seminário de iniciação científica
ISSN: 2358-6052
ISSN 2558-6052

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA GERAÇÃO DE MODELOS PREDITIVOS DE GERAÇÃO DE ENERGIA ELÉTRICA: UM ESTUDO DE CASO NOS *CAMPI* IFMG

CARVALHO[1], Ana Caroline Coutinho; CASTRO[2], Lucas dos Anjos de; GIACULI JUNIOR[3], Calebe; FARIA[4], Felipe Lopes de Melo;

1 Ana Caroline Coutinho Carvalho, bolsista CNPq, Curso de Engenharia da Computação, IFMG *Campus* Bambuí, Bambuí-MG; ccarol690@gmail.com
2 Lucas dos Anjos de Castro, Curso de Engenharia da Computação, IFMG *Campus* Bambuí, Valinhos-SP
3 Calebe Giaculi Junior, Pesquisador do IFMG, *Campus* Bambuí; calebe.giaculi@ifmg.edu.br
4 Felipe Lopes de Melo Faria, Pesquisador do IFMG, *Campus* Bambuí; felipe.faria@ifmg.edu.br

RESUMO

O planejamento de gastos de energia elétrica vem se mostrando cada vez mais ser uma tarefa de difícil realização. Já que o ambiente em que grandes instituições operam está constantemente sujeito a mudanças inesperadas, é comum que busquem entender seu consumo de recursos para encaixá-los em suas receitas. O presente trabalho tem como objetivo a aplicação de técnicas de mineração de dados como os algoritmos KNN (k - *Nearest Neighbor*), Regressão Linear Simples e *Random Forests* a fim de gerar modelos preditivos que contribuam na análise de dados sobre a geração de energia das usinas fotovoltaicas de diversos *campi* do IFMG. Os *campi* que foram e que serão analisados são: Bambuí, Ouro Preto e Ribeirão das Neves. As bases de dados analisadas dispõem informações sobre a geração de energia elétrica mês a mês e foram retiradas do portal *SolarView Pro*. O portal possibilita a visualização de seus dados com as granularidades diária, mensal ou anual. Em um primeiro experimento, foram utilizados dados dos *campi* Bambuí, Ouro Preto e Ribeirão das Neves com bases de dados datadas de 2018 a 2020 contendo 36 instâncias e 1 atributo, exceto o *Campus* Bambuí que possui apenas 25 instâncias e 1 atributo. O atributo é caracterizado como data, já as instâncias são os valores de cada mês contendo a geração de energia dada em KWh. Para a análise dos modelos preditivos, serão utilizadas as técnicas de validação MAPE, MAE e RMSE. O trabalho apresenta pesquisas correlatas a fim de observar métricas, técnicas e parâmetros empregados no estado da arte. Emprega-se a ferramenta WEKA para mineração de dados, explorando seus recursos para edição e visualização de dados além de sua biblioteca de algoritmos. Espera-se que este trabalho possa ajudar em tomadas de decisões futuras e assim podendo gerar economia através do uso mais eficiente da usina fotovoltaica.

Palavras-Chaves: mineração de dados; energia fotovoltaica; WEKA.

INTRODUÇÃO:

O planejamento de gastos vem se mostrando cada vez mais ser uma tarefa de difícil realização. Já que o ambiente em que grandes instituições operam está constantemente sujeito a mudanças inesperadas, é comum que busquem entender seu consumo de recursos para encaixá-los em suas receitas. Em organizações de ensino ocorre um paralelo entre constantes demandas por vagas e resultados com a oferta de novos cursos e recursos atrelados a suas fontes de receita (QUEIROZ; QUEIROZ; HÉKIS, 2011).

Métricas para planejamento em si, apresentam todo um campo de estudo robusto e grande gama de problemáticas. Ressaltam-se, à medida que avança a ciência, preocupações com o meio ambiente através da utilização de fontes de energias limpas. Outro ponto importante é que, na medida em que as civilizações crescem e prosperam, acompanha-se uma crescente demanda por energia elétrica. Constata-se que em instituições com um relevante tempo de existência ocorrem volumes consideráveis de informações armazenadas em consequência de documentação, porém, muitas vezes, estes dados não são aproveitados após seu uso burocrático.

seminário de iniciação científica
ISNN: 2358-6052
ISSN 2558-6052

Com os avanços tecnológicos atuais, cada vez mais, as corporações buscam armazenar de forma digital o maior número possível de informações. Encontrando assim, um problema recorrente: grandes bases de dados de difícil entendimento. Mesmo com a atual grande capacidade de processamento de informação, é trabalhoso atrelar sentido ao que é analisado, de maneira a tornar o processamento útil. Um processo de Mineração de Dados apresenta resultados que técnicas estatísticas habituais não são capazes de identificar (SCHUCH *et al.*, 2010).

O presente trabalho propõe a continuação de uma outra pesquisa que teve como foco de estudo dados de geração e de consumo de energia do IFMG *Campus* Bambuí nos anos de 2019 e 2020. Estes dados podem ser encontrados no portal IFMG - *Campus* Bambuí.

METODOLOGIA:

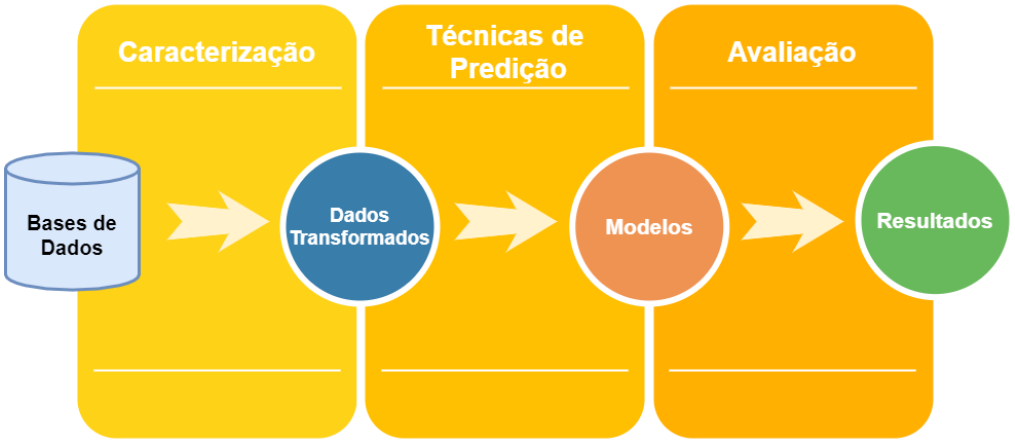
O presente trabalho caracteriza-se quanto a natureza como uma pesquisa aplicada, visando o emprego prático e imediato do conhecimento (MORESI, 2003). Os objetivos se enquadram na classificação de Gil (2002), como pesquisa descritiva, uma vez que se visa descrever e entender as relações das usinas fotovoltaicas de alguns *campi* do IFMG, de maneira a estabelecer relações entre as variáveis que compõem o sistema, além da utilização dos dados de geração de energia.

Quanto aos procedimentos, primeiramente, classifica-se a pesquisa como experimental por trabalhar com variáveis que influenciam os objetos de estudo de maneira a observar os efeitos decorridos (GIL, 2002). Posteriormente, os procedimentos classificam-se também como um estudo de caso por ter como objeto de estudo as usinas fotovoltaicas exclusivamente dos *campi* IFMG (GIL, 2002).

A problemática do trabalho se categoriza conforme explorado por Moresi (2003), como quantitativa por lidar com dados numéricos, conter a aplicação de algoritmos, controlar a precisão por meio de métodos estatísticos, como as medidas de acurácia e promover a análise desses números para tomar conclusões. Por fim, conforme explicado por Wazlawick (2017), o trabalho se adequa à apresentação de algo presumivelmente melhor por comparar a aplicação de diferentes algoritmos para predição.

Para a realização do trabalho, definiram-se três etapas (Figura 1). Na etapa Caracterização foram realizados a escolha, o estudo e a caracterização das bases de dados, contendo os dados de geração de energia elétrica. Na etapa Técnicas de Predição foram aplicadas técnicas computacionais para predição, utilizando os algoritmos k - *Nearest Neighbors* (KNN), *Random Forests* e Regressão Linear Simples. Na etapa Avaliação, realizaram-se a análise dos resultados, utilizando-se de medidas de acurácia dos modelos de predição gerados pelos algoritmos.

Figura 1 - Etapas para realização do trabalho



Fonte: Adaptado de FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996.

Caracterização das Bases de Dados

Os dados possuem apenas uma categoria: informações de geração de energia elétrica que é

ISSN 2558-6052

dada em KWh (*Kilo Watt* hora). Estes dados encontram-se no portal *SolarView Pro*, e o seu acesso foi concedido pelo professor coordenador da pesquisa, variando desde o início da instalação e operação das usinas fotovoltaicas. O portal possibilita o acesso aos dados em diversas granularidades diferentes de tempo, contendo informações de medições diárias, mensais e anuais.

O primeiro experimento realizado, buscou a comparação dos comportamentos dos algoritmos sobre distintas bases de dados de geração de energia elétrica para cada *campus,* possuindo um atributo: a data. Além de 36 instâncias, exceto Bambuí com 25 instâncias. A escolha de cada base se deu pela disponibilidade, de forma a encontrar dificuldades para obtenção de um período de tempo significativo com dados ininterruptos, ou seja, bases com menos dados faltosos, além de analisar quais os *campi* que possuíam dados sobre os mesmos anos em comum.

Neste primeiro experimento foi decidido então estudar as bases de geração de energia agrupadas mensalmente dos *campi* Bambuí, Ouro Preto e Ribeirão das Neves. Em especial, a base Ribeirão das Neves apresentou alguns dados faltosos que foram preenchidos automaticamente através de um *script* que substitui leituras vazias pela média daquele atributo. O experimento foi feito com as datas de janeiro de 2018 a dezembro de 2020.

Técnicas de Predição

Os algoritmos que se mostram promissores dentro do estado da arte e que serão empregados no trabalho são o KNN, Regressão Linear Simples e *Random Forests*. Para o algoritmo KNN foram usados valores para o parâmetro k sendo: 1, 3, 5, 7 e 10. Já na aplicação do *Random Forests* mostra-se necessária a regulagem do parâmetro: *ntree*, definindo o número de árvores da floresta. No trabalho de Breiman (2001), sugere-se os valores para o *ntree* de 500 e 1000. Já na Regressão Linear Simples nenhum parâmetro foi alterado.

Avaliação

A avaliação da precisão preditiva dos modelos de regressão será aferida pela técnica de validação por divisão dos dados em conjuntos de treinos e testes. Como abordado por Santos *et al.* (2019), este tipo de validação expõe se o modelo avaliado e obtém boa performance tanto nos treinamentos quanto em testes, visando aprimorar sua capacidade na predição de valores em situações mais generalizadas. Os autores apontam que dentre as proporções mais utilizadas para a divisão dos dados estão 60:40, 70:30 e 80:20, de maneira que o primeiro valor representa a porcentagem de divisão dos dados para a criação do conjunto de treino e o segundo para o conjunto de testes. A medição da acurácia preditiva é necessária para se verificar quanto os modelos se ajustam aos dados, no presente estudo, a proporção de 70:30 será empregada por ser a indicada pela ferramenta utilizada como divisão padrão, através da opção *Evaluate on held out training*, conforme a documentação (PENTAHO, 2014).

Sendo evidenciados pelo seu amplo uso na literatura, as medidas de acurácia MAPE ou *Mean Absolute Percentage Error* (DE MYTTENAERE *et al*., 2016), MAE ou *Mean Absolute Error* (TAREEN *et al.*, 2019) e RMSE ou *Root Mean Squared Error* (RAFIQUE *et al*., 2020) foram utilizadas.

De Myttenaere *et al.* (2016) explicam que o MAPE é amplamente utilizado por apresentar uma interpretação muito intuitiva do erro e possui ampla relevância nos campos de finanças e na definição de preços para produtos, sendo que, quanto menor o valor indicado menor o erro apresentado pelo modelo. Os autores ainda fomentam o seu uso como medida de validação para predição do consumo de energia elétrica, ponto reforçado ao analisar-se a quantidade de trabalhos científicos apresentados no estado da arte que empregam a métrica.

O MAE, como explorado por Tareen *et al.* (2019), avalia a extensão média dos erros sobre as predições, ignorando sua direção e atribuindo peso igual para as diferenças entre os valores reais e preditos e é expresso na grandeza do valor predito. Os autores indicam que quanto menor o valor obtido de MAE, mais preciso é o modelo.

Já em Rafique *et al.* (2020, p. 4, tradução nossa), é evidenciado que “O RMSE mede a

seminário de iniciação científica
ISNN: 2358-6052
ISSN 2558-6052

discrepância entre os valores de concentração de *radon* medidos e preditos. Valores menores de RMSE indicam incongruências menores". Além disso, de maneira semelhante ao MAE, o RMSE é medido na grandeza da variável predita, ou seja, em um estudo de demanda elétrica é expressa em KWh. Cada medida tem suas particularidades e se destaca em um aspecto diferente, sendo difícil eleger se uma pode ser superior a outra. Em um trabalho comparativo entre medidas, Chai e Draxler (2014) destacam que o ideal é realizar uma combinação entre métricas diferentes. As medidas empregadas são definidas como:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|\bar{y}i - yi|}{yi} \qquad MAE = \frac{1}{N}\sum_{i=1}^{N}|\overline{yi} - yi| \qquad RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|\overline{yi} - yi|^2}$$

Sendo que N representa a quantidade de registros de teste, e $\bar{y}i$ e yi são, respectivamente, os valores preditos e reais.

RESULTADOS E DISCUSSÕES:

Com os algoritmos executados resultou-se nas seguintes tabelas onde mostra de forma destacada qual o melhor e o pior resultado das técnicas, como uma análise inicial deste trabalho, uma vez que ele não está concluído. As análises foram feitas através do valor do MAPE (Erro Percentual Absoluto Médio). De Myttenaere *et al*. (2016) explicam que o MAPE é amplamente utilizado por apresentar uma interpretação muito intuitiva do erro, sendo que, quanto menor o valor indicado menor o erro apresentado pelo modelo.

Tabela 1: Resultado base dados Bambuí

Métricas	Regressão Linear	Random Forests ntree = 500	Random Forests ntree = 1000	KNN 1	KNN 5	KNN 10
MAE (KWh)	328,19	304,86	**303,48**	301,24	**403,64**	341,58
MAPE	13,64%	12,34%	**12,28%**	13,58%	**16,33%**	14,64%
RMSE (KWh)	392,41	386,32	**385,98**	328,70	**499,34**	385,05

Fonte: Elaborado pela autora, 2022.

Tabela 2: Resultado base dados Ouro Preto

Métricas	Regressão Linear	Random Forests ntree = 500	Random Forests ntree = 1000	KNN 1	KNN 5	KNN 10
MAE (KWh)	**307,20**	285,17	287,03	301,65	245,48	**231,54**
MAPE	**12,55%**	11,22%	11,36%	12,35%	9,79%	**9,20%**
RMSE (KWh)	**348,68**	340,53	340,85	381,18	302,23	**295,15**

Fonte: Elaborado pela autora, 2022.

ISSN 2558-6052

Tabela 3: Resultado base dados Ribeirão Preto

Métricas	Regressão Linear	Random Forests ntree = 500	Random Forests ntree = 1000	KNN 1	KNN 5	KNN 10
MAE (KWh)	**711,54**	563,69	571,73	**420,44**	490,80	509,78
MAPE	**38,10%**	33,07%	33,27%	**27,13%**	30,79%	31,50%
RMSE (KWh)	**1038,84**	873,58	876,50	**748,78**	830,05	841,97

Fonte: Elaborado pela autora, 2022.

Analisando o *Campus* Bambuí percebe-se que o melhor resultado foi com a técnica *Random Forests*, tanto com o parâmetro *ntree* valendo 500 e 1000 os valores foram bem próximos um do outro, onde MAPE = 12,28%, MAE = 303,48 e RMSE = 385,98. Já o pior resultado foi com a técnica KNN com o valor do parâmetro k valendo 5 onde o MAPE =16,33%, MAE = 403,64 e RMSE = 499,34.

O *Campus* Ouro Preto obteve o melhor resultado com técnica KNN com o valor do parâmetro k valendo 10, onde o MAPE = 9,20%, MAE = 231,54 e RMSE = 295,15. Já seu pior resultado foi com a técnica Regressão Linear com o MAPE = 12,55%, MAE = 307,20 e RMSE = 348,68.

O *Campus* Ribeirão das Neves teve seus valores maiores aos dos outros *campi* possivelmente por se tratar de uma base que teve seus dados preenchidos de forma automática, influenciando nos resultados. A técnica que teve o melhor resultado foi o KNN com seu parâmetro k valendo 1, onde MAPE = 27,13%, MAE = 420,44 e RMSE = 748,78. Já o pior resultado foi com a técnica Regressão Linear com o MAPE = 38,10%, MAE = 711,54 e RMSE = 1038,84.

CONCLUSÕES:

O presente trabalho aplica técnicas de mineração de dados sobre informações de geração de energia elétrica, buscando avaliar o poder preditivo dos modelos gerados e a possibilidade de auxiliar os *campi* a entender o perfil deste recurso. O experimento conduzido teve dois resultados negativos para a técnica Regressão Linear Simples levando dois *campi* a terem como seus piores resultados: Ouro Preto e Ribeirão das Neves e um *campi* com a técnica KNN com o parâmetro k valendo 5: Bambuí. Já os melhores resultados foram com as técnicas *Random Forests* com *ntree* valendo 1000: Bambuí e KNN valendo 10 e 1: Ouro Preto e Ribeirão das Neves, respectivamente.

Dados meteorológicos serão introduzidos ao estudo como uma tentativa de enriquecer a análise proposta, uma vez que a análise apresentada neste trabalho ainda não está concluída. As informações meteorológicas são disponibilizadas através do Banco de Dados Meteorológicos do Instituto Nacional de Meteorologia. Sendo assim, já está em andamento um novo experimento, que irá incluir dados meteorológicos como atributos do primeiro experimento e logo em seguida também será feito o segundo experimento contendo os *campi* Governador Valadares, Ouro Preto e São João Evangelista com os anos de 2017 a 2021 com os atributos ano e geração de energia, posteriormente com dados meteorológicos como novos atributos.

Ao fim do trabalho, é esperado que as informações provenientes das predições auxiliem os *campi* a se planejarem quanto a geração de energia, pretendendo fornecer uma previsão para se tornarem autossuficientes em energia elétrica, desejando inspirar a utilização de energias limpas. Espera-se que em uma próxima pesquisa possa reduzir a granularidade mensal para diária.

seminário de iniciação científica
ISNN: 2358-6052
ISSN 2558-6052

REFERÊNCIAS BIBLIOGRÁFICAS:

CHAI, Tianfeng; DRAXLER, Roland R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. **Geoscientific model development,** v. 7, n. 3, p. 1247-1250, 2014. Disponível em: https://gmd.copernicus.org/articles/7/1247/2014/. Acesso em 23 abr. 2022. D. W. Aha. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. **International Journal of Man-Machine Studies**, 36(2):267-287, 1992. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/002073739290018G. Acesso em: 07 abr. 2022.

DE MYTTENAERE, Arnaud De; GOLDEN, Boris; LE GRAND, Bénédicte; ROSSI, Fabrice. Mean absolute percentage error for regression models. **Neurocomputing**, v. 192, p. 38-48, 2016. Disponível em: https://arxiv.org/abs/1605.02541. Acesso em 12 abr. 2022.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2002.

JAWAD, Muhammad; NADEEM, Malik Sajjad Ahmed; SHIM, Seong-O; KHAN, Ishtiaq Rasool; SHAHEEN, Aliya; HABIB, Nazneen; HUSSAIN, Lal; AZIZ, Wajid. Machine Learning Based Cost Effective Electricity Load Forecasting Model Using Correlated Meteorological Parameters. **IEEE Access**, [S.L.], v. 8, p. 146847-146864, 20 ago. 2020. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/access.2020.3014086. Acesso em: 04 abr. 2022.

MORESI, Eduardo. Metodologia da pesquisa. Brasília: **Universidade Católica de Brasília**, v. 108, p. 24, 2003. Disponível em: http://www.inf.ufes.br/~pdcosta/ensino/2010-2-metodologia-de-pesquisa/MetodologiaPesquisa-Moresi2003.pdf. Acesso em: 31 mar. 2022.

PENTAHO. **Time Series Analysis and Forecasting with Weka.** Software. Versão 1.0.27.

QUEIROZ, Jamerson Viegas; QUEIROZ, Fernanda Cristina Barbosa Pereira; HÉKIS, Hélio Roberto. Gestão estratégica e financeira das Instituições de ensino superior: um estudo de caso. **Iberoamerican Journal Of Industrial Engineering**, Florianópolis, v. 3, n. 1, p.98-117, jul. 2011. Semestral. Disponível em: http://incubadora.periodicos.ufsc.br/index.php/IJIE/article/view/504/pdf. Acesso em: 30 mar. 2022.

RAFIQUE, Muhammad; TAREEN, Aleem Dad Khan; MIR, Adil Aslim; NADEEM, Malik Sajjad Ahmed; ASIM, Khawaja M.; KEARFOTT, Kimberlee Jane. Delegated Regressor, A Robust Approach for Automated Anomaly Detection in the Soil Radon Time Series Data. **Scientific Reports**, [S.L.], v. 10, n. 1, 20 fev. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-020-59881-9. Acesso em: 20 abr. 2021.

SANTOS, Hellen Geremias dos; NASCIMENTO, Carla Ferreira do; IZBICKI, Rafael; DUARTE, Yeda Aparecida de Oliveira; CHIAVEGATTO FILHO, Alexandre Dias Porto. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. **Cadernos de Saúde Pública**, [S.L.], v. 35, n. 7, 2019. UNIFESP (SciELO). http://dx.doi.org/10.1590/0102-311x00050818. Acesso em: 10 abr. 2022.

SCHUCH, Regis; DILL, Sérgio Luis; SUASEN, Paulo Sérgio; PADOIN, Edson Luis; CAMPOS, Mauricio de. Mineração de dados em uma subestação de energia elétrica. In: **Proceedings of the 9th Brazilian Conference on Dynamics, Control and Their Applications–dincon**. 2010. p. 804. Disponível em: http://sbmac.locaweb.com.br/dincon/trabalhos/PDF/energy/68015.pdf. Acesso em: 30 mar. 2022.

TAREEN, Aleem Dad Khan; ASIM, Khawaja M.; KEARFOTT, Kimberlee Jane; RAFIQUE, Muhammad; NADEEM, Malik Sajjad Ahmed; IQBAL, Talat; RAHMAN, Saeed Ur. Automated anomalous behaviour detection in soil radon gas prior to earthquakes using computational intelligence techniques. **Journal of Environmental Radioactivity**, [S.L.], v. 203, p. 48-54, jul. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.jenvrad.2019.03.003. Acesso em: 17 abr. 2022.

ISSN 2558-6052

TERRA, Guilherme Saad. **Uma Metodologia de Mineração de Dados para previsão de Cargas**. Rio de Janeiro, 2003. Disponível em: http://www.coc.ufrj.br/pt/teses-de-doutorado/147-2003/943-guilherme-saad-terra. Acesso em: 02 abr. 2022.

WAZLAWICK, Raul. **Metodologia de pesquisa para ciência da computação.** Elsevier Brasil, 2017.

WEKA. **The University of Waikato.** Software. Versão 3.8.5. [S.L.], 2016. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/. Acesso em: 12 nov. 2021.