Redução de Dimensionalidade Via Análise de Componentes Principais de Variáveis Inerentes à Geração de Energia Hidrelétrica

Gabriel de Campos* Mateus G. Santos* Pedro Paulo C. Viana**
Marcelo O. Fonseca Guilherme S. Bastos***

** Instituto de Engenharia de Sistemas e Tecnologia da Informação, Universidade Federal de Itajubá, Itajubá-MG, Brasil.*
*** Jirau Energia, Porto Velho-RO, Brasil.*

E-mails: gabrieldecampos.98@gmail.com, mateus.gabriel@unifei.edu.br, pedro.viana@jirauenergia.com.br, marcelo.fonseca@jirauenergia.com.br, sousa@unifei.edu.br

Abstract: The increased amount of monitored data allows more complex analysis and optimization techniques. As a result of the high dimensional databases, the computational cost has become a critical point. This paper proposes the application of the Principal Component Analysis (P.C.A.) to reduce high dimensional databases, applied to a hydroelectrical generation dataset, and evaluate its impacts on supervised machine learning techniques prediction. The findings of this study showed that by applying the P.C.A. the reduced dataset was able to preserve most of the data variability of original data and achieved a similar performance rate on machine learning techniques, compared to the non-reduced dataset.

Resumo: A crescente quantidade de dados disponíveis vem possibilitando técnicas de otimização e análises mais complexas. Como consequência da alta dimensionalidade, o custo computacional envolvido vem se tornando um ponto crítico. Diante deste cenário, este artigo propõe a aplicação do método de Análise de Componentes Principais (A.C.P.) para redução de dimensionalidade no ambiente de geração hidrelétrica, e avalia seu impacto no desempenho de técnicas de aprendizado de máquina supervisionado na predição do consumo. Ao aplicar a A.C.P., verificou-se que os dados reduzidos mantiveram a variabilidade de informações dos dados originais e obtiveram desempenho de predição semelhante ao obtido sem redução.

Keywords: Principal component analysis; Data reduction; Hydropower; Artificial intelligence; systems optimization.

Palavras-chaves: Análise de componentes principais; Redução de dimensionalidade; Energia hidrelétrica; Inteligência artificial; Otimização de sistemas.

1. INTRODUÇÃO

O uso de energias renováveis vem ganhando grande destaque com a crescente demanda de eletricidade e com o aumento da preocupação dos impactos ambientais de sua geração (IRENA, 2021). Destas, a energia proveniente de hidrelétricas se destaca por ser uma das mais antigas fontes de baixa emissão de carbono e, dentre as renováveis, a mais presente na matriz energética mundial. No Brasil, a parcela de geração de energia por hidrelétricas corresponde a aproximadamente 65% da geração total, ou 76,7% do total de energia de origem renovável (Ritchie et al. 2020).

A fim de otimizar a produção de energia hidrelétrica, diversos estudos ao longo das décadas buscaram técnicas para aumentar a eficiência da geração, reduzir seus custos e impactos ambientais, além de formas de gerenciar o despacho das Unidades Geradoras (U.G.) visando a otimização do planejamento energético de curto prazo, longo prazo ou em tempo real.

Devido a crescente complexidade da modelagem e contínuo aumento da quantidade de variáveis monitoradas, métodos que utilizam de Inteligência Artificial (IA) vêm ganhando espaço na busca pela otimização na operação de usinas hidrelétricas. No entanto, o uso dessas técnicas gera um grande custo computacional, por vezes tornando inviável seu desenvolvimento em computadores convencionais devido à alta dimensionalidade dos dados analisados (Feng et al. 2019). Isto é, a grande quantidade de variáveis analisadas gera um alto custo de tempo de execução do algoritmo e elevando consumo de memória de armazenamento e de trabalho.

Neste contexto, o presente trabalho propõe a utilização do método de redução de dimensionalidade por Análise de Componentes Principais (A.C.P.) voltado para a otimização da geração de energia hidrelétrica. Para fins de verificação do desempenho, técnicas de IA foram aplicadas para predição do

consumo de Serviços Auxiliares (S.A.) de uma Usina Hidrelétrica (UHE) com ambos os conjuntos de dados: sem redução e com redução.

A técnica de A.C.P. é comumente aplicada no pré-processamento dos dados para aplicação de técnicas de IA (Brunton et al. 2019), reduzindo a dimensionalidade dos dados analisados e, consequentemente, o custo computacional dos modelos de IA. A A.C.P. realiza uma transformação dos dados originais para um novo conjunto de variáveis denominadas Componentes Principais (C.P.) (Abdi et al. 2010), que serão utilizadas nas análises subsequentes.

Este artigo apresenta inicialmente uma revisão dos conceitos básicos de A.C.P. e de algumas técnicas de IA, nas seções 2 e 3, seguidos por um estudo de caso e resultados obtidos, nas seções 4 e 5. Para o estudo de caso, foram utilizados dados de geração e consumo interno da UHE de Jirau, Rondônia, entre 2017 e 2020.

2. ANÁLISE DE COMPONENTES PRINCIPAIS

Técnicas de redução de dimensionalidade, como A.C.P., são utilizadas no processamento de dados de alta dimensionalidade, sendo parte do conjunto de ferramentas de estatística multivariada (Abdi et al. 2010). Estas técnicas por si só podem ser aplicadas diretamente para a análise de relação entre múltiplas variáveis, compressão dos dados, ou ainda no pré-processamento de dados para análises subsequentes, como modelos de predição (Sousa et al. 2007), análise de imagens (Celik, 2009), entre outros.

O conjunto de variáveis reduzidas não possui significado físico, mas, de forma mais eficiente, preserva a variabilidade de informações presentes nos dados originais.

Variações do método A.C.P. vêm sendo aplicadas com objetivo de otimizar a redução de dimensionalidade, atribuindo características como maior rejeição a outliers (Candes et al. 2011), redução de ruido (Yata et al. 2012) ou para lidar com Big Data (Zhang et al. 2016). Outras técnicas de redução de dimensionalidade comumente utilizadas são: Análise de Fatores (Yong et al. 2013), regressão por redução de posto (Chen et al. 2012) e análise de agrupamento (Borland et al. 2001).

2.1 Conceitos Básicos de A.C.P.

Para realizar a redução de dimensionalidade por A.C.P., é calculado um novo conjunto de variáveis, C.P., obtidas através da combinação linear dos dados originais. A primeira componente, CP_1, possuirá a maior variabilidade de informações do conjunto original possível. A segunda, CP_2, deve ser ortogonal a primeira e agregar a maior variabilidade possível à nova base. Para um conjunto de n variáveis originais linearmente independentes, a n-ésima componente, CP_n, possuirá a menor variabilidade e será ortogonal as demais. O conjunto das n C.P. agregará 100% da variabilidade dos dados originais, tratando-se então de uma transformação de base, sem redução de dimensionalidade.

A Figura 1 é uma representação gráfica de um conjunto de dados de duas dimensões, x e y (a). Os respectivos autovetores das C.P. são representados em azul e, em vermelho, são representadas as elipsoides dos três primeiros desvios padrões (b).

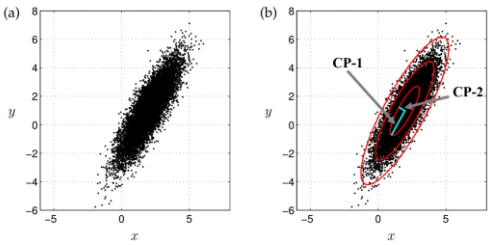

Fig. 1 Representação gráfica das C.P. – Adaptado (Brunton et al. 2019).

Neste exemplo, a CP_1 possui uma variabilidade significativamente maior quando comparada a CP_2 . As C.P. de menor variabilidade podem ser excluídas de análises posteriores de acordo com o critério de seleção adotado, de tal forma a reduzir a dimensionalidade dos dados e preservar a variabilidade de informações. Desta forma, podemos realizar a transformação de base, de n variáveis originais para p C.P., onde $p \leq n$.

Considerando X a matriz $m\,x\,n$ de dados originais, de n variáveis originais $[X_1\ X_2\ \ldots\ X_n]$ e m amostras:

$$X = [X_1\ X_2\ \ldots\ X_n] = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \quad (1)$$

E Z uma matriz $m\,x\,n$ de dados transformados $[Z1, Z2, \ldots, Zn]$, com m amostras para n C.P.,

$$Z = [Z_1\ Z_2\ \ldots\ Z_n] = \begin{bmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & \ddots & \vdots \\ z_{m1} & \cdots & z_{mn} \end{bmatrix} \quad (2)$$

Onde Z tem potencial de redução de dimensionalidade para p C.P.

$$Z' = [Z_1\ Z_2\ \ldots\ Z_p] \quad (3)$$

A transformação linear de uma amostra i , $1 \leq i \leq m$, para a nova base de C.P. é dada por

$$\begin{cases} z_{i1} = x_{i1}\ v_{11} + x_{i2}\ v_{12} + \cdots + x_{in}\ v_{1n} \\ z_{i2} = x_{i1}\ v_{21} + x_{i2}\ v_{22} + \cdots + x_{in}\ v_{2n} \\ \vdots \\ z_{in} = x_{i1}\ v_{n1} + x_{i2}\ v_{n2} + \cdots + x_{in}\ v_{nn} \end{cases} \quad (4)$$

Ou

$$Z = X * V^T \quad (5)$$

A matriz Z, matriz de dados convertidos em termos das C.P., é denominada matriz de escores, e V é a matriz $n\ x\ n$ de pesos v_{ij} da combinação linear, $1 \leq i, j \leq n$.

2.2 Solução das Componentes Principais

O método de A.C.P. tem por essência calcular a matriz de pesos V. Para isso, diferentes variações desta técnica aplicam diferentes algoritmos, a fim de otimizar características desejadas, conforme mencionado anteriormente nesta seção.

Para o método A.C.P. clássico, dado o conjunto de n variáveis originais e m amostras definido em (1), calcula-se sua respectiva matriz $n\ x\ n$ de covariância, Σ, onde a covariância entre a *i*-ésima e *j*-ésima variáveis é dada por

$$\sigma_{ij}^2 = \frac{1}{m}\sum_{k=1}^{m}(x_{k\,i} - \bar{x}_i)\left(x_{k\,j} - \bar{x}_j\right) \tag{6}$$

e $\bar{x}_i$ e $\bar{x}_j$ são as médias das variáveis *i* e *j*, respectivamente. A solução dos pesos V é dada pelos autovetores e autovalores de Σ.

$$\Sigma\, V = \lambda\, V \tag{7}$$

$$(\Sigma - \lambda\, \mathrm{I}) = 0 \tag{8}$$

Onde I é matriz identidade de ordem n, V o vetor de n autovetores v_i e λ é o vetor de n autovalores λ_i, para $1 \leq i \leq n$ e $\lambda_1 \leq \lambda_2 \ldots \leq \lambda_n$ (Johnson et al. 1998).

As relações entre os autovalores e autovetores com as C.P. e suas variâncias são deduzidas utilizando a técnica de decomposição por valores singulares, como demonstra Brunton et al. (2019).

Desta forma, o autovetor v_i está associado aos pesos da combinação linear das variáveis originais para as C.P. Já o autovalor λ_i, por sua vez está relacionado à variância, σ_i^2, dos dados transformados Z_i, referente a variável CP_i.

$$\sigma_i^2 = \lambda_i = \frac{1}{m}\sum_{k=1}^{m}(z_{k\,i} - \bar{z}_i)^2 \tag{9}$$

A contribuição de cada C.P. é expressa em termos proporcionais da variância, obtida pela razão entre a variância da componente e a soma das variâncias das n C.P. Esta proporção é denominada Explicação da Variância Proporcional (E.V.P.):

$$EVP_i = \frac{\lambda_i}{\sum_{k=1}^{n}\lambda_k} \tag{10}$$

Já o valor acumulado das variâncias proporcional das i C.P. anteriores, incluindo a CP_i, é denominado Explicação da Variância Proporcional Acumulada (E.V.P.A.):

$$EVPA_i = \sum_{k=1}^{i} EVP_i \tag{11}$$

Ao se utilizar A.C.P. com um conjunto de variáveis X com escalas muito diferentes, é recomendada a normalização prévia dos dados, tanto durante a A.C.P. como nas análises posteriores, para evitar a dominância de variáveis de maiores escalas sobre as de menores escalas.

A variável normalizada X_i' é dada por

$$X_i' = \frac{X_i - \bar{x}_i}{\sigma_i} \tag{12}$$

Onde $\bar{x}_i$ e σ_i são, respectivamente, a média e o desvio padrão das *m* amostras de X_i.

Ao aplicar as variáveis normalizadas para o cálculo da matriz de covariância, tem-se que:

$$\sigma_{ij}^2 = \frac{1}{m}\sum_{k=1}^{m}(x'_{k\,i} - 0)\left(x'_{k\,j} - 0\right) \tag{13}$$

$$\sigma_{ij}^2 = \frac{1}{m}\sum_{k=1}^{m}\left(\frac{x_{ki} - \bar{x}_i}{\sigma_i}\right)\left(\frac{x_{kj} - \bar{x}_j}{\sigma_j}\right) \tag{14}$$

$$\sigma_{ij}^2 = \frac{1}{m}\frac{\sum_{k=1}^{m}(x_{k\,i} - \bar{x}_i)\left(x_{k\,j} - \bar{x}_j\right)}{\sigma_i\,\sigma_j} \tag{15}$$

A equação da covariância com dados normalizados (15) é equivalente à correlação, r_{ij}, entre as variáveis i e j:

$$r_{ij} = \frac{1}{m}\frac{\sum_{k=1}^{m}(x_{k\,i} - \bar{x}_i)\left(x_{k\,j} - \bar{x}_j\right)}{\sigma_i\,\sigma_j} = \frac{\sigma_{ij}^2}{\sigma_i\,\sigma_j} \tag{16}$$

Portanto, para cálculo dos autovalores e autovetores dos dados normalizados, utiliza-se a matriz de correlação R, $n\ x\ n$, dos dados sem normalização.

$$R\, V = \lambda\, V \tag{17}$$

$$(R - \lambda\, \mathrm{I}) = 0 \tag{18}$$

Desta forma, diferenciam-se dois métodos de obtenção da solução da A.C.P.: por matriz de covariância, em (7) e (8), e por matriz de correlação, em (17) e (18).

2.3 Resultados e Desempenho das C.P.

Inicialmente, a quantidade de C.P. p e a quantidade de variáveis originais n, são iguais, $p = n$, para dados originais linearmente independentes. No entanto, ao manter a dimensionalidade dos dados, não haverá mudança no custo computacional das análises subsequentes. Desta forma, tem-se por objetivo $p \ll n$, obtido ao descartar as componentes de menor variabilidade.

Um indicativo do potencial de redução de dimensionalidade dos dados originais é obtido através da análise das correlações das variáveis originais. Dados originais com correlação forte são mais propícios a serem representados com menos variáveis, ao passo que a baixa correlação entre variáveis pode, em casos extremos, levar um novo conjunto de dados onde as C.P. são as próprias variáveis originais.

O potencial de redução da nova base deve ser avaliado de forma a diminuir ao máximo a dimensionalidade, sem perda significativa das informações. No entanto, esta é uma questão ainda em aberto (Abdi et al. 2010).

A seguir, algumas das regras utilizadas para definir a redução de C.P. apresentadas por Savegnago et al. (2011):

- E.V.P.A. acima de 80%, ou algum valor pré-definido. Este valor pode variar de acordo com a referência utilizada;
- Selecionar C.P. com autovalores maiores que a média de autovalores.

3. TÉCNICAS DE APRENDIZADO DE MÁQUINA

A fim de avaliar e comparar o desempenho dos dados originais e as diferentes combinações de dados reduzidos, optou-se por aplicar algumas das técnicas de IA mais comuns.

Por não se tratar do foco do presente trabalho, as técnicas utilizadas não foram otimizadas.

3.1 Introdução

Técnicas de IA vêm sendo largamente aplicadas nos mais diversos campos profissionais e acadêmicos. Pode-se citar a crescente utilização na área da saúde, engenharia, economia, entre outras.

No entanto, o conceito de IA é amplo, podendo ser definida como ciência e engenharia que produz máquinas inteligentes que lidem com o mundo de forma não-pior que um humano (Dobrev et al. 2005), sem necessariamente envolver aprendizado.

As técnicas de IA que possuem algoritmos com aprendizado utilizando grandes conjuntos de dados são denominadas de técnicas de Machine Learning (ML), ou aprendizado de máquina (Choi et al. 2020).

Dentro da área de aprendizado existem diversos outros segmentos, especializados em problemas e soluções especificas. Neste trabalho serão utilizadas técnicas de aprendizado de máquina supervisionado.

As técnicas de aprendizado supervisionado buscam criar um modelo de predição através de um histórico de dados de entrada e saída (James et al. 2013). Este modelo tem por objetivo estimar novos valores de saída para novos dados de entrada.

3.2 Técnicas Utilizadas

Para escolher as técnicas de ML a serem utilizadas, deve-se atentar às características dos dados utilizados, da aplicação, entre outros pontos.

Para a aplicação abordada neste artigo, foram escolhidas técnicas de aprendizado por regressão supervisionada (Choi et al. 2020), isto é, utiliza-se de dados históricos de entrada e saída para criação do modelo, cuja saída é numérica quantitativa.

3.2.1 Árvore de decisão

O método de árvore de decisão inicia eu um nó de origem que se divide em outros dois nós. A decisão de para qual nó prosseguir é tomada a partir de comparações das variáveis de entrada com valores obtidos durante o aprendizado. Os nós seguintes seguem a mesma estrutura de decisão. Por fim, após percorrer um número máximo de nós, definido pela profundidade da árvore, um valor final é obtido como resultado da última decisão. (James et al. 2013)

3.2.2 Multi Layer Perceptron (M.L.P.)

A técnica M.L.P. é um dos tipos mais simples de rede neural artificial, onde existem nós de entrada, nós de saída e, interligando estes extremos, existem grupos de nós separados em camadas. Cada camada interna, chamadas de camada oculta, possui nós que são excitados pela combinação dos nós da camada anterior de acordo com sua função de ativação. A excitação do nó determina seu sinal de saída que irá excitar os nós da camada seguinte (Hastie et al. 2009).

3.2.3 Regressão Linear

O método de regressão linear utiliza de uma combinação linear das variáveis de entrada para predizer o valor de saída. Os pesos das combinações lineares são calculados de forma a minimizar o erro da aproximação (James et al. 2013).

3.2.4 Vizinhança

Na técnica de Vizinhança, ou K-vizinhos mais próximos, o valor de previsão é estimado com base na média das K amostras mais próximas (James et al. 2013).

3.3 Performance do Modelo

Para avaliar o desempenho do modelo gerado, são utilizadas métricas estatísticas do erro entre os dados históricos de saída e a previsão de saída.

Considerando os dados de saída y, a saída prevista $\hat{y}$, e n a quantidade de amostras de y e $\hat{y}$, algumas das métricas comumente utilizadas são:

- Erro médio absoluto (E.M.A.):

$$EMA = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{19}$$

- Coeficiente de determinação - R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{20}$$

Para analisar o impacto no tempo de execução dos algoritmos, utilizou-se o Contador de Performance, disponibilizado no sistema operacional como ferramenta de monitoramento de performance.

4. ESTUDO DE CASO

Localizada a 120 km de Porto Velho, Rondônia, a hidrelétrica de Jirau foi inaugurada em 2012, fazendo parte do complexo de hidrelétricas do rio Madeira. Atualmente conta com 50 Unidades Geradoras e capacidade instalada de 3750 MW de potência. (Jirau Energia, 2019)

Trata-se de uma UHE tipo fio d'agua, isto é, possui um reservatório apenas com capacidade de regulação diária ou semanal. Para controle de nível do reservatório durante período de cheia, a usina conta com 18 vãos vertedouros e 1 vertedouro de troncos, necessário para desviar a grande quantidade de troncos presente no rio durante este período.

As U.G. estão agrupadas em 13 ilhas, das quais 12 possuem 4 U.G. cada e uma contém 2 U.G. Cada ilha conta com 2 quadros de distribuição para S.A., num total de 26 quadros, responsáveis por alimentar serviços como iluminação, sistemas de proteção e controle, excitação das máquinas, operação dos vertedouros, entre diversos outros.

É importante destacar que a carga dos S.A. não está uniformemente distribuída entre os quadros de distribuição. Por este motivo, algumas ilhas têm maior consumo por prover energia para serviços gerais da hidrelétrica, como iluminação geral, enquanto outras, têm o consumo atrelado apenas às operações da própria ilha.

Ao todo, tem-se como dados monitorados: geração por U.G. (50), vazão por U.G. (50), consumo de S.A. por quadro de distribuição (26) e vazão vertida por vão (18), além das variáveis monitoradas globais: vazão do vertedouro de tronco (1), nível do reservatório (1) e queda bruta (1), totalizando 147 variáveis.

Dados históricos da UHE de Jirau foram utilizados para aplicação das técnicas apresentadas nas seções anteriores. Por se tratar de uma usina de grande porte e com grande quantidade de U.G., esta possui uma grande complexidade em sua operação e otimização.

Para simplificar a análise, os dados foram agrupados por ilha. A A.C.P. foi aplicada separadamente para cada ilha, de forma a reduzir a dimensionalidade de seus dados de entrada. Dentre os dados disponíveis para cada ilha, foram utilizados para este trabalho:

- Dados de entrada: X1 - Geração total da ilha (MWh); X2 - U.G. ativas; X3 - vazão total da ilha (m3/s); X4 - nível do reservatório (m); X5 - queda bruta (m);
- Dado de saída: Y - Consumo total da ilha (MWh).

Tais variáveis foram selecionadas a fim de predizer o consumo total da ilha na etapa de verificação por ML e, desta forma, possibilitar a otimização da operação da UHE em estudos futuros.

Os dados da Ilha 4 foram selecionados para análise detalhada devido à menor influência do consumo dos S.A. de fatores externos não mensurados pelos dados de entrada.

Todos os procedimentos, resultados e gráficos foram obtidos e gerados utilizando linguagem de programação Python, com as bibliotecas: Numpy, Matplotlib, Pandas, Seaborn, Scikit-Learn e Time.

4.1 Tratamento e Análise de Dados

Apesar do início das operações da UHE em 2012, foram utilizados apenas dados obtidos entre os anos de 2017 e 2020, período no qual a usina contava com as 50 U.G. em operação comercial (as últimas U.G. entraram em operação em dezembro de 2016) e dados amostrados a cada 1 hora.

Para garantir a compatibilidade entre as medições de variáveis instantâneas (vazão, queda bruta, nível e U.G. ativas) e variáveis incrementais (consumo e geração), foram utilizados apenas amostras de dados estáveis, cuja quantidade de U.G. ativas permaneceu constante na amostra anterior e na posterior. Sob este critério de seleção, o conjunto de dados da Ilha 4 passou a conter 29026 amostras das variáveis de entrada e saída mencionadas.

Após o tratamento inicial e remoção de dados espúrios, foram analisadas as correlações entre os dados analisados, apresentadas na Tabela 1. A forte correlação entre os dados indica um grande potencial de redução de dimensionalidade.

Tabela 1. Coeficientes de correlação das variáveis originais

	X_1	X_2	X_3	X_4	X_5	Y
X_1	1,000	0,95	0,986	0,780	0,664	0,904
X_2	0,950	1,000	0,970	0,655	0,531	0,962
X_3	0,986	0,970	1,000	0,726	0,577	0,925
X_4	0,780	0,655	0,726	1,000	0,913	0,622
X_5	0,664	0,531	0,577	0,913	1,000	0,511
Y	0,904	0,962	0,925	0,622	0,511	1,000

4.2 Aplicação da Análise de Componentes Principais

A A.C.P. foi aplicada utilizando os dados de entrada apresentados na subseção anterior, totalizando 5 variáveis originais. Por possuir grandes diferenças entre as escalas das variáveis, o método foi aplicado com dados normalizados, utilizando a matriz de correlação, presentes na Tabela 1.

Através dos autovalores e autovetores da matriz de correlação, obteve-se os pesos das combinações lineares e as variâncias para as respectivas C.P.

4.3 Verificação com Inteligência Artificial

A fim de comparação, as técnicas de ML apresentadas na seção 3.2, foram aplicadas a ambos os conjuntos de dados: Originais e reduzidos.

Para os dados originais, as variáveis de entrada foram inseridas como entrada do modelo. Para dados reduzidos, a matriz de escore foi aplicada como entrada, variando a quantidade de C.P. utilizada. Em ambos os casos, utilizou-se a variável de saída original como saída do modelo.

O conjunto de dados foi dividido entre dados de treino e dados de teste, com proporção de 75% e 25%, respectivamente.

Para análise do desempenho das técnicas, foram calculados os coeficientes de E.M.A. e R^2 dos dados teste. Para a análise do impacto no tempo de execução dos algoritmos, mediu-se o tempo gasto em cada etapa do algoritmo: Normalização dos dados, Cálculo A.C.P., Aprendizado e Predição.

5. RESULTADOS

5.1 Resultado da Análise de Componentes Principais

Ao aplicar a A.C.P. nos dados de estudo, obteve-se os autovetores e autovalores da matriz de correlação. Os pesos das combinações lineares das variáveis originais, correspondentes aos autovetores e autovalores, são apresentados na Tabela 2.

Tabela 2. Autovetores ou pesos das combinações lineares

	CP_1	CP_2	CP_3	CP_4	CP_5
X_1	0,479	-0,228	0,015	-0,592	0,606
X_2	0,451	-0,423	-0,322	0,696	0,172
X_3	0,467	-0,348	0,156	-0,256	-0,755
X_4	0,439	0,465	0,702	0,298	0,101
X_5	0,394	0,657	-0,616	-0,104	-0,151

Pela tabela, pode-se observar quais variáveis estão mais representadas em cada C.P. e se a relação é diretamente proporcional, peso positivo, ou inversamente proporcional, peso negativo.

Os valores das variâncias da nova base de C.P., correspondente aos autovalores da matriz de correlação, são apresentados na Tabela 3, juntamente com os valores de E.V.P. e E.V.P.A. Estes valores são representados graficamente na Figura 2.

Tabela 3. Autovalores ou variância das C.P.

	CP_1	CP_2	CP_3	CP_4	CP_5
Variância	4,116	0,767	0,076	0,035	0,006
E.V.P.	82,3%	15,3%	1,5%	0,7%	0,1%
E.V.P.A.	82,3%	97,7%	99,2%	99,9%	100%

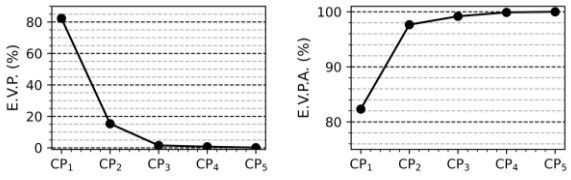

Fig. 2 E.V.P. e E.V.P.A. da Ilha 4.

Pode-se verificar que, apenas a CP_1 é responsável por 82,3 % da variância dos dados originais. Juntamente à CP_2, responsável por 15,3% da variabilidade, as duas primeiras C.P. acumulam 97,7% da variabilidade total.

Pelos critérios de seleção apresentados na subseção 2.3, e definindo um coeficiente E.V.P.A. maior que 90%, pode-se definir o uso de duas C.P. para este conjunto de dados.

Desta forma, 97,7% da variabilidade das 5 variáveis originais é preservada apenas com 2 C.P., o que representa uma redução de 60% da dimensionalidade.

Em análise das demais ilhas, verificou-se que a E.V.P.A. para 1 componente variou entre 76,7% e 88,0% a depender da ilha analisada. No entanto, essa amplitude é atenuada ao utilizar 2 ou mais C.P., variando entre 97,1% e 97,9% com o uso de duas componentes. Esta oscilação é verificada graficamente na Figura 3.

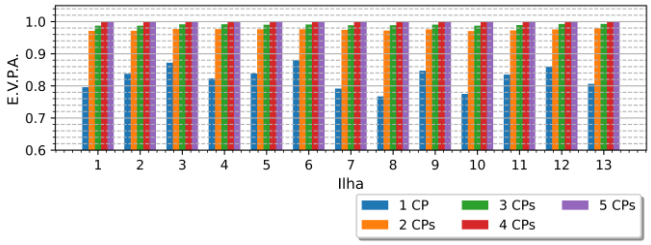

Fig. 3 E.V.P.A. das ilhas.

5.2 Resultados das Técnicas de Inteligência Artificial

A métrica de performance R^2, apresentada na Tabela 4, e sua respectiva representação gráfica na Figura 4, evidencia um ganho significativo no coeficiente de 1 para 2 C.P. utilizadas na entrada do modelo de ML. No entanto, o ganho de performance ao utilizar as demais C.P. é consideravelmente inferior para todas as técnicas aplicadas, justificado pelos E.V.P. significativamente menores das componentes 3, 4 e 5.

Tabela 4. Coeficientes de Dispersão R^2

	Árvore	M.L.P.	Linear	Vizinhança
1 C.P.	0,84	0,765	0,765	0,827
2 C.P.	0,95	0,891	0,891	0,969
3 C.P.	0,958	0,903	0,903	0,984
4 C.P.	0,958	0,926	0,926	0,984
5 C.P.	0,958	0,926	0,926	0,984
Original	0,967	0,926	0,926	0,984

Em termos proporcionais, apenas com 2 C.P. foram obtidos coeficientes de dispersão R^2 que atingem 96% a 98% dos coeficientes obtidos com os dados originais.

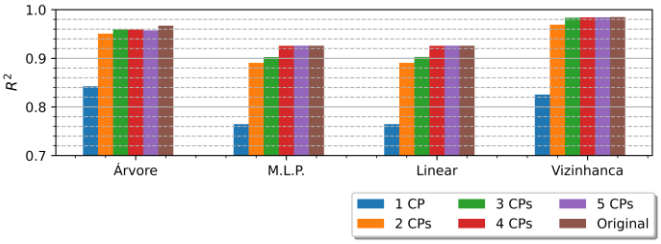

Fig. 4 Coeficientes de dispersão R^2.

Os valores do E.M.A., presentes na Tabela 5 e Figura 5, corroboram com os resultados apresentados pelo coeficiente R^2, indicando uma queda substancial no erro ao utilizar a 2ª C.P., seguida por uma queda menor para as demais C.P.

Tabela 5. Erro Médio Absoluto E.M.A.

	Árvore	M.L.P.	Linear	Vizinhança
1 C.P.	0,162	0,253	0,253	0,165
2 C.P.	0,085	0,171	0,171	0,053
3 C.P.	0,075	0,155	0,155	0,033
4 C.P.	0,075	0,125	0,125	0,032
5 C.P.	0,075	0,125	0,125	0,032
Original	0,064	0,125	0,125	0,032

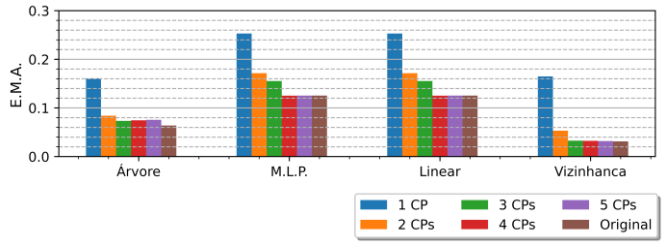

Fig. 5 Coeficientes E.M.A da Ilha 4.

Outro ponto a se observar é a diferença entre os desempenhos das técnicas aplicadas. O método de aprendizado Vizinhança se mostrou com melhor performance R^2 e E.M.A. para o cenário de previsão proposto, seguida pela técnica Árvore. Por fim, as técnicas M.L.P. e Regressão Linear obtiveram um desempenho inferior e similar entre si.

Em análise das demais ilhas, é possível identificar uma performance inferior das técnicas em ilhas cujo consumo dos S.A. incluem fatores externos não monitorados. Os coeficientes R^2 da técnica Vizinhança de todas as ilhas, representados na Figura 6, indicam uma performance expressivamente inferior nas Ilhas 1 e 2, as quais os S.A. incluem a maior parte dos serviços gerais da usina, estes não mensurados nas variáveis utilizadas. No entanto, a queda no desempenho também foi observada ao se utilizar os dados sem redução.

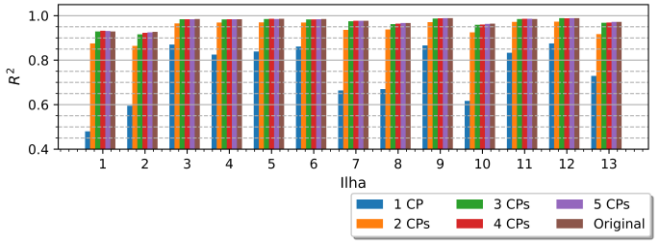

Fig. 6 R^2 (Vizinhança) das ilhas.

A Tabela 6 é referente ao Tempo de Aprendizado, de Predição e Tempo Total do algoritmo, aplicadas 2 C.P. Os valores são apresentados proporcionalmente ao tempo gasto nos algoritmos com dados sem redução de dimensionalidade.

Tabela 6. Tempo de Execução - 2 C.P.

	Árvore	M.L.P.	Linear	Vizinhança
Aprendizado	0,8225	0,3521	0,7574	0,3252
Predição	0,9849	0,9128	0,6381	0,4468
Total	0,8497	0,3579	1,1955	0,3988

A Tabela 7 apresenta a fração do tempo gasto no cálculo das C.P. em relação ao Tempo Total do algoritmo, considerando o uso de 2 C.P.

Tabela 7. Custo A.C.P. - 2 C.P.

	Árvore	M.L.P.	Linear	Vizinhança
A.C.P.	2,46%	0,46%	28,63%	2,99%

Pela Tabela 6 pode-se observar reduções nos Tempos de Aprendizado de todos os métodos avaliados. Também foram obtidas reduções significativas no Tempo de Predição nos métodos Linear e Vizinhança.

Em análise do Tempo Total, foram observadas reduções significativas nos métodos de M.L.P., Vizinhança e, em menor grau, no método Árvore. No entanto, para o método Linear, obteve-se um Tempo Total superior ao observado com dados sem redução. Este comportamento é justificado por se tratar de um método com tempo de execução expressivamente inferior aos demais. Desta forma, o tempo gasto na A.C.P., constante para todos os métodos, se torna predominante, conforme Tabela 7, sendo um custo superior aos ganhos obtidos pela redução.

6. CONCLUSÃO

Apesar das metodologias de A.C.P. serem utilizadas em outros setores, a aplicação para o pré-tratamento de dados para a previsão de consumo dos serviços auxiliares de uma UHE representa uma contribuição bem-vinda para a área uma vez que envolvem grandes quantidades de dados organizados em sistemas multivariáveis.

Diante deste cenário, o presente trabalho propôs a aplicação da técnica de redução de dimensionalidade por A.C.P. nas variáveis inerentes a geração hidrelétrica. Para verificar a viabilidade dessa técnica na etapa de pré-tratamento, algumas das técnicas de aprendizado de máquina mais comuns foram aplicadas para estimar o consumo das ilhas de geração.

Com a A.C.P., aplicando uma redução de 60% na dimensionalidade dos dados, foi possível preservar de 97,1% a 97,9% da variabilidade dos dados originais em todas as ilhas.

Em análise do desempenho das IA, as métricas R^2 e E.M.A. indicaram performance semelhante entre o conjunto de dados originais e os dados com 60% de redução. Para o coeficiente R^2, atingiu-se de 96% a 98% da performance original para a Ilha 4. Estes resultados mostraram a viabilidade de A.C.P. do ponto de vista do desempenho das predições no cenário proposto.

Por fim, com redução de 60% na dimensionalidade, obteve-se uma diminuição expressiva no tempo de execução dos algoritmos nas etapas de aprendizado e de predição. No entanto, observou-se que, para técnicas de menor tempo de

execução, o custo adicionado pelo cálculo da A.C.P. superou o ganho da redução de dimensionalidade, passando a apresentar um impacto negativo ao tempo total do algoritmo.

AGRADECIMENTOS

O autor agradece a Jirau Energia, que através do projeto regulamentado pela Agência Nacional de Energia Elétrica – ANEEL e desenvolvido no âmbito do Programa de P&D da Jirau Energia (Energia Sustentável do Brasil S.A.) (PD 06631-0011/2020), cedeu os dados para a realização do estudo de caso apresentado.

REFERÊNCIAS

Abdi, H., Williams, L. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 2(4), 433-459.

Borland, J., Hirschberg, J., Lye J. (2001). Data reduction of discrete responses: An application of cluster analysis. *Applied Economics Letters*, 8(3), 149-153.

Brunton, S.L., Kutz, J.N. (2019). *Data-driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*, Cambridge: Cambridge University Press.

Candes, E.J., Li, X., Ma, Y., Wright, J. (2011). Robust principal component analysis?. *Journal of the ACM*, 58(3).

Celik, T. (2009). Unsupervised change detection in satellite images Using principal component analysis and K-means clustering. *IEEE Geoscience and Remote Sensing Letters*, 6(4), 772-776.

Chen, L., Huang, J.Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500), 1533-1545.

Choi, R.Y., Coyner, A.S., Kalpathy-Cramer, J., Chiang, M.F., Campbell, J.P. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science & Technology*, 9(2).

Dobrev, D. (2005). A definition of artificial intelligence. *Mathematica Balkanica, New Series*, 19(1-2), 67-74.

Feng, Z., Niu, W., Cheng, C. (2019). China's large-scale hydropower system: Operation characteristics, modeling challenge and dimensionality reduction possibilities, *Renewable Energy*, 136(1), 805-818.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, Prediction Springer Series in Statistics*, Ed. 2. Springer New York, Springer Series in Statistics, New York, NY.

Hongyu, K., Sandanielo, V.L.M., Junior, G.J.O. (2016). Análise de componentes principais: Resumo teórico, aplicação e interpretação. *Engineering and Science*, 5(1), 83-90.

International Renewable Energy Agency (2021). *IRENA's Energy Transition Support to Strengthen Climate Action*, International Renewable Energy Agency, Abu Dhabi.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*, Ed. 2. Springer New York, Springer Series in Statistics, New York, NY.

Jirau Energia (2019). Energia Sustentável do Brasil, Dados Técnicos. Disponível em : https://esbr.com.br/a-usina#dados-tecnicos (Acesso em : 15 de nov. de 2021)

Johnson, R.A., Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*, Madison: Prentice Hall International.

Ritchie, H., Roser, M. (2020). Energy. *Our world in Data.*

Savegnago, R.P., Caetano, S.L., Ramos, S.B., Nascimento, G.B., Schmidt, G.S., Ledur, M.C., Munari, D.P. (2011). Estimates of genetic parameters, and cluster and principal components analyses of breeding values related to egg production traits in a white leghorn population. *Poultry Science*, 90(10), 2174-2188.

Sousa, S.I.V., Matins, F.G., Alvim-Ferraz, M.C.M., Pereira, M.C. (2007). Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling & Software*, 22(1), 97-103.

Yata, K., Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis*, 105(1), 193-215.

Yong, A.G., Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79-94.

Zhang. T., Yang, B. (2016). Big data dimension reduction using PCA. *2016 IEEE International Conference on Smart Cloud (SmartCloud)*, 152-157.