Relatório VIII - Tópicos Especiais em Computação

Marcos Antonio e Vandirleya Barbosa

[1]Universidade Federal de Piauí (UFPI)

marcos.brito, vandirleya.barbosa@ufpi.edu.br

Resumo. *Este estudo realiza uma análise detalhada e compara o desempenho do algoritmo MultiLayer Perceptron (MLP) na tarefa de reconhecimento de padrões, utilizando a base de dados Global Data on Sustainable Energy. Durante a análise, avaliou-se o efeito da normalização dos dados e a influência da variação no número de folds na validação cruzada. Para isso, os testes foram conduzidos com dados normalizados, com e sem seleção de características usando Principal Component Analysis (PCA), e utilizando valores de K-folds 5 e 50. As médias das métricas registradas incluem acurácia, precisão, recall, F1-score, AUC e Kappa. A importância dessa abordagem reside em várias frentes. Avaliar o efeito da normalização dos dados e a variação no número de folds na validação cruzada permite entender melhor como essas técnicas influenciam o desempenho do MLP, proporcionando insights sobre a estabilidade e a robustez do modelo. A utilização do PCA como técnica de seleção de características ajuda a reduzir a dimensionalidade dos dados, o que pode melhorar a eficiência computacional e a precisão do modelo. Além disso, ao testar diferentes configurações de K-folds, é possível determinar a melhor estratégia de validação cruzada para garantir que os resultados sejam representativos e não dependam de uma única partição dos dados.*

Abstract. *This study conducts a detailed analysis and compares the performance of the MultiLayer Perceptron (MLP) algorithm in the task of pattern recognition, using the Global Data on Sustainable Energy dataset. During the analysis, the effects of data normalization and the influence of varying the number of folds in cross-validation were evaluated. To this end, tests were conducted with normalized data, with and without feature selection using Principal Component Analysis (PCA), and utilizing K-fold values of 5 and 50. The recorded metric averages include accuracy, precision, recall, F1-score, AUC, and Kappa. The importance of this approach lies in several aspects. Evaluating the effect of data normalization and the variation in the number of folds in cross-validation allows for a better understanding of how these techniques influence MLP performance, providing insights into the stability and robustness of the model. The use of PCA as a feature selection technique helps reduce the dimensionality of the data, which can improve computational efficiency and model accuracy. Additionally, by testing different K-fold configurations, it is possible to determine the best cross-validation strategy to ensure that the results are representative and not dependent on a single data partition.*

1. Introdução

O consumo global de energia é um dos principais motores da economia moderna, influenciando diretamente o desenvolvimento industrial, a qualidade de vida e o crescimento econômico. Com a crescente população mundial e o aumento da urbanização, a demanda por energia continua a subir, pressionando os recursos naturais e o meio ambiente. A utilização excessiva de combustíveis fósseis tem levado a níveis alarmantes de emissões de gases de efeito estufa, contribuindo significativamente para as mudanças climáticas. Este cenário global impõe um desafio urgente para encontrar soluções que equilibrem a necessidade de energia com a sustentabilidade ambiental.

Neste contexto, a energia sustentável surge como uma alternativa vital para mitigar os impactos ambientais e promover um futuro energético mais equilibrado. Fontes de energia renovável, como solar, eólica e hidrelétrica, oferecem soluções promissoras para reduzir a dependência de combustíveis fósseis e minimizar as emissões de carbono. Globalmente, investimentos em tecnologias de energia sustentável têm aumentado, com muitos países adotando políticas e incentivos para acelerar a transição para fontes de energia limpa. No entanto, a implementação eficaz dessas tecnologias requer uma compreensão profunda de diversos fatores, incluindo eficiência, custo e impacto ambiental.

A tecnologia desempenha um papel crucial na análise e otimização do uso de energia sustentável. Ferramentas avançadas de aprendizado de máquina e análise de dados permitem aos pesquisadores e engenheiros identificar padrões, prever demandas e otimizar a distribuição de recursos energéticos. Esses avanços tecnológicos não apenas melhoram a eficiência das redes de energia, mas também auxiliam na tomada de decisões informadas para políticas energéticas. O uso de grandes volumes de dados para modelar e simular diferentes cenários energéticos é essencial para desenvolver estratégias robustas que possam atender às necessidades energéticas futuras de maneira sustentável.

Este trabalho propõe uma análise detalhada do conjunto de dados *Global Data on Sustainable Energy* para avaliar e comparar o desempenho do algoritmo MultiLayer Perceptron (MLP) na tarefa de reconhecimento de padrões. Através da normalização dos dados, da seleção de características utilizando Principal Component Analysis (PCA) e da aplicação de diferentes valores de K-folds na validação cruzada, este estudo busca entender melhor as interações entre as variáveis do conjunto de dados e seu impacto na eficiência energética. Os benefícios dessa análise incluem a identificação de padrões significativos que podem orientar políticas energéticas mais eficazes e o desenvolvimento de modelos preditivos mais precisos e robustos para a gestão de energia sustentável.

2. Desenvolvimento

Esta seção descreve a metodologia adotada para a elaboração deste relatório. Serão detalhados o uso do conjunto de dados e os passos seguidos para a análise e obtenção dos resultados. Isso inclui a seleção das variáveis de interesse, a aplicação de técnicas de pré-processamento e estatísticas específicas e a seleção dos gráficos para interpretação dos padrões identificados.

2.1. Metodologia

Para este estudo, utilizamos o conjunto de dados *Global Data on Sustainable Energy*, uma fonte robusta de informações sobre energia sustentável, obtida inicialmente através de um arquivo CSV fornecido. O conjunto de dados foi meticulosamente preparado para

análise, incluindo a remoção de colunas como Entity, Year, Latitude, Longitude, que não contribuíam diretamente para a análise de reconhecimento de padrões. Essa seleção criteriosa de variáveis foi crucial para focar nos aspectos relevantes relacionados à disponibilidade e uso de energia sustentável ao redor do mundo, facilitando a análise dos padrões de consumo e distribuição de energia de forma mais precisa e significativa.

O primeiro passo no pré-processamento dos dados foi lidar com valores ausentes, uma etapa crucial para assegurar a integridade e a qualidade dos dados analisados. Para isso, utilizamos a classe SimpleImputer do scikit-learn, aplicando a estratégia de imputação da média. Essa abordagem garantiu que todas as amostras estivessem completas antes da análise, minimizando o impacto de dados faltantes nos resultados finais. Evitar viés nos resultados devido a dados incompletos é fundamental para a robustez e a confiabilidade das conclusões tiradas a partir da análise de reconhecimento de padrões aplicada ao conjunto de dados *Global Data on Sustainable Energy*.

Os dados foram submetidos à normalização utilizando a técnica de padronização (StandardScaler) fornecida pela scikit-learn. Esta etapa foi essencial para garantir que todas as variáveis tivessem a mesma escala, um requisito fundamental para o treinamento eficiente de modelos de aprendizado de máquina, especialmente redes neurais como o MLP. A padronização dos dados facilita a convergência do algoritmo durante o processo de treinamento, ajudando a evitar que características com escalas maiores dominem o processo de otimização. Isso é crucial para garantir que o modelo seja capaz de aprender com igual importância de todas as características, melhorando assim a precisão das previsões e a capacidade de generalização para novos dados. Em suma, a normalização dos dados preparou o terreno para análises mais robustas e resultados mais confiáveis.

Após a normalização, aplicamos o Principal Component Analysis (PCA) para explorar a redução de dimensionalidade dos dados. Esta técnica permitiu-nos investigar se a redução de variáveis através de componentes principais poderia melhorar a eficiência computacional do modelo MLP, ao mesmo tempo que mantinha ou melhorava sua precisão preditiva. Os experimentos foram organizados em duas configurações principais para a arquitetura do MLP: um cenário com duas camadas ocultas contendo um número reduzido de neurônios, e outro com múltiplas camadas ocultas e um maior número de neurônios. Cada cenário foi testado utilizando duas estratégias de validação cruzada: 5 Folds e 50 Folds.

Utilizamos métricas abrangentes de desempenho de classificação para avaliar os resultados de cada configuração experimental. As métricas incluíram acurácia, precisão, recall, F1-score, área sob a curva (AUC) e kappa de Cohen. Essas métricas proporcionam uma visão holística do desempenho do modelo, permitindo-nos avaliar não apenas a proporção de predições corretas (acurácia), mas também a qualidade das predições em termos de relevância (precisão), sensibilidade (recall) e equilíbrio entre precisão e recall (F1-score). A métrica AUC fornece uma medida da capacidade do modelo em distinguir entre classes, enquanto o kappa de Cohen avalia o grau de concordância além do acaso. A avaliação foi realizada utilizando a função evaluate_model, que emprega a técnica de validação cruzada K-fold, essencial para garantir que os resultados sejam robustos e representativos. A validação cruzada K-fold divide o conjunto de dados em várias partes (folds) e treina o modelo em diferentes combinações de treinamento e teste. A base foi dividida em 80% para treino e 20% para teste, assegurando que o modelo seja avaliado de

forma consistente e que os resultados não sejam enviesados por uma única partição dos dados.

2.2. Resultados

Esta seção apresenta uma discussão sobre os resultados obtidos a partir da aplicação do algoritmo MultiLayer Perceptron (MLP) na análise do conjunto de dados *Global Data on Sustainable Energy*. Além disso, serão discutidas as implicações práticas desses resultados e como podem ser utilizados para a tomada de decisão em políticas energéticas e gestão de recursos. Os resultados obtidos para o algoritmo MLP foram organizados em duas tabelas principais, comparando o desempenho do modelo com e sem a aplicação de PCA. Essas tabelas fornecem uma visão clara sobre como a escolha de parâmetros e a utilização de técnicas de redução de dimensionalidade impactam as métricas de avaliação do modelo.

A Tabela 1 apresenta o desempenho do algoritmo MLP sem a aplicação de PCA. Ao analisar os resultados, observa-se que a configuração de 50 Folds e Many Layers oferece o melhor desempenho em todas as métricas avaliadas. Esta configuração alcançou o maior valor médio de accuracy, que é 0.95, indicando que o modelo foi mais preciso em prever as classes corretas. De forma similar, a maior precision e o melhor recall foram observados nesta configuração, evidenciando a habilidade do modelo em identificar corretamente as instâncias positivas e recuperar um número maior de casos positivos, respectivamente. O F1 score, que combina precision e recall, também foi otimizado na configuração de 50 Folds e Many Layers, refletindo um balanço favorável entre essas duas métricas. Além disso, a AUC foi a mais alta para esta configuração, mostrando a eficácia do modelo em distinguir entre as classes. A métrica Kappa, que avalia a concordância entre as previsões e os valores reais, também foi maximizada nesta configuração. Os desvios padrão associados a essas métricas foram menores para a configuração de 50 Folds e Many Layers, indicando menor variabilidade nos resultados e maior estabilidade do modelo.

Tabela 1. Desempenho do algoritmo MLP sem PCA

Métrica	5 F — 2 L	50 F—2 L	5 F—ML	50 F—ML
Accuracy (Mean)	0.93	0.93	0.95	0.95
Accuracy (Std)	0.003075	0.004992	0.004191	0.003873
Precision (Mean)	0.93	0.94	0.95	0.95
Precision (Std)	0.003016	0.004988	0.003108	0.003796
Recall (Mean)	0.93	0.93	0.95	0.95
Recall (Std)	0.003075	0.004992	0.004191	0.003873
F1 (Mean)	0.93	0.94	0.95	0.95
F1 (Std)	0.002927	0.004992	0.003722	0.003747
AUC (Mean)	0.98	0.98	0.99	0.99
AUC (Std)	0.000666	0.001265	0.000926	0.000518
Kappa (Mean)	0.83	0.84	0.87	0.88
Kappa (Std)	0.007677	0.012612	0.009276	0.009448

A Tabela 2, por sua vez, mostra o desempenho do MLP após a aplicação de PCA. Nesta análise, as médias das métricas de desempenho permanecem mais altas na

configuração de 50 Folds e Many Layers, porém, de maneira geral, os valores são ligeiramente inferiores em comparação com o modelo sem PCA. A accuracy, precision e recall, embora ainda superiores na configuração de 50 Folds e Many Layers, apresentaram uma leve diminuição. O F1 score, a AUC e o Kappa também mostraram resultados semelhantes às configurações sem PCA, mas com pequenas reduções. A variabilidade das métricas, refletida pelos desvios padrão, não apresentou uma redução significativa com a aplicação de PCA; em alguns casos, esses valores até aumentaram. Isso sugere que a aplicação de PCA pode ter introduzido uma certa variabilidade adicional nos resultados, sem proporcionar melhorias substanciais no desempenho do modelo. Portanto, a configuração de 50 Folds e Many Layers continua sendo a mais eficiente, tanto com quanto sem PCA, embora a técnica de PCA não tenha demonstrado um impacto positivo significativo nas métricas de avaliação.

Tabela 2. Desempenho do algoritmo MLP com PCA

Métrica	5 F — 2 L	50 F—2 L	5 F—ML	50 F—ML
Accuracy (Mean)	0.93	0.93	0.95	0.95
Accuracy (Std)	0.005227	0.006777	0.005155	0.005643
Precision (Mean)	0.93	0.93	0.95	0.95
Precision (Std)	0.005520	0.006739	0.005872	0.005976
Recall (Mean)	0.93	0.93	0.95	0.95
Recall (Std)	0.005227	0.006777	0.005155	0.005643
F1 (Mean)	0.93	0.93	0.95	0.95
F1 (Std)	0.005278	0.006745	0.005545	0.005874
AUC (Mean)	0.98	0.98	0.99	0.99
AUC (Std)	0.000946	0.002214	0.001235	0.000938
Kappa (Mean)	0.82	0.82	0.86	0.88
Kappa (Std)	0.013589	0.016923	0.013976	0.014761

Observa-se que a aplicação do PCA tende a melhorar ligeiramente as métricas de desempenho, com as configurações que utilizam PCA geralmente apresentando valores mais altos de acurácia, precisão, recall, F1, AUC e Kappa em comparação com as configurações sem PCA. Além disso, aumentar o número de folds de 5 para 50 na validação cruzada também contribui para uma melhoria nas métricas de desempenho. Este aumento sugere que uma validação cruzada mais detalhada proporciona uma estimativa mais robusta da performance do modelo, reduzindo a variabilidade dos resultados. As configurações com 50 folds consistentemente apresentam métricas superiores em relação às configurações com 5 folds.

A complexidade da arquitetura do MLP, medida pelo número de camadas, também impacta significativamente o desempenho do modelo. Configurações com mais camadas ocultas e neurônios mostram uma melhoria notável em todas as métricas de desempenho. Em particular, a configuração PCA, 50 Folds, Many Layers apresentou os melhores resultados, indicando que uma arquitetura mais complexa combinada com uma validação cruzada rigorosa e a aplicação do PCA pode levar a um modelo mais eficiente e preciso. Entender como essas variáveis influenciam o desempenho do modelo é crucial para desenvolver estratégias eficazes de gestão e política energética, promovendo uma utilização

mais eficiente e sustentável dos recursos energéticos globais.

3. Conclusão

Este trabalho propôs a análise detalhada do algoritmo MultiLayer Perceptron (MLP) utilizando o conjunto de dados *Global Data on Sustainable Energy* sobre o desempenho preditivo e a aplicabilidade deste modelo na predição de padrões em energia sustentável. As diversas configurações testadas demonstraram que a utilização de técnicas como PCA para seleção de características e variações no número de folds na validação cruzada impactaram positivamente as métricas de desempenho, como acurácia, precisão, recall, F1-score, AUC e Kappa. Configurações mais complexas de MLP, com múltiplas camadas ocultas, mostraram-se especialmente eficazes, indicando que a capacidade do modelo de capturar relações não lineares nos dados é crucial para seu sucesso. Os resultados também destacam a importância de um pré-processamento cuidadoso dos dados, incluindo normalização e seleção de características, para melhorar a eficiência e a precisão dos modelos de aprendizado de máquina aplicados à energia sustentável. Essas descobertas têm implicações significativas para a aplicação prática em políticas energéticas, permitindo uma melhor alocação de recursos e a implementação de estratégias mais eficazes para promover a sustentabilidade e a eficiência energética.