PREVENDO INCÊNDIOS FLORESTAIS COM MINERAÇÃO DE DADOS: UMA ABORDAGEM UTILIZANDO ÁRVORE DE DECISÃO

PREDICTING WILDFIRES WITH DATA MINING: AN APPROACH USING DECISION TREE

Ede Miranda Junior
https://orcid.org/0000-0001-5151-6742
Centro Paula Souza – Fatec Indaiatuba/SP
ede.miranda@fatec.sp.gov.br

Orientadora: Profa Dra. Maria das Graças J. M. Tomazela
https://orcid.org/0000-0002-5471-2658
Centro Paula Souza – Fatec Indaiatuba/SP
graca.tomazela@fatec.sp.gov.br

RESUMO: Os incêndios florestais têm sido um dos conhecidos e desafiantes problemas da humanidade nos últimos tempos, eles podem trazer grandes danos ambientais e econômicos a uma determinada área, a maior parte disso advêm de atividades humanas e são ainda mais agravantes em determinados períodos dos anos. Mas por meio da evolução tecnológica e dos conhecimentos cada vez mais avançados em processamento e análise de dados, é possível empregar técnicas de mineração de dados para ajudar a prever incêndios florestais utilizando-se de dados climatológicos e de sensoriamento remoto. Esse estudo teve como objetivo, analisar e compreender as técnicas preditivas em mineração de dados para auxiliar na prevenção de incêndios florestais. A Metodologia usada para essa pesquisa foi a experimental. Foram coletados dados climatológicos do OpenWeather e do satélite ERA5, dados de focos de incêndio da base do INPE, e NDVI dos satélites Landsat 7 e Landsat 8, para análise na ferramenta Weka. Como resultado dessa pesquisa foi possível concluir como os dados de climatológicos e de sensoriamento remoto combinados podem ser uma importante fonte de dados para geração de informações, principalmente para monitoramento e previsão de incêndios florestais. Por meio desse estudo foi possível gerar um modelo de classificação utilizando do algoritmo J48 para prever incêndios, com uma acurácia de mais de 84,9%. Espera-se, com o conhecimento gerado sob os dados analisados, subsidiar os planos formulados por ambientalistas de forma que possam utilizar das análises realizadas para atividades mais assertivas quanto à prevenção de incêndios e, consequentemente, a preservação da fauna e flora.

ABSTRACT: Forest fires have been one of the known and challenging problems of humanity in recent times, they can cause great environmental and economic damage to a certain area, most of which come from human activities and are even more aggravating in certain periods of the years. But through technological evolution and increasingly advanced knowledge in data processing and analysis, it is possible to employ data mining techniques to help predict forest fires using climatological data and remote sensing. This study aimed to analyze and understand the predictive techniques in data mining to help prevent forest fires. The methodology used for this research was experimental. Climatological data were collected from OpenWeather and the ERA5 satellite, fires data from INPE base, and NDVI from the Landsat 7 and Landsat 8 satellites, for analysis in the Weka tool. As a result of this research it was possible to conclude how the combined climatological and remote sensing data can be an important source of data for generating information, mainly for monitoring

and forecasting forest fires. Through this study it was possible to generate a classification model using the J48 algorithm to predict fires, with an accuracy of more than 84.9%. With the knowledge generated from the analyzed data, is expected to subsidize the plans formulated by environmentalists so that they can use the analyzes carried out for more assertive activities regarding the prevention of fires and, consequently, the preservation of fauna and flora.

PALAVRAS-CHAVE: Mineração de dados. Incêndios florestais. Predição. Classificação.

KEYWORDS: Data mining. Wildfires. Prediction. Classification.

1 INTRODUÇÃO

Incêndios Florestais são eventos que geram grande impacto social e ambiental devido às queimadas de grandes proporções que isso pode causar em áreas sensíveis e de grande valor para o meio ambiente. Segundo Batista (2000), um incêndio ocorre quando materiais inflamáveis são expostos a materiais acesos, ou seja, para que um incêndio florestal ocorra é necessária uma fonte de calor ativa, resultado de uma combustão.

Segundo Ramos (1995), existem métodos de prevenção de incêndios florestais, o qual varia de região, tipo de bioma e climatologia, dentre as principais atividades dentro de um plano de combate a incêndio florestais em uma determinada área estão: 1) Conhecimento das causas dos incêndios, 2) Caracterização da área; 3) Mapa planialtimétrico/planimétrico da região, 4) Prevenção das fontes de fogo, e; 5) Educação e comunicação da população.

Com a evolução do poder de processamento dos computadores, modelos computacionais para resoluções de problemas estatísticos e matemáticos estão sendo cada vez mais usados, auxiliando a análise de grande quantidade de dados. Uma dessas técnicas é a mineração de dados, que recentemente tem sido explorada para o auxílio do combate ao fogo.

As áreas florestais abrigam uma grande biodiversidade além de apresentar um importante papel ecológico na absorção de carbono e também fornecer oxigênio à atmosfera. Assim, trata-se de uma área valiosa sendo importante controlar e prevenir incêndios nessas áreas. De acordo com o Soares, Santos e Batista (2008), o controle das fontes de risco depende do conhecimento de como essas fontes operam no local, quando e onde os incêndios costumam ocorrer. Os programas de prevenção dependem do registro dessas informações sobre frequência e locais onde os incêndios ocorrem e também de dados como época e extensão das queimadas.

Partindo desse contexto o objetivo dessa pesquisa foi auxiliar a prevenção de incêndio florestais fazendo uso das técnicas preditivas de mineração de dados, visando a analisar com profundidade os padrões de ocorrência de incêndio e, assim, subsidiar os planos formulados por ambientalistas de forma que possam utilizar das análises realizadas para atribuir atividades mais assertivas quanto à prevenção de incêndios e consequentemente preservar a fauna e flora local.

2 METODOLOGIA

Neste trabalho foi utilizada a pesquisa experimental, que implica em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e de observação dos efeitos que a variável produz no objeto (GIL, 2007) As variáveis de controle utilizadas foram: especificação do tipo de poda da árvore de decisão; utilização de diferentes conjuntos de treinamento e teste; utilização de diferentes métodos de seleção de atributos.

Para se realizar esse estudo utilizou-se como ferramenta principal para mineração de dados, o Weka. Ela contém algoritmos e ferramentas de visualização que suportam o aprendizado de máquina e suporta as principais tarefas de mineração de dados,

Para a extração de dados geo-espaciais foi utilizado o Google Earth Engine. Essa plataforma gratuita combina um catálogo de vários petabytes de imagens de satélite e conjuntos de dados geo-espaciais.

Para a primeira etapa no processo de KDD, o pré-processamento, foi utilizada a ferramenta Google Colab. Essa ferramenta oferece, de forma gratuita, um ambiente em nuvem utilizando o ambiente Jupyter Notebook por meio da linguagem Python. Nessa ferramenta foi realizada a limpeza dos dados: removendo dados inconsistentes e faltantes.

2 DESENVOLVIMENTO

O processo de KDD é constituído de várias etapas operacionais, normalmente definidas como: pré-processamento, mineração de dados e pós-processamento. As atividades de pré-processamento podem ser divididas em: 1. **Limpeza** dos dados: etapa na qual são eliminados ruídos e dados inconsistentes; 2. **Integração** dos dados: etapa em que diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados; 3. **Seleção**: etapa na qual são selecionados os atributos que interessam ao processo de descoberta de conhecimento; 4. **Transformação** dos dados: etapa em que os dados são transformados num formato apropriado para aplicação de algoritmos de

mineração. As atividades de pós-processamento são: 1. **Avaliação**: etapa em que são identificados os padrões interessantes; 2. **Visualização dos Resultados**: etapa na qual são utilizadas técnicas de representação do conhecimento para apresentar ao usuário o conhecimento minerado (AMO, 2004; HAN, KAMBER, PEI, 2011).

A mineração de dados, etapa essencial do processo de KDD, é composta por tarefas, definidas também como funcionalidades ou métodos. Consistem na especificação do que se está buscando nos dados, que tipo de regularidades ou categoria de padrões tem-se interesse em encontrar, Para Tsai (2013) as tarefas mais importantes são classificação, predição, associação, *clusterização e* sumarização.

Neste trabalho foi utilizada a tarefa de classificação que é realizada em dois passos: 1) encontrar um modelo para o atributo alvo (também chamado de atributo meta ou classe) como uma função dos valores dos outros atributos; 2) associar as instâncias com classes não conhecidas a uma determinada classe com a maior precisão possível.

A Indução por Árvore de Decisão é uma das principais técnicas de mineração de dados utilizada para a tarefa de classificação (GOLDSCHMIDT e PASSOS, 2005). Essa importância se dá pela sua expressividade simbólica, pois, diferentemente de outras técnicas, é possível entender a estrutura preditiva do modelo, ou seja, quais atributos e seus respectivos valores são mais determinantes para a previsão da classe.

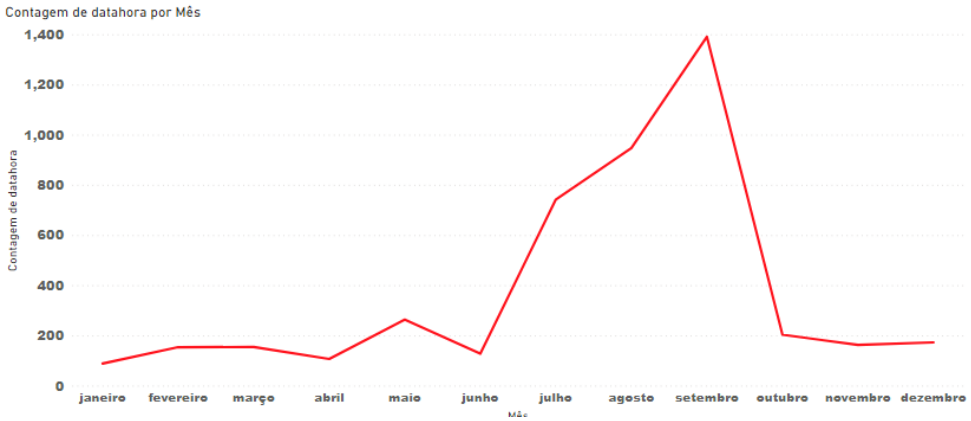
A partir da exploração do conceitos teóricos sobre mineração de dados e sobre como os incêndios florestais se comportam e afetam o meio ambiente, a cadeia alimentar e o ciclo da água., foi realizado um processo de KDD, cujos resultados são apresentados a seguir.

4 RESULTADOS OBTIDOS

A região escolhida para esse experimento foi a cidade de Castilho no extremo oeste do estado de São Paulo e a cidade de Três Lagoas no estado de Mato Grosso do Sul por apresentar uma diversidade biológica devido a sua área de transição entre a Mata Atlântica e o Cerrado, além da cidade de Castilho abrigar o Parque Estadual do Aguapeí, em que apresenta características parecidas com o do Pantanal.

O período dos dados são de 2013 a 2018. O ano de 2017 dentre os anos coletados foi o ano que registrou mais focos de incêndios, e os meses que registraram mais focos durante esse período foram os meses de setembro, agosto e julho, respectivamente, conforme figura 1, devido ao período de estiagem.

Figura 1: Focos de incêndios por mês.



Fonte: Autor

O Quadro 1 descreve os 12 atributos utilizados nesse experimento e as fontes desses atributos.

Quadro 1**:** Atributos utilizados no experimento.

Atributos	Fonte
Mês (1-12)	OpenWeather
Média da temperatura local (°C)	OpenWeather
Média da pressão atmosférica local (hPa)	OpenWeather
Média da umidade local (%)	OpenWeather
Média da velocidade do vento local (m/s)	OpenWeather
Direção do vento local (º)	OpenWeather
Média da cobertura das nuvens local (%)	OpenWeather
Descrição do tempo local (0-1)	OpenWeather
Média do NDVI local (0-1)	Landsat 7 e 8
Média da temperatura do ar local (°C)	ERA5
Média do ponto de orvalho local (°C)	ERA5
Incêndio (True/False)	INPE

Fonte: Autor

Na primeira etapa do experimento após a aquisição dos dados, foram realizadas as atividades de pré-processamento: limpeza, seleção e integração dos dados utilizando a ferramenta Google Colab, aplicando as bibliotecas Pandas e Numpy, e em seguida, para a análise no Weka, foi gerado uma tabela no formato .csv.

No Weka houve a necessidade de realizar o balanceamento das classes, pois os dados meteorológicos coletados do OpenWeather, são dados coletados de hora em hora, ou seja, por dia são feitos 24 registros no *dataset*. Cada linha do *dataset* representa o

horário do registro com um intervalo de 1 hora para o próximo registro. Foi realizada uma busca no conjunto de dados do INPE para verificar em quais horários houve incêndios e integrá-los junto ao conjunto de dados do OpenWeather, sendo *True* para os casos que tiveram incêndios, e *False* para os casos que não houve. Sendo assim o conjunto de dados possui mais dados *False* do que *True.* Foi utilizado o recurso *ClassBalancer* do próprio Weka para corrigir essa disparidade e prover uma melhor classificação.

Na sequência foi realizada a mineração de dados em si. Foi aplicado o algoritmo J48, utilizando do método de seleção de atributo “razão ganho de informação”, implementada pelo próprio algoritmo. O método de treinamento e teste utilizado para primeira execução foi o de *cross-validation* com 11 grupos (*folds*). Na primeira rodada utilizando as configurações padrões do algoritmo, foi obtido uma árvore com 1481 nós. Porém esse modelo estava sobre ajustado para esse conjunto de dados, o que seria falho para prever novos dados. Para a segunda rodada foi ajustado para 30 o número mínimo de objetos por nó. Os resultados são apresentados a seguir na tabela 1 por meio da matriz de confusão.

Tabela 1: Matriz de confusão

n = 55331	Previsto True	Previsto False	Total
Real True	22982,33	4683,17	27665,5
Real Falso	3648,25	24017,25	27665,5
Total	26630,58	28703,42	

Fonte: Autor

Com essa nova rodada a árvore ficou com 689 nós e 345 nós folhas e o classificador obteve uma acurácia de 84,9%. Foram classificadas 46999,58 instâncias corretamente e 8331.42 instâncias incorretamente, representando 84,94% e 15,06% respectivamente. As taxas de Verdadeiro Positivo (VP) para a classe *False* obteve uma acurácia de 0,868 e para *True* 0,831, gerando um peso médio de 0,849. A precisão do modelo para a classe *False* obteve uma acurácia de 83,7% e para *True* 86,3%, o que indica uma boa capacidade preditiva do modelo. Os ramos mais relevantes e que previram mais casos de incêndios são apresentados nas figuras de 2 e 3.

Figura 2: Ramo da árvore com maior número de previsão de incêndios.

```
Media da umidade <= 54
|   Media do ponto de orvalho <= 14.900263
|   |   Media do NDVI <= 0.331702
|   |   |   Media da pressao <= 1014.3
|   |   |   |   Mes <= 9
|   |   |   |   |   Media da temperatura do ar > 23.371881
|   |   |   |   |   |   Media da temperatura do ar <= 29.708795
|   |   |   |   |   |   |   Media do NDVI <= 0.316514
|   |   |   |   |   |   |   |   Media do ponto de orvalho <= 13.311273
|   |   |   |   |   |   |   |   |   Media do NDVI <= 0.283723
|   |   |   |   |   |   |   |   |   |   Media da temperatura do ar <= 29.364618: TRUE (2828.97/99.66)
```

Fonte: Extraído da ferramenta Weka

Na figura 2, nos casos em que houve umidade inferior/igual a 54%, ponto de orvalho inferior/igual a 13,3°C, NVDI inferior/igual a 0,283, pressão inferior/igual a 1014,3 hPa, temperatura do ar maior que 23,3°C e inferior/igual a 29,3°C, nos meses menores que setembro o modelo pode prever 2828,97 focos de incêndios e errou 99,66.

A umidade relativa do ar quando muito baixa afeta diretamente a vegetação, o que reflete no baixo número do NDVI, que quanto mais próximo de 0, significa que mais seca está a vegetação. Outros pontos que corroboram para as previsões do ramo, são o ponto de orvalho e a pressão. Quanto menor a temperatura do ponto de orvalho menos umidade haverá no ar, causando assim a baixa umidade, enquanto que a baixa pressão eleva a temperatura da superfície. Com isso os riscos de incêndio aumentam devido ao estado da vegetação e das condições meteorológicas.

Figura 3: Ramo da árvore com maior número de previsão de incêndios.

```
Media da umidade <= 54
|   Media do ponto de orvalho <= 14.900263
|   |   Media do NDVI <= 0.331702
|   |   |   Media da pressao > 1014.3
|   |   |   |   Media da cobertura das nuvens <= 97
|   |   |   |   |   Media da temperatura do ar > 18.415277
|   |   |   |   |   |   Media da umidade > 24
|   |   |   |   |   |   |   Direção do vento <= 268
|   |   |   |   |   |   |   |   Media da velocidade do vento <= 5.21
|   |   |   |   |   |   |   |   |   Media da temperatura do ar <= 28.386713
|   |   |   |   |   |   |   |   |   |   Media da cobertura das nuvens <= 12
|   |   |   |   |   |   |   |   |   |   |   Mes > 6
|   |   |   |   |   |   |   |   |   |   |   |   Media da temperatura do ar > 20.40481
|   |   |   |   |   |   |   |   |   |   |   |   |   Media do NDVI <= 0.264752: TRUE (757.14/85.5)
```

Fonte: Extraído da ferramenta Weka.

No ramo da figura 3, nos casos em que houve umidade maior que 24% e inferior/igual a 54%, ponto de orvalho inferior/igual a 14,9°C, NVDI inferior/igual a 0,264, pressão superior a 1014,3 hPa, cobertura das nuvens inferior/igual a 12%, temperatura do ar maior que 20,4°C, direção do vento em 268° e velocidade do vento inferior/igual a 5,21 m/s nos meses maiores que junho o modelo pode prever 757,14 focos de incêndios e errou 85,5.

O modelo fez uso dos mesmos atributos do ramo anterior, contendo pequenas variações, no entanto nesse ramo foi acrescido o atributo de direção do vento e velocidade

do vento. O classificador aponta que durante os períodos mais secos, a partir de junho, os ventos em 268º, ou seja, os ventos que seguem entre oeste e sudoeste influenciaram os focos de incêndio estando em uma velocidade próxima a 5 m/s.

5 CONSIDERAÇÕES FINAIS

Por meio da combinação de várias fontes de dados (estações meteorológicas, sensoriamento remoto etc.) foi possível criar um modelo de alta acurácia para a previsão de incêndios. Verificou-se que as causas dos incêndios florestais advêm de condições meteorológicas, como a umidade, a pressão, o ponto de orvalho, e em situações em que a vegetação já se encontra seca, nos casos das métricas de NDVI. Sendo assim modelos como esses podem ser usados para emissões de alertas de probabilidade de incêndios, tanto para as brigadas de incêndios e bombeiros, quanto para a população que vivem próximas as áreas com alta densidade de vegetação.

REFERÊNCIAS

AMO, S. **Técnicas de mineração de dados**. Uberlândia, MG: Minas: Universidade Federal de Uberlândia, 2004. Disponível em: <http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf>. Acesso em: 03 fev. 2011.

BATISTA, A. C. Mapas de risco: uma alternativa para o planejamento de controle de incêndios florestais. **Floresta**, 2000.

GIL, A. C. **Métodos e técnicas de pesquisa social**. São Paulo: Atlas, 2007.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: Um Guia Prático**. Rio de Janeiro, Brasil: Elsevier Editora Ltda., 2005.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. São Francisco, EUA: Morgan Kaufmann Publishers, 2011.

RAMOS, P. C. M. **SISTEMA NACIONAL DE PREVENÇÃO E COMBATE AOS INCÊNDIOS**. I FORUM NACIONAL SOBRE INCÊNDIOS FLORESTAIS. Piracicaba, São Paulo, Brasil: [s.n.]. 1995.

SOARES, R. V.; SANTOS, J. F.; BATISTA, A. C. Some Details of Forest Fire Statistics in Brazil. **Floresta**, 2008.

TSAI, H. H. Knowledge management vs. data mining: research trend, forecast and citation approach. **Expert Systems with Applications**, v. 40, n. 8, p. 3160–3173, 2013.