e dessa planilha, somente os dados considerados válidos (sem suspeitas) foram selecionados, totalizando 201836 instâncias.

Em seguida, foi realizada a correção de inconsistências nos dados em relação às casas decimais, provavelmente ocasionadas durante a criação das planilhas pelas estações.

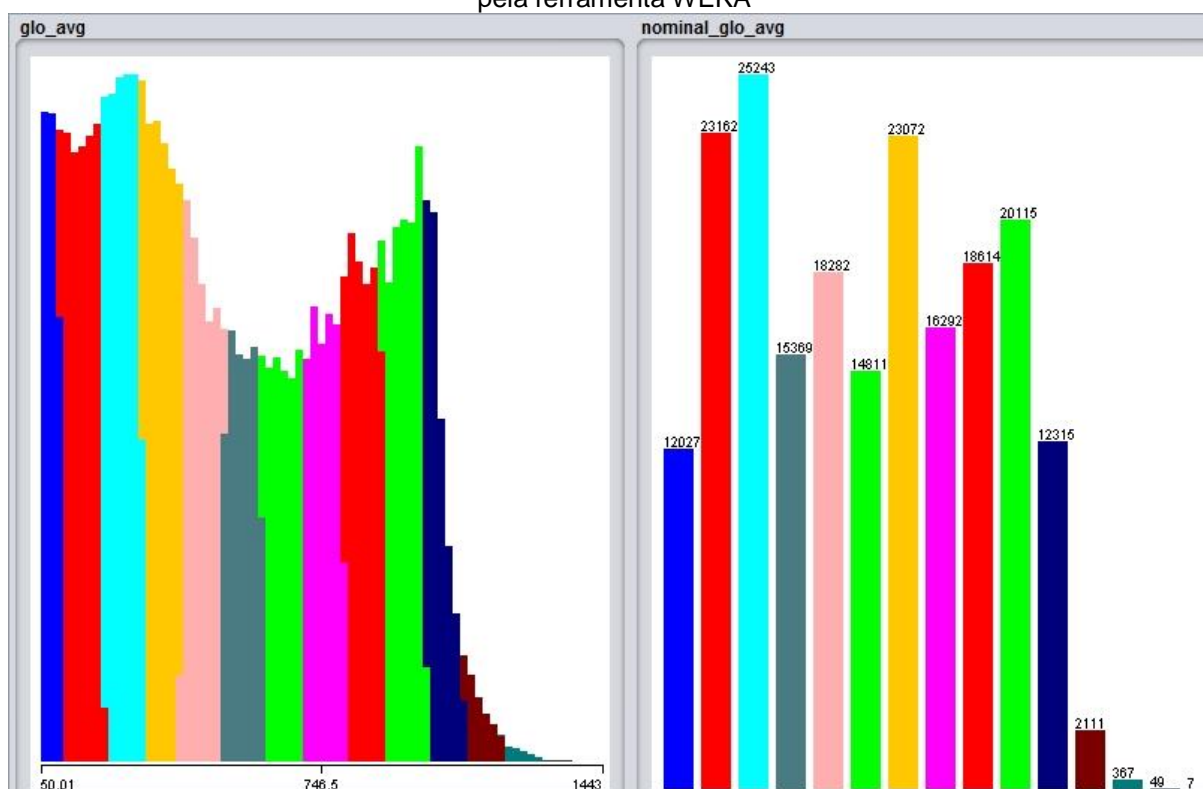
A abordagem desenvolvida neste trabalho apresenta um modelo de prazo muito curto com objetivo de estimar tanto os dados válidos de 1 minuto quanto de 30

minutos. Modelos com prazo muito curto, de segundos a minutos, são utilizados para o controle e segurança do sistema de gerenciamento de energia da rede elétrica e qualidade/confiabilidade da energia de uso final (Voyant, 2017). Assim, uma base de dados em uma série temporal de 30 minutos foi gerada a partir da base de dados de 1 minuto. Essa transformação foi realizada calculando a média a cada conjunto de 30 valores dos dados válidos, e resultou em 6722 instâncias.

Nesta etapa, além da extrapolação temporal, também realizou-se a discretização dos dados da radiação solar global, pois o algoritmo de classificação manipula apenas valores de uma classe especificada. Considerando que o valor mínimo dos dados válidos para a radiação solar global é 50.01 Wm^{-2} e o valor máximo é 1443 Wm^{-2} , os dados foram discretizados em intervalos de 100 Wm^{-2} , portanto os valores para a radiação solar passam a ter uma representação nominal correspondente ao respectivo intervalo, ou classes.

A Figura 3.5 mostra a representação nominal da radiação solar global (*nominal_glo_avg*) em relação à sua representação numérica (*glo_avg*) para os dados de 1 minuto.

Figura 3.5 – Representação da radiação solar global em classes para os dados de 1 minuto gerada pela ferramenta WEKA



Fonte: Autoria própria.

3.3 DEFINIÇÃO DOS ATRIBUTOS

A definição dos atributos consiste na seleção do subconjunto de variáveis consideradas relevantes para o domínio do problema. Sendo também uma etapa interativa e iterativa, por conseguinte, vários subconjuntos são treinados e testados com intuito de encontrar aquele(s) que proporciona(m) resultados mais precisos na estimativa da radiação solar global.

Dentre os conjuntos de variáveis anemométricas, meteorológicas e radiométricas, e os dados de coleta da estação de Petrolina, as variáveis apresentadas na Tabela 3.1 foram selecionadas como candidatas a preditores de radiação solar.

Tabela 3.1 – Variáveis candidatas a preditores

Variável	Descrição (unidades)	Tipo
<i>month</i>	Mês de coleta	Dados de coleta
<i>lw_avg</i>	Média da radiação de onda longa descendente (Wm^{-2} , <i>Watts</i> por metro quadrado)	Radiométrica
<i>lux_avg</i>	Média da iluminância (kLux)	Radiométrica
<i>tp_sfc</i>	Temperatura do ar na superfície (°C)	Meteorológica
<i>humid</i>	Umidade relativa do ar (%)	Meteorológica
<i>press</i>	Pressão atmosférica em (mb, milibares)	Meteorológica

Fonte: Autoria própria.

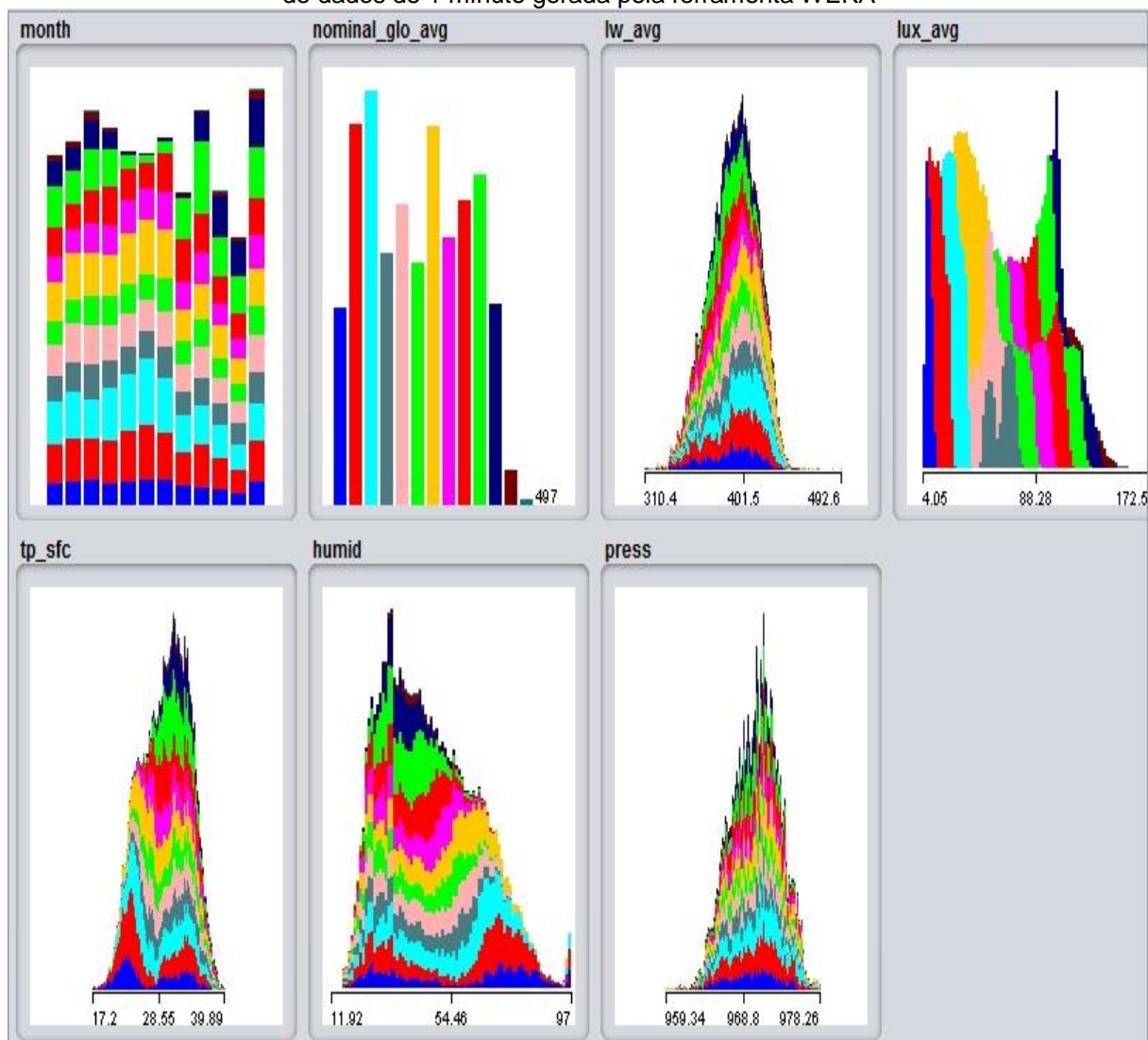
3.4 CLASSIFICAÇÃO

Na etapa de classificação, as variáveis selecionadas (Tabela 3.1) têm seus respectivos valores submetidos ao algoritmo de classificação J48 da ferramenta WEKA.

Os experimentos foram realizados com a base de dados a série temporal de 1 minuto e a base de dados com série temporal extrapolada em 30 minutos. A Figura

3.6 demonstra a relação entre a classe objetivo, radiação solar global discretizada em intervalos iguais (*nominal_glo_avg*), e todas as variáveis candidatas a preditores para a base de dados de 1 minuto.

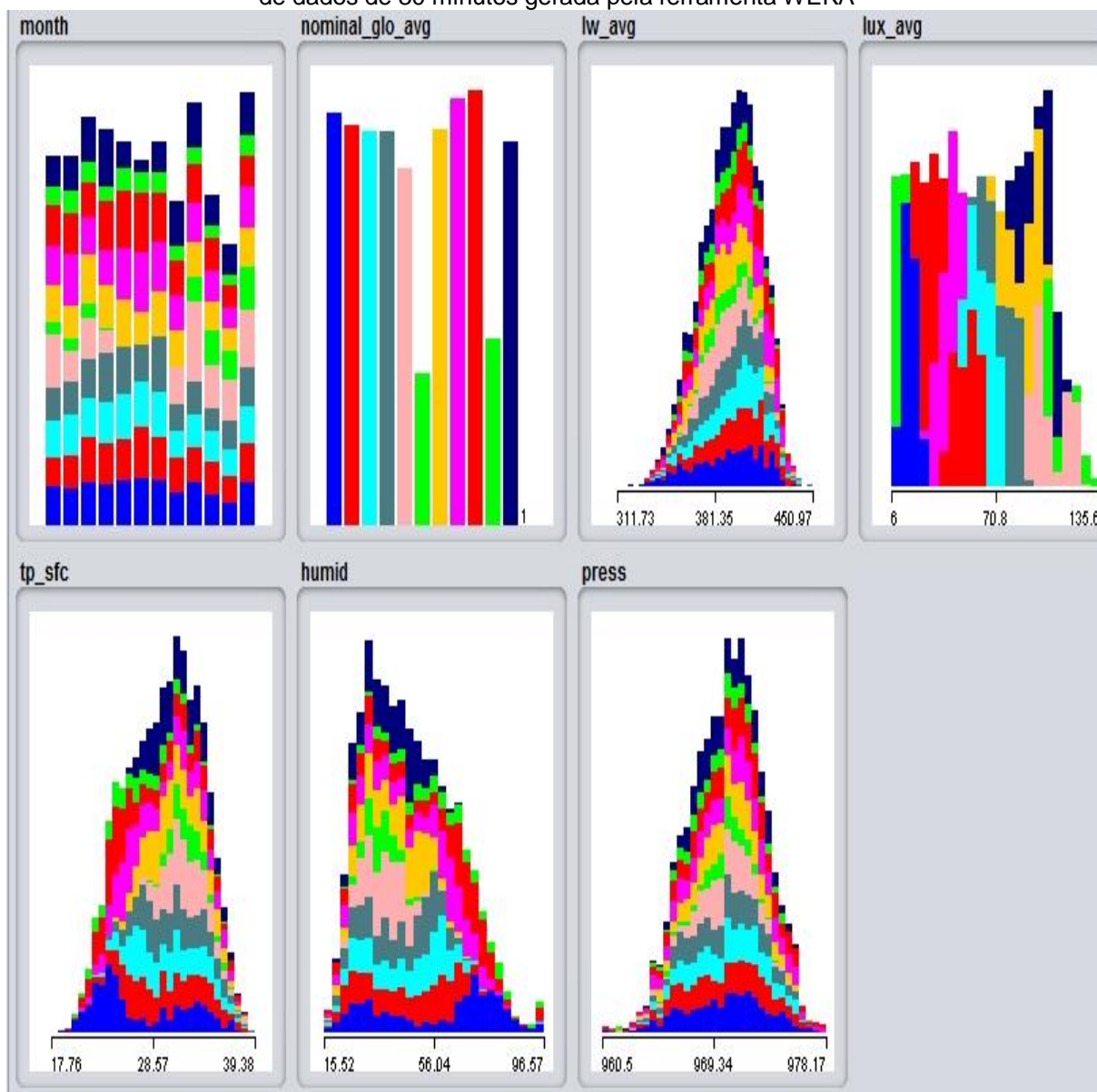
Figura 3.6 – Relação entre a radiação solar global e as variáveis candidatas a preditores para a base de dados de 1 minuto gerada pela ferramenta WEKA



Fonte: Autoria própria.

A Figura 3.7 demonstra a relação entre a radiação solar global e todas as variáveis candidatas a preditores para a base de dados de 30 minutos.

Figura 3.7 – Relação entre a radiação solar global e as variáveis candidatas a preditores para a base de dados de 30 minutos gerada pela ferramenta WEKA



Fonte: Autoria própria.

Nas Figuras 3.6 e 3.7, a radiação solar global está discretizada em classes, onde cada cor representa uma classe e a quantidade de valores para cada uma. Nos atributos (variáveis) preditores, a radiação solar global está distribuída em intervalos, como, por exemplo, no atributo mês (*month*). Nos atributos radiação de onda longa descendente (*lw_avg*), iluminância (*lux_avg*), temperatura (*tp_sfc*), umidade (*humid*) e pressão (*press*) apresentam-se seus valores mínimo, médio e máximo e pequenos intervalos com suas respectivas classes.

Durante os experimentos, todas as variáveis foram avaliadas individualmente na tentativa de entender a relação entre elas e a radiação solar global. Segundo

Guarnieri (2006), variáveis correlacionáveis com a radiação solar global, como fluxos de energia e temperatura, podem ser utilizadas para a estimativa da radiação solar.

3.5 AVALIAÇÃO DO MODELO

A avaliação do desempenho do modelo de classificação serve como parâmetro para as etapas iterativas do processo de KDD. Dessa forma, o método *cross-validation* (validação cruzada) com k partições, ou *folds*, foi empregado para avaliar o modelo para estimativa da radiação solar global.

Nesse método, os dados são divididos randomicamente em k subconjuntos (*folds*) mutuamente exclusivos de tamanhos aproximadamente iguais. Em seguida, k iterações de treinamento e teste são realizadas, de modo que enquanto uma partição é testada as demais são treinadas. Assim, todas as partições são treinadas e testadas. O número k de partições é definido manualmente e a estimativa do erro é a média dos erros obtidos em k iterações. Para Witten et al. (2016) e Han, Kamber, e Pei (2011), 10 partições ($k=10$) é uma boa maneira para obter a melhor estimativa do erro. Este trabalho adotou, conseqüentemente, 10 partições.

Como os valores da radiação solar global foram discretizados em atributos categóricos, ou nominais, a utilização de métricas estatísticas bastante empregadas para esse domínio não foram utilizadas. A acurácia do modelo, proximidade entre o valor medido e o valor obtido experimentalmente, é definida pela exatidão do algoritmo J48 em relação ao número de instâncias classificadas corretamente e o número total de instâncias, e é expressa pela Equação 3.1:

$$Acurácia = 100 \times \frac{n^{\circ} \text{ de instâncias corretamente classificadas}}{n^{\circ} \text{ total de instâncias}} \quad (3.1)$$

4 RESULTADOS E DISCUSSÕES

Nesta seção são apresentados e discutidos os resultados obtidos da classificação utilizando o algoritmo J48 da ferramenta WEKA para estimativa da radiação solar global para a base de dados de série temporal de 1 minuto e para a base de dados de série temporal de 30 minutos.

Inicialmente, as variáveis candidatas a preditores foram avaliadas individualmente com intuito de entender a influência que exercem sobre a radiação solar global e analisar a correlação entre elas. A Tabela 4.1 apresenta os resultados dos modelos para a estimativa da radiação solar global para 1 minuto utilizando somente uma variável de entrada.

Tabela 4.1 – Resultados dos modelos para estimativa da radiação solar global para 1 minuto com uma variável de entrada

Variável De Entrada	Acurácia (%)
<i>Month</i>	14
<i>lw_avg</i>	15
lux_avg	65
<i>tp_sfc</i>	19
<i>Humid</i>	20
<i>Press</i>	19

Fonte: Autoria própria.

A Tabela 4.2 demonstra os resultados dos modelos para a estimativa da radiação solar global para 30 minutos utilizando somente uma variável de entrada.

Tabela 4.2 – Resultados dos modelos para estimativa da radiação solar global para 30 minutos com uma variável de entrada

Variável De Entrada	Acurácia (%)
<i>Month</i>	14
<i>lw_avg</i>	12
lux_avg	62
<i>tp_sfc</i>	16
<i>Humid</i>	19
<i>Press</i>	11

Fonte: Autoria própria.

De acordo com os resultados dos modelos vistos nas Tabelas 4.1 e 4.2, a variável que exerce maior influência para a estimativa da radiação solar global é a iluminância (*lux_avg*) com 65% e 62% de acurácia para 1 minuto e 30 minutos, respectivamente. A outra variável radiométrica empregada, radiação de onda longa descendente (*lw_avg*), não demonstra tanta influência individualmente. Dentre as variáveis meteorológicas, a umidade relativa do ar (*humid*) demonstra maior relação com a radiação solar global, 20% de acurácia para o modelo de 1 minuto e 19% para o de 30 minutos.

Após a avaliação individual das variáveis, vários experimentos foram realizados de modo que as variáveis foram combinadas entre si conforme seu tipo (radiométrica, meteorológica e dados de coleta).

A Tabela 4.3 demonstra os resultados dos modelos para a estimativa da radiação solar global para 1 minuto utilizando algumas variáveis de entrada.

Tabela 4.3 – Resultados dos modelos para estimativa da radiação solar global para 1 minuto com algumas variáveis de entrada

Variáveis De Entrada	Acurácia (%)
<i>month, lw_avg, lux_avg, tp_sfc, humid, press</i>	94
<i>month, lw_avg, lux_avg</i>	93
<i>lw_avg, lux_avg</i>	67
<i>month, tp_sfc, humid, press</i>	54
<i>tp_sfc, humid, press</i>	42
<i>month, lux_avg, tp_sfc, humid, press</i>	93
<i>lux_avg, tp_sfc, humid, press</i>	81
<i>month, lw_avg, tp_sfc, humid, press</i>	60
<i>lw_avg, tp_sfc, humid, press</i>	52

Fonte: Autoria própria.

Para o modelo de 1 minuto, o melhor resultado obtido apresenta acurácia de 94% e utiliza todas as variáveis candidatas. No entanto, com menos variáveis de entrada, outros dois modelos, com destaque para o modelo com combinação das variáveis radiométricas e dados de coleta (*lw_avg*, *lux_avg*, *month*), resultaram em 93% de acurácia. O melhor modelo sem empregar nenhuma variável radiométrica apresenta 54% de precisão.

A Tabela 4.4 demonstra os resultados dos modelos para a estimativa da radiação solar global para 30 minutos utilizando algumas variáveis de entrada.

Tabela 4.4 – Resultados dos modelos para estimativa da radiação solar global para 30 minutos com algumas variáveis de entrada

Variáveis De Entrada	Acurácia (%)
<i>month, lw_avg, lux_avg, tp_sfc, humid, press</i>	91
<i>month, lw_avg, lux_avg</i>	91
<i>lw_avg, lux_avg</i>	62
<i>month, tp_sfc, humid, press</i>	23
<i>tp_sfc, humid, press</i>	20
<i>month, lux_avg, tp_sfc, humid, press</i>	90
<i>lux_avg, tp_sfc, humid, press</i>	69
<i>month, lw_avg, tp_sfc, humid, press</i>	27
<i>lw_avg, tp_sfc, humid, press</i>	24

Fonte: Autoria própria.

Na estimativa da radiação solar global para 30 minutos, os melhores modelos obtiveram 91% de precisão: o primeiro modelo utiliza todas as variáveis de entrada, e o outro aplica-se apenas mês (*month*), radiação de onda longa descendente (*lw_avg*) e iluminância (*lux_avg*). Para essa série temporal, o modelo sem nenhuma

variável radiométrica que demonstrou melhor precisão teve apenas 23% de acurácia.

A partir das Tabelas 4.3 e 4.4, constatou-se que os melhores resultados obtidos possuem a iluminância e o mês de coleta como variáveis de entrada determinantes. Por conseguinte, dois experimentos foram realizados para cada série temporal utilizando apenas mês (*month*) e iluminância (*lux_avg*) como variáveis de entrada, conforme visto na Tabela 4.5.

Tabela 4.5 – Resultados dos modelos para estimativa da radiação solar global com mês e iluminância como variáveis de entrada

Intervalo de Estimativa	Variáveis de entrada	Acurácia (%)
1 minuto	month, lux_avg	91
30 minutos	month, lux_avg	91

Fonte: Autoria própria.

Para ambas as séries temporais, o resultado obtido para a estimativa da radiação solar global teve acurácia de 91%. Dessa forma, com uma quantidade menor de variáveis de entrada (somente mês e iluminância) foi possível obter modelos com precisão significativa, semelhante aos melhores modelos apresentados nas Tabelas 4.3 e 4.4, e redução do custo computacional durante a etapa de Classificação.

5 CONSIDERAÇÕES FINAIS

A restrição de informações de cunho científico a respeito da disponibilidade da energia solar é um dos enormes obstáculos para a exploração e desenvolvimento comercial das tecnologias de sistemas solares. Todavia este trabalho, ao atingir seus objetivos propostos, propicia informações relevantes para o planejamento de tais sistemas, como os fotovoltaicos e os CSPs.

O esforço demandado nas etapas de Pré-processamento e Transformação dos Dados e Definição dos Atributos aplicados neste trabalho evidencia que a qualidade dos dados disponíveis é um fator que dificulta o desenvolvimento de pesquisas científicas na área.

Os modelos de estimativa com prazo muito curto de 1 minuto e 30 minutos para os dados válidos do ano de 2015, da Estação Solarimétrica e Anemométrica de Petrolina da rede SONDA, mostraram-se satisfatoriamente avaliados através do método *cross-validation* (validação cruzada) com 10 partições. Ambos os modelos utilizaram, inicialmente, 6 variáveis de entrada e têm estimativas para valores discretizados em 100 Wm^{-2} . Portanto, fazem-se necessários estudos e investigações mais detalhadas acerca do potencial elétrico da energia solar com o propósito de alcançar resultados mais específicos.

A análise das variáveis de entrada mostrou-se uma abordagem adequada para o domínio do problema, tendo em vista, que a seleção de variáveis com maior relação com a radiação solar global apresentou resultados significativos durante a classificação (acima de 90%) e a redução de variáveis de entrada para o modelo demonstrou redução do custo computacional.

Como propostas de trabalhos futuros, pretende-se estimar saída dos sistemas solares a partir da previsão da radiação solar global e a utilização de outras técnicas de classificação vistas na literatura, como redes neurais artificiais e SVMs, que são bastante aplicadas nesse domínio, as quais também estão disponíveis na ferramenta WEKA, visando um modelo que proporcione o melhor desempenho, tendo em vista os resultados e o custo computacional.

REFERÊNCIAS

- AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA. **Atlas de Energia Elétrica do Brasil**. 3. ed. Brasília, 2008. 236 p. Disponível em: <<http://www2.aneel.gov.br/arquivos/PDF/atlas3ed.pdf>>. Acesso em: 8 fev. 2017.
- ANGELIS-DIMAKIS, A. et al. Methods and tools to evaluate the availability of renewable energy sources. **Renewable and Sustainable Energy Reviews**, Elsevier, v. 15, n. 2, p.1182-1200, 2011.
- BARROS, R. C. et al. Automatic Design of Decision-Tree Algorithms with Evolutionary Algorithms. **Evolutionary Computation**, MIT Press Journals, v. 21, n. 4, p. 659-684, Nov. 2013.
- BASGALUPP, M. P. **LEGAL-Tree**: Um algoritmo genético multi-objetivo lexicográfico para indução de árvores de decisão. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – Universidade de São Paulo, São Carlos, 2010.
- DEMIRTAS, M. et al. Prediction of Solar Radiation Using Meteorological Data. In: **2012 International Conference on Renewable Energy Research and Applications (ICRERA)**, Nov. 2012, Nagasaki, Japan. **IEEE Conference Publications** ... Japan: IEEE, Mar. 2013, p.1-4. Disponível em: <<http://ieeexplore.ieee.org/abstract/document/6477329/>> Acesso em: 8 mar. 2017.
- FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases: an overview. **American Association for Artificial Intelligence**, Menlo Park, USA, v. 17, n. 3, p. 37-54, 1996. Disponível em: <<http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf>>. Acesso em: 2 fev. 2017.
- FUNCEME. Fundação Cearense de Meteorologia e Recursos Hídricos. Disponível em: <<http://www.funceme.br/>> Acesso em: 5 jan. 2017.
- GUARNIERI, A. R. **Emprego de redes neurais artificiais e regressão linear múltipla no refinamento das previsões de radiação solar do modelo ETA**. Dissertação (Mestrado em Meteorologia) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2006.
- HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. 3rd ed. Waltham, USA: Elsevier, 2011. 703 p. Disponível em: <<https://books.google.com.br/books?hl=pt-BR&lr=&id=pQws07tdpjoC&oi=fnd&pg=PP1&dq=data+mining+concepts+and+techniques&ots=tyLvZWqy1W&sig=np--Gig64bo5ljkBh8wcfM6DU#v=onepage&q=data%20mining%20concepts%20and%20techniques&f=false>>. Acesso em: 2 fev. 2017.
- KANNAN, N.; VAKEESAN, D. Solar energy for future world: - A review. **Renewable and Sustainable Energy Reviews**, Elsevier, v. 62, p. 1092-1105, 2016.
- LAROSE, D. T. **Discovering knowledge in data: an introduction to data mining**. 2nd ed. New Jersey, USA: IEEE, 2014. 316 p. Disponível em:

<<https://books.google.com.br/books?hl=pt-BR&lr=&id=UGu8AwAAQBAJ&oi=fnd&pg=PT22&dq=DISCOVERING+KNOWLEDGE+IN+DATA+An+Introduction+to+Data+Mining&ots=zruRkcPJAN&sig=mVHUx2ymJvGHuV3rSLkiw2UNvHc#v=onepage&q=DISCOVERING%20KNOWLEDGE%20IN%20DATA%20An%20Introductio>>. Acesso em: 2 fev. 2017.

LIMA, F. J. L. **Previsão de irradiação solar no Nordeste do Brasil empregando o modelo WRF ajustado por redes neurais artificiais (RNAs)**. Tese (Doutorado em Meteorologia) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2015.

MAIMON, O.; ROKACH, L. (Ed.). **Data Mining and Knowledge Discovery Handbook**. 2nd ed. New York, USA: Springer, 2010. 1306 p. Disponível em: <<https://books.google.com.br/books?id=aHIsT6LBI0C&printsec=frontcover#v=onepage&q&f=false>> Acesso em: 8 mar. 2017.

MELZI, F. N. et al. Hourly solar irradiance forecasting based on machine learning models. In: **15th IEEE International Conference on Machine Learning and Applications (IEEE ICMLA'16)**, Nov. 2016, Anaheim, USA. **IEEE Conference Publications...** USA: IEEE, 2 Feb. 2017, p. 441-446. Disponível em: <<http://ieeexplore.ieee.org/document/7838182/>> Acesso em: 8 mar. 2017.

MINISTÉRIO DE MINAS E ENERGIA. Empresa de Pesquisa Energética. **Balanco Energético Nacional 2016**: Ano base 2015. Rio de Janeiro: EPE, 2016, 292 p. Disponível em: <<https://ben.epe.gov.br/>> Acesso em: 8 mar. 2017.

MORI, H.; TAKAHASHI, A. A. Data mining method for selecting input variables for forecasting model of global solar radiation. In: **2012 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)**, May. 2012, Orlando, USA. **IEEE Conference Publications...** USA: IEEE, 27 Aug. 2012, p.1-6. Disponível em: <<http://ieeexplore.ieee.org/document/6281569/>> Acesso em: 8 mar. 2017.

PACHECO, F. Energias Renováveis: breves conceitos. **Conjuntura e Planejamento**, Salvador, v. 149, p. 4-11, 2006.

PEREIRA, E. B. et al. **Atlas Brasileiro de Energia Solar**. 1. ed. São José do Campos: INPE, 2006. 60 p. Disponível em: <<http://sonda.ccst.inpe.br/publicacoes/index.html>> Acesso em: 22 nov. 2016.

QUINLAN, J. R. Improved use of continuous attributes in C4. 5. **Journal of Artificial Intelligence Research**, v.4, p. 77-90, 1996. Disponível em: <<http://www.jair.org/papers/paper279.html>> Acesso em: 8 mar. 2017.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, Boston, Springer, v.1, n.1, p. 81-106, mar 1986. Disponível em: <<https://link.springer.com/article/10.1023/A:1022643204877>> Acesso em: 8 mar. 2017.

SILVA, M. P. S. Mineração de dados: Conceitos, aplicações e experimentos com weka. **Sociedade Brasileira de Computação**. 2004.

SONDAa. Sistema de Organização Nacional de Dados Ambientais. **REDE SONDA**. Disponível em: <<http://sonda.ccst.inpe.br/index.html>>. Acesso em: 23 fev. 2017.

SONDAb. Sistema de Organização Nacional de Dados Ambientais. **BASE DE DADOS**. Disponível em: <<http://sonda.ccst.inpe.br/basedados/index.html>>. Acesso em: 23 fev. 2017.

SONDAc. Sistema de Organização Nacional de Dados Ambientais. **ESTAÇÃO DE PETROLINA**. Disponível em: <http://sonda.ccst.inpe.br/basedados/graficos/ambientais/PTR/2015/val_PTR15ED.html>. Acesso em: 23 fev. 2017.

SONDAd. Sistema de Organização Nacional de Dados Ambientais. **VALIDAÇÃO DOS DADOS**. Disponível em: <<http://sonda.ccst.inpe.br/infos/validacao.html>>. Acesso em: 23 fev. 2017.

TOLMASQUIM, M. T. (Coord.). **Energia Renovável: Hidráulica, Biomassa, Eólica, Oceânica**. Rio de Janeiro: EPE, 2016. Disponível em: <<http://www.epe.gov.br/Paginas/default.aspx>> Acesso em: 2 fev. 2017.

VOYANT, C. et al. Machine learning methods for solar radiation forecasting: A review. **Renewable Energy**, Elsevier, v. 105, p. 569-582, 2017. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0960148116311648>> Acesso em: 28 abr. 2017.

WASU, A. A.; KARIYA, H. M.; TOTE, S. S. Evaluating renewable energy using data mining techniques in developing India. **International Journal of Scientific & Engineering Research**, v. 4, p. 232-236, Dec. 2013. Disponível em: <http://www.ijser.org/researchpaper/Evaluating_renewable_energy_using_data_mining_techniques_in_developing_India.pdf>. Acesso em: 2017 fev. 01.

WITTEN, I. H. et al. **Data Mining: Practical Machine Learning Tools and Techniques**. 4th ed. Cambridge, USA: Charlotte Kent, 2016. 621 p. Disponível em: <https://books.google.com.br/books?hl=pt-BR&lr=&id=1SylCgAAQBAJ&oi=fnd&pg=PP1&dq=Data+Mining_+Practical+Machine+Learning+Tools+and+Techniques-Morgan+Kaufmann&ots=8HHNrioBze&sig=FvBaPoP7JJGsF62DdHICjG-XIQg#v=onepage&q=Data%20Mining_%20Practical%20Machine%20>. Acesso em: 20 dez. 2016.

WEKA. Waikato Environment for Knowledge Analysis. **Weka 3: Data Mining Software in Java**. Disponível em: <<http://www.cs.waikato.ac.nz/ml/weka/>> Acesso em: 4 mar. 2017.

YADAV, A. K.; MALIK, H.; CHANDEL, S. S. ANN based prediction of daily global solar radiation for photovoltaics applications. In: **12th Annual IEEE India Conference (INDICON-2015)**, Dec. 2015, Jamia Millia Islamia, India. **IEEE Conference Publications...** India: IEEE, 31 Mar. 2016, p.1-5. Disponível em: <<http://ieeexplore.ieee.org/document/7443186/>> Acesso em: 8 mar. 2017.

APÊNDICE A – SELEÇÃO DOS DADOS VÁLIDOS

```

import csv
import sys

# Lê os dados das planilhas em formato .CSV e transforma em list
def read(filepath):
    with open(filepath) as file:
        reader = csv.reader(file, delimiter=';')
        data=[]
        for row in reader:
            data.append(row)
        return data

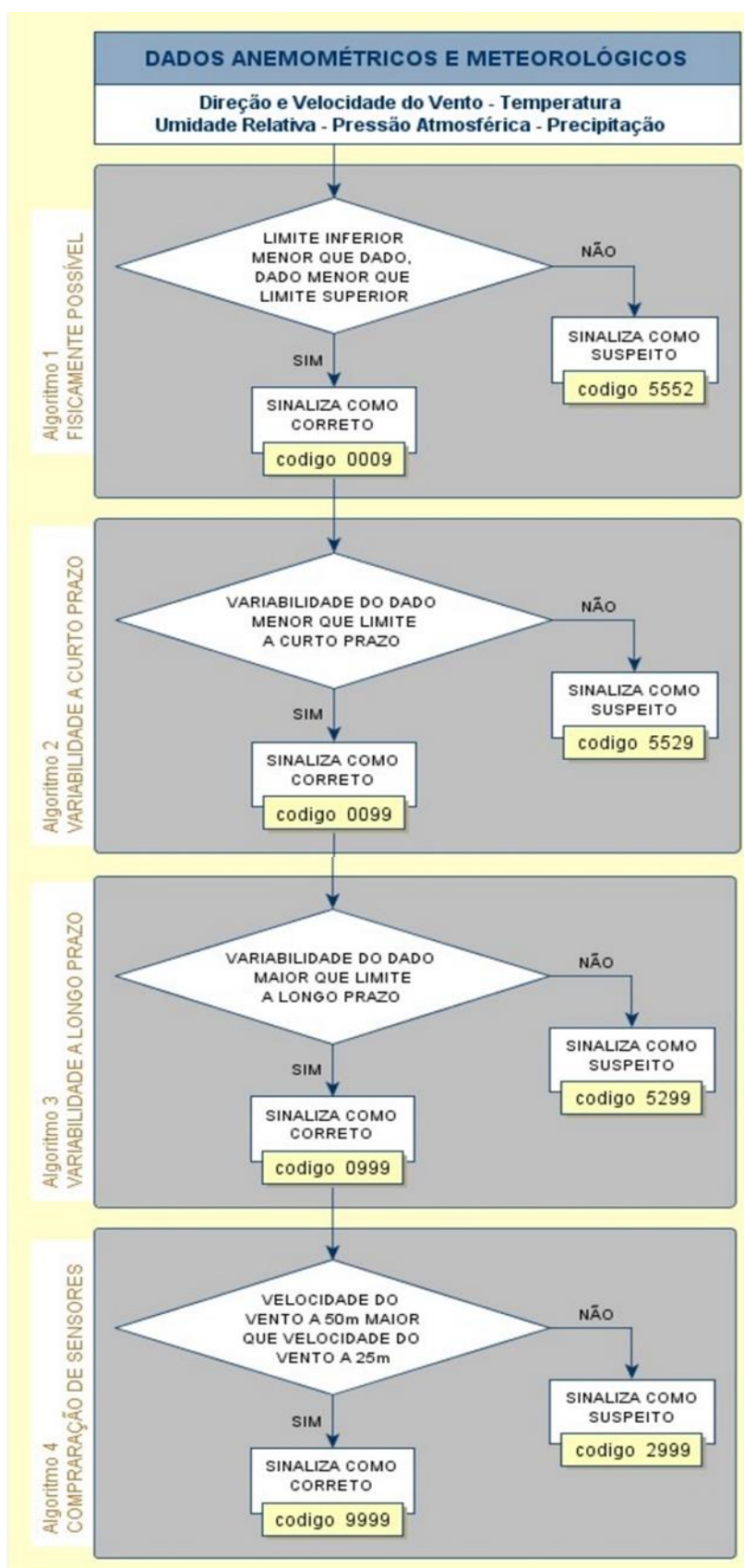
# Seleciona os dados considerados válidos através da planilha que
# sinaliza os dados suspeitos
def select_data(sinalized_data, data, filename):
    valid_data = []
    for i in range(len(sinalized_data)):
        # Colunas dos valores das variáveis coletadas pela estação
        for j in range(4,16):
            # Sinalização dos dados válidos
            if sinalized_data[i][j] != '009' and sinalized_data[i][j]
            != '099' and sinalized_data[i][j] != '999':
                break
            else:
                valid_data.append(data[i])
    with open(filename + '.csv','w',newline='\n') as csvfile:
        writer = csv.writer(csvfile, delimiter=',')
        writer.writerows(valid_data)
    csvfile.close()

if __name__ == '__main__':
    file1, file2, filename = sys.argv[1:4]

    sinalized_data = read(file1)
    data = read(file2)
    select_data(sinalized_data,data,filename)

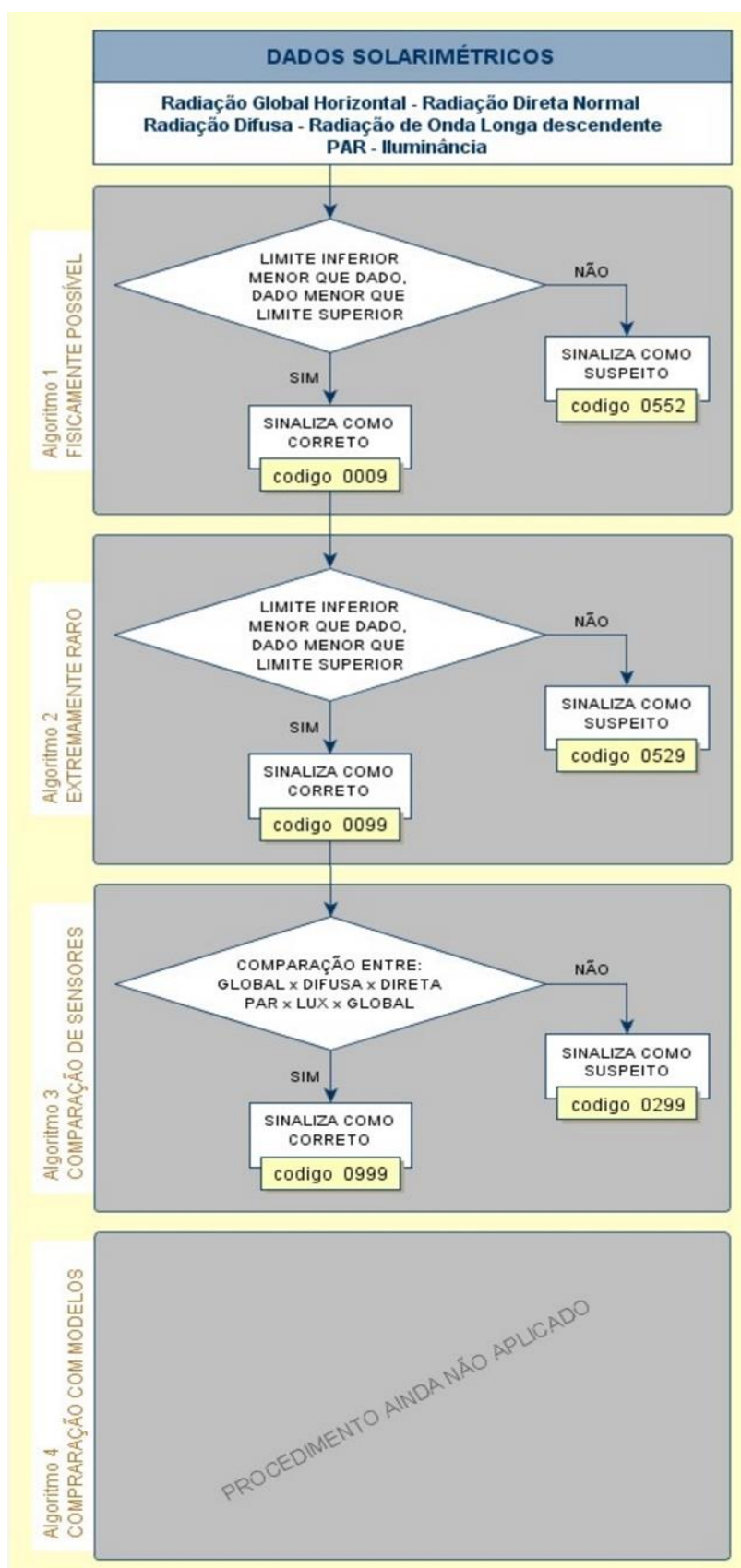
```

ANEXO A – FLUXOGRAMA DO PROCESSO DE VALIDAÇÃO PARA DADOS ANEMOMÉTRICOS E METEOROLÓGICOS



Fonte: SONDA (2017d)

ANEXO B – FLUXOGRAMA DO PROCESSO DE VALIDAÇÃO PARA DADOS SOLARIMÉTRICOS



Fonte: SONDA (2017d)