UNIVERSIDADE FEDERAL DO OESTE DO PARÁ – UFOPA
INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS – IEG
BACHARELADO INTERDISCIPLINAR EM CIÊNCIA E TECNOLOGIA

MAURO SÉRGIO DOS SANTOS MOURA

RESOLUÇÃO DE PROBLEMAS DE APRENDIZADO SUPERVISIONADO COM DEEP LEARNING COM VISTAS À APLICAÇÃO EM SISTEMAS DE ENERGIA

SANTARÉM
2019

MAURO SÉRGIO DOS SANTOS MOURA

RESOLUÇÃO DE PROBLEMAS DE APRENDIZADO SUPERVISIONADO COM DEEP LEARNING COM VISTAS À APLICAÇÃO EM SISTEMAS DE ENERGIA

Trabalho de Conclusão de Curso apresentado ao Bacharelado Interdisciplinar em Ciência e Tecnologia para obtenção do grau de Bacharel em Ciência e Tecnologia na Universidade Federal do Oeste do Pará, Instituto de Engenharia e Geociências.

Orientador: Anderson Alvarenga De Moura Meneses

SANTARÉM

2019

SERVIÇO PÚBLICO FEDERAL
UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
PRÓ-REITORIA DE PESQUISA, PÓS-GRADUAÇÃO E INOVAÇÃO TECNOLÓGICA - PROPPIT
DIRETORIA DE PESQUISA
PROGRAMA INSTITUCIONAL DE BOLSAS DE INICIAÇÃO CIENTÍFICA

RELATÓRIO TÉCNICO-CIENTÍFICO PIBIC/PIBITI

1. IDENTIFICAÇÃO
Bolsista: Mauro Sérgio dos Santos Moura
E-mail: maurosergiostm@gmail.com Telefone: (93) 99229-7669
Título do Plano de Trabalho: Resolução de Problemas de Aprendizado Supervisionado com Deep Learning com vistas à aplicação em Sistemas de Energia.
Título do Projeto ao qual está vinculado o plano de trabalho: Inteligência Computacional Aplicada a Recursos Energéticos: Modelagem e Análise de Padrões de Consumo e Geração de Energia Solar Fotovoltaica.
Orientador: Anderson Alvarenga de Moura Meneses
E-mail do orientador: anderson_meneses@hotmail.com Telefone: (93) 991269294
Instituto: Instituto de Engenharia e Geociências
Bolsa: (X) PIBIC/UFOPA ( ) PIBIC/FAPESPA ( ) PIBIC/CNPq ( ) PIBITI/UFOPA ( ) PIBITI/CNPq ( ) PIBIC-AF/CNPq ( ) PIBIC-AF/UFOPA ( ) PIBIC-AF/UFOPA – Indígena ( ) PIBIC-AF/UFOPA - Quilombola
Vigência de atuação do bolsista: 2018-2019

2. INTRODUÇÃO

Inteligência Artificial (IA) é uma área de estudos onde o objetivo é encontrar ou solucionar problemas que usam máquinas para realizarem tarefas cognitivas [1]. Dentro do campo da IA, existe uma área específica de aprendizado de máquina, chamada de *Machine Learnig* (ML) é uma área onde se busca encontrar modelos computacionais para resolução de problemas com grande foco na estatística. Por fim, *Deep Learning* (DL) é uma subárea do ML que tem se tornado notável, principalmente por ser capaz de obter melhor performance em diversos problemas, assim como conseguir avanços em diversas áreas [2], o *Deep* se dá pela maior quantidade de camadas profundas, que trouxe grandes avanços na solução de problemas [3]. Utilizando DL é possível realizar tarefas simples como classificação e regressão, com grande adaptação aos dados, e principalmente com a possibilidade de aplicação em grandes quantidades de dados.

Hoje em dia, é possível conseguir dados de diversos lugares, tais como dados de temperatura, pressão, corrente elétrica, do mesmo modo que é possível criar modelos de ML capazes de encontrar padrões entre esses dados e realizar predições. Dados como de Eficiência Energética possuem grande utilidade e aplicabilidade dentro de todas as áreas, principalmente na indústria, para automação de processos e otimização de uso de recursos.

Em [4] pode-se observar o uso de técnicas de ML para realizar uma estimativa da performance energética de uma construção residencial, esse caso mostra também que é possível aplicar técnicas de classificação em *datasets* relacionados à eficiência energética, conseguindo bons resultados. Além disso, no site da Universidade da Califórnia UCI *Machine Learning*, é possível encontrar diversos conjuntos de dados, em específico um de classificação chamado *Wine*, e um de regressão chamado *Energy Efficiency*.

Esses dados requerem técnicas de pré-processamento para que fiquem no mesmo intervalo, já que são grandezas distintas. Para isso, funções de redimensionamento de dados podem ser utilizadas, tais como `MinMaxScaler`, que está contida dentro da biblioteca scikit-learn da linguagem de programação Python, técnicas como podem ser utilizadas em conjunto *Principal Components Analysis* [5] aperfeiçoar a escolha dos parâmetros principais.

A grande importância da introdução à aprendizado supervisionado para tarefas de classificação e regressão em pesquisas de graduação, utilizando os *datasets Wine e Energy Efficiency*, se dá pela necessidade de conhecer o funcionamento de Redes Neurais Artificiais (RNAs) e suas aplicações, o que influencia diretamente na possibilidade de modelagem de problemas reais.

3. OBJETIVOS

O objetivo geral do projeto foi determinar quais, dentre as arquiteturas de RNAS profundas testadas, são as melhores para a classificação de dados dos *datasets Wine* e *Energy Efficiency*. Houve uma alteração devido aos *datasets* serem de tarefas diferentes, sendo o *Wine* uma classificação e o *Energy Eficiency* uma regressão, logo precisam de abordagens diferentes, assim como métricas.

Tendo como objetivos específicos a implementação e realização de testes com RNAs profundas aplicadas aos *datasets* anteriormente mencionados, que foi realizada sem alterações. Para a implementação das Métricas de Desempenho, houve adição das métricas de regressão *Mean Squared Error* e Coeficiente de Determinação (R^2). Implementação do código para geração dos gráficos para Análise Exploratória Visual dos resultados das métricas foi realizada sem alterações. Realização do teste de Kruskal-Wallis e do teste de Dunn para verificar se há diferença estatisticamente significativa entre as arquiteturas dos modelos de redes testados, foi unicamente executada para o *Energy Efficiency* já que os resultados do *Wine* não necessitaram de análise profunda.

4. METODOLOGIA

Preparação dos dados

Os *datasets Wine* e *Energy Efficiency* foram obtidos no site UCI *Machine Learning*. O dataset Wine consiste em valores obtidos de análises químicas de tipos de vinho para realizar uma classificação, para este, foi utilizado uma técnica de pré-processamento, implementada a partir da biblioteca scikit-learn chamada de *MinMaxScaler* para que os dados fiquem no mesmo intervalo, ou seja, entre 0 e 1. O *dataset Energy Efficiency* consiste em dados estruturais de casas modeladas no software AutoCAD Ecotec, com o objetivo de determinar os valores de *Cooling Load* (CL) e *Heating Load* (HL) a partir da regressão, nesse caso foram realizadas 5 técnicas de pré-processamento, sendo elas, dados sem tratamento (caso 1), dados com `MinMaxScaler` (caso 2), dados com PCA com n = 2 (caso 3), PCA com n = 4 (caso 4), PCA com n = 6 (caso 5). A divisão do *dataset* foi de 90% para treinamento e 10% para teste.

Os parâmetros do *dataset Wine*, assim como seu *range* são apresentados na Tabela 1.

Tabela 1 – Parâmetros e *ranges* do *dataset Wine*

Parâmetros	**Range**
Classe	1 até 3
Álcool	11.03 até 14.83

Ácido Málico	0.74 até 5.8
Cinza	1.36 até 3.23
Alcalinidade das cinzas	10.6 até 30
Magnésio	70 até 162
Fenol total	0.98 até 3.88
Flavonoides	0.34 até 5.08
Não-flavonoide Fenol	0.13 até 0.66
Proanthocyanins	0.41 até 3.58
Intensidade da cor	1.28 até 13
Hue	0.48 até 1.71
OD280/OD315 de vinhos diluídos	1.27 até 4
Prolina	278 até 1680

Os parâmetros do *dataset Energy Efficiency* são apresentados na Tabela 2, assim como os nomes das variáveis de entrada e saída e seus respectivos números de valores possíveis.

Tabela 2 – Parâmetros do *dataset Energy Efficiency*

Parâmetros	**Variáveis de Entrada ou Saída**	**Número de valores possíveis**
X1	Compactação relativa	12
X2	Área de superfície	12
X3	Área de parede	7
X4	Área do telhado	4
X5	Altura Geral	2
X6	Orientação	4
X7	Área Envidraçada	4
X8	Distribuição da Área Envidraçada	6
Y1	Carga de Aquecimento	586
Y2	Carga de Resfriamento	636

Implementação da RNA

A RNA foi implementada a partir da biblioteca Keras utilizando para isso a linguagem de programação Python 3.5. Desse modo criando duas RNAs com o modelo *MultiLayer Perceptron* (MLP) diferentes, uma para cada caso. Em todos os casos, foram criadas 3 camadas ocultas utilizando a função ReLU e o algoritmo de otimização Adam. Os parâmetros de rede foram definidos empiricamente e foi selecionado para cada modelo separadamente.

Para o *Wine* uma RNA MLP de classificação foi desenvolvida, cada camada oculta contém 7 neurônios, utilizando como função de saída a função *Softmax*, *batch size* de 128 e 500 épocas.

No caso do *Energy Efficiency*, uma RNA MLP de regressão foi desenvolvida, a camada de saída possui a função Linear, as camadas ocultas contêm 128 neurônios cada, o *batch size* é de 30 e foram executadas 450 épocas.

Para garantir a reprodutibilidade dos dados a técnica de validação cruzada [6] foi executada, fazendo a rede neural ser testada 10 vezes a utilizando com dados do *dataset*.

Métricas de Desempenho

As métricas foram implementadas a partir da biblioteca scikit-learn, sendo utilizadas para avaliar o desempenho da RNA. Para a avaliação da classificação (*Wine*), foram utilizadas as métricas Acuraria, Precisão, Recall e F1-Score e para a regressão (*Energy Efficiency*) foram utilizadas MSE e R^2.

Análise Exploratória Visual dos Resultados

A análise exploratória visual dos dados foi feita a partir de boxplots e tabelas das métricas, nesse caso foram implementadas a partir das bibliotecas pandas, matplotlib e seaborn.

Análise Estatística

O teste de Kruskall-Wallis foi realizado para verificar se há diferença significativa entre os resultados obtidos com um valor de p = 0.05. O teste de Dunn foi executado logo após, caso haja diferença estatisticamente significativa, para a comparação pareada dos resultados. O teste foi realizado utilizando o software SAS/STAT® 15.1 *University Edition*.

5. RESULTADOS OBTIDOS

Dois *datasets* foram resolvidos utilizando técnicas de DL e cada um deles será abordado separadamente a seguir.

Wine

No caso do *dataset Wine*, os resultados foram encontrados a partir da execução da RNA 10 vezes. As métricas, média, desvio, dentre outras informações obtidas são apresentadas na Tabela 3.

Tabela 3 – Resultados das métricas do *dataset Wine*

	Acurácia	Precisão	Recall Score	F1 Score
mean	0,983	0,983	0,984	0,982
std	0,027	0,029	0,027	0,030
min	0,941	0,917	0,933	0,930
25%	0,958	0,969	0,969	0,961
50%	1,000	1,000	1,000	1,000

Como pode ser observado na Tabela 3, o *dataset* é de fácil resolução para a RNA desenvolvida, assim como percebe-se que a partir do segundo quartil os resultados foram 1, indicando 100% de acertos, não necessitando comparação com outros modelos. Vale ressaltar que em teoria atingir 100% nas métricas é impossível, porém, devido à pouca complexidade do *dataset* o resultado é possível. Esse caso mostra o poder de resolução de uma RNA MLP, mesmo sendo uma classificação de multiclasses.

Energy Efficiency

Para o *dataset Energy Efficiency*, foram feitas 4 análises, sendo duas para a saída de HL e duas para CL. Essas análises são apresentadas na forma de boxplots e tabelas.

Resultados da métrica MSE para CL.

O boxplot dos resultados da métrica MSE para CL são apresentados na Figura 1.

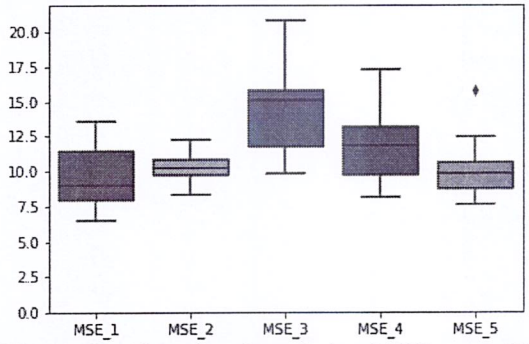


Figura 1 – Boxplot da métrica MSE para CL.

A estatística descritiva dos dados na Figura 1 são apresentados na Tabela 4.

Tabela 4 – Resultado da métrica MSE para CL

	MSE_1	MSE_2	MSE_3	MSE_4	MSE_5
mean	9.562	10.188	14.464	11.878	10.236
std	2.427	1.116	3.453	2.910	2.403
min	6.447	8.337	9.816	8.061	7.539
max	13.552	12.237	20.799	17.290	15.816

Resultados da métrica R^2 para CL
Os boxplots dos resultados de R^2 para CL são apresentados na Figura 2.

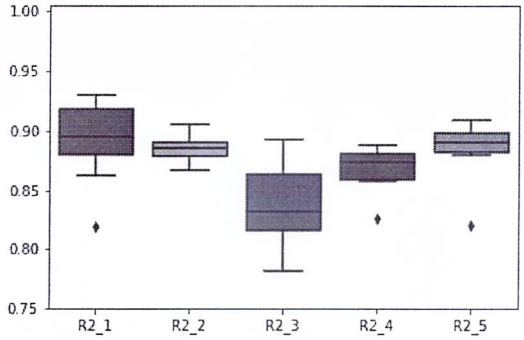


Figura 2 – Boxplot da métrica R^2 para CL.

A estatística descritiva dos dados na Figura 2 são mostrados na Tabela 5.

Tabela 5 – Resultados de R^2 para CL.

	R2_1	R2_2	R2_3	R2_4	R2_5
mean	0.892	0.885	0.838	0.868	0.886
std	0.034	0.011	0.034	0.018	0.025
min	0.819	0.867	0.782	0.827	0.820
max	0.931	0.906	0.892	0.888	0.909

Resultados da métrica MSE para HL
O boxplot dos resultados do MSE para HL é apresentado na Figura 3.

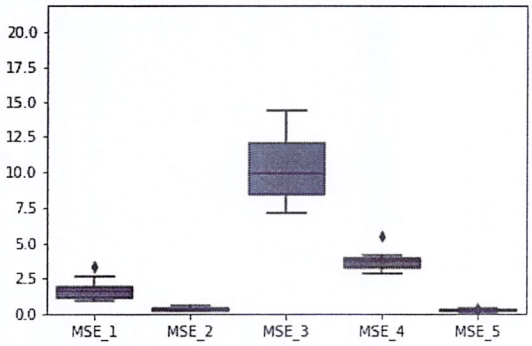


Figura 3 – Boxplot da métrica MSE para HL.

A estatística descritiva dos dados mostrados na Figura 3 são mostrados na Tabela 6.

Tabela 6 – Resultados de MSE para HL

	MSE_1	MSE_2	MSE_3	MSE_4	MSE_5
mean	1.724	0.336	10.291	3.756	0.246
std	0.777	0.118	2.411	0.756	0.097
min	0.921	0.189	7.123	2.856	0.144
max	3.344	0.577	14.374	5.490	0.452

Resultados da métrica R^2 para HL

O boxplot dos resultados de R^2 para HL é mostrado na Figura 4.

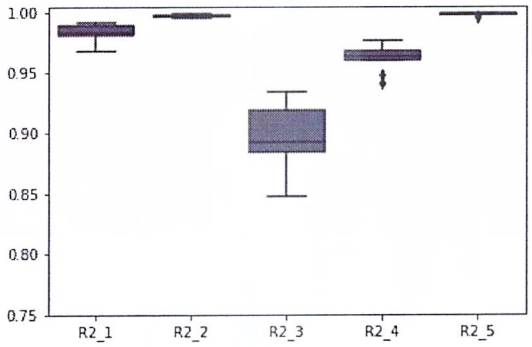


Figura 4 – Boxplot da métrica R^2 para HL.

Os resultados do boxplot de R^2 para HL são mostrados na Tabela 7.

Tabela 7 – Resultados de HL para R^2

	R2_1	R2_2	R2_3	R2_4	R2_5
mean	0.983	0.997	0.897	0.962	0.998
std	0.008	0.001	0.026	0.010	0.001
min	0.968	0.995	0.847	0.941	0.995
max	0.991	0.998	0.934	0.976	0.999

Testes Estatísticos

A partir dos resultados obtidos, foi realizado o teste de Kruskall-Wallis para verificar se há diferença estatisticamente significativa entre os resultados, assim, os resultados obtidos são apresentados na Tabela 8.

Tabela 8 – Resultados do teste de Kruskall-Wallis

Kruskal-Wallis	Heat - MSE	Heat -R^2	Cool - MSE	Cool- R^2
χ^2	45.0682	45.1360	13.5200	19.0071
p-value	<.0001	<.0001	0.0090	0.0008

Os valores da variável *p-values* estarem abaixo de 0.05 significa que realmente existe diferença significativa entre os casos. Desse modo, o teste de Dunn foi utilizado para realizar a comparação pareada entre cada caso. Os testes são apresentados a seguir contendo somente os casos que apresentaram diferença significativa.

Cooling Load

A Tabela 9 mostra os resultados do teste de Dunn para CL com MSE.

Tabela 9 – Resultados do teste de Dunn para CL usando a métrica MSE

Comparison number	Group comparisons	Difference in average ranks	Cutoff at alpha = 0.05	Significance difference = **
2	Cool_1-Cool_3	20.7	18.2996	**
9	Cool_3-Cool_5	18.4	18.2996	**

A Tabela 10 mostra os resultados do teste de Dunn para CL com R^2.

Tabela 10 – Resultados do teste de Dunn para CL utilizando R^2

Comparison number	Group comparisons	Difference in average ranks	Cutoff at alpha = 0.05	Significance difference = **
2	Cool_1-Cool_3	22.7	18.2996	**
5	Cool_2-Cool_3	19.3	18.2996	**
9	Cool_3-Cool_5	22.4	18.2996	**

Heating Load

A Tabela 11 apresenta o teste de Dunn para HL com a métrica MSE.

Tabela 11 – Resultados do teste de Dunn para HL usando a métrica MSE

Comparison number	Group comparisons	Difference in average ranks	Cutoff at alpha = 0.05	Significance difference = **
2	Heat_1-Heat_3	19.7	18.2996	**
5	Heat_2-Heat_3	32.4	18.2996	**
9	Heat_3-Heat_5	37.6	18.2996	**
10	Heat_4-Heat_5	27.3	18.2996	**

A Tabela 12 mostra os resultados do teste de Dunn para HL com a métrica R^2.

Tabela 12 – Resultados do teste de Dunn para HL com a métrica R^2

Comparison number	Group comparisons	Difference in average ranks	Cutoff at alpha = 0.05	Significance difference = **
2	Heat_1-Heat_3	19,6	18,2996	**
4	Heat_1-Heat_5	18,3	18,2996	**
5	Heat_2-Heat_3	32,1	18,2996	**
6	Heat_2-Heat_4	21,7	18,2996	**
9	Heat_3-Heat_5	37,9	18,2996	**
10	Heat_4-Heat_5	27,5	18,2996	**

Discussão

A partir das tabelas 4 a 7, é possível observar que o caso com os piores resultados é o caso 3, o mesmo pode ser utilizado como referência para analisar os outros casos, de outro lado, é notável que HL obteve resultados melhores que CL, isso pode ser devido à relação entre as variáveis ser mais forte para HL que para CL, porém é necessária uma análise separada de cada saída para obtenção de melhores resultados.

Da Tabela 9, podemos ver que os casos 1 e 5 apresentam diferença significativa com relação ao caso 3, o que indica bons resultados para MSE no caso de CL. A Tabela 10 mostra que os casos 1, 2 e 5 tem diferenças significativas para o caso 3, o que significa bons resultados. Na Tabela 11, é apresentada a diferença significativa do caso 5 para os casos 3 e 4, porém os casos 1 e 2 não apresentam diferença ao caso 4, o que mantém o 5 como o caso com melhores valores de métrica. Os resultados da Tabela 12 mostram que o caso 5 tem diferença significativa dos casos 1, 3 e 4 o que

reforça o que as tabelas anteriores apresentam, sendo o caso 5 com melhores resultados.

Por fim, considerando todos os testes realizados, o pré-processamento que apresentou os melhores resultados foi o caso 5 (PCA com n = 6), já que o mesmo mantém diferença significativa com o pior dos casos em todos os testes, enquanto os outros casos não mantiveram.

6. PUBLICAÇÕES TÉCNICO-CIENTÍFICAS.

O trabalho realizado no *dataset Energy Efficiency* deu pauta para o desenvolvimento de um artigo completo aprovado no XXII Encontro Nacional de Modelagem Computacional, sendo de autoria do bolsista Mauro Sérgio dos Santos Moura e do orientador Anderson Alvarenga de Moura Meneses. Vale ressaltar que o evento ocorrerá de 08 a 11 de outubro de 2019, logo o artigo ainda não foi publicado. Por isso, o Anexo A possui somente a primeira página com o resumo, pois o trabalho se encontra em fase de revisão pelo comitê técnico do evento.

7. PRINCIPAIS PROBLEMAS E DIFICULDADES PARA A REALIZAÇÃO DAS ATIVIDADES

O maior problema encontrado durante a aplicação do plano trabalho foi a realização da regressão a partir do *Deep Learning* com duas saídas simultâneas no caso do *dataset Energy Efficiency*, já que as implementações conhecidas da validação cruzada são normalmente para uma única saída, sendo necessário implementação manual.

8. REFERÊNCIAS

[1] S. Haykin, *Neural Networks - A Comprehensive Foundation*. Peason Prentice Hall, 2005.

[2] F. Chollet, *Deep Learning with Phyton*. Maning Publications Co., 2018.

[3] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

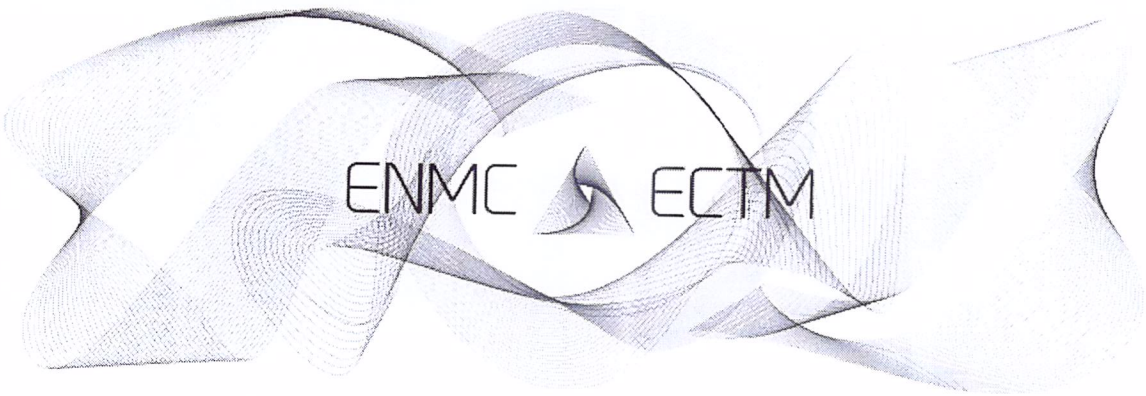
[4] A. Tsanas, and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools", Energy and Buildings, vol 49, pp. 560-567, 2012.

[5] I. T. Jolliffe, *Principal Component Analysis*, Second Ed. New York: Springer, 2002.

[6] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence," vol. 5, p. 7, 1995.

9. ANEXOS

ANEXO A – ARTIGO ENMC (SOMENTE PRIMEIRA PÁGINA)



08 a 11 de Outubro de 2019
Juiz de Fora - MG

REGRESSION OF HEATING AND COOLING LOAD DATA IN RESIDENTIAL BUILDINGS USING DEEP NEURAL NETWORKS

Mauro Sérgio dos Santos Moura[1] – maurosergiostm@gmail.com
Anderson Alvarenga de Moura Meneses[2] – anderson.meneses@ufopa.edu.br
[1] Universidade Federal do Oeste do Pará, Laboratório de Inteligência Computacional, Instituto de Engenharia e Geociências – Santarém, PA, Brasil
[2] Programa de Pós Graduação em Sociedade Natureza e Desenvolvimento, Universidade Federal do Oeste do Pará – Santarém, PA, Brasil

Abstract. *Energy Efficiency addresses the efficient consumption of energy, which is extremely important nowadays. Deep Learning allows tasks such as classification and regression with massive amounts of data with an enormous potential of application to Energy Efficiency. The objective of present work is to use a Deep Neural Network (DNN) for regression of Energy Efficiency data, in particular for the output variables Heating Load (HL) and Cooling Load (CL) in residential buildings. The values obtained in the regression can be used to estimate HL and CL given structural data of residential buildings. The use of deep learning for regression tasks presents itself as a viable solution in comparison to linear regression, with highly accurate results. A DNN was developed and the dataset was tested with different types of preprocessing and the results were evaluated using the metrics Mean Squared Error (MSE) and Coefficient of Determination (R^2). The mean MSE results range from 9.562 to 14.464 for CL and from 0.246 to 10.291 for HL. The mean R^2 results range from 0.838 to 0.892 for CL and from 0.897 to 0.998 for HL.*

Keywords: *Energy Efficiency, Deep Learning, Regression, Cooling Load, Heating Load.*

1. INTRODUCTION

Nowadays Energy Efficiency is very important for many reasons. According to IEA (2018), "the global energy demand rose by 1.9% in 2017", which means more consumption, therefore increasing the requirement of saving energy. Among Energy Efficiency topics, the determination of Heating Load (HL) and Cooling Load (CL) is particularly important. Such values can be acquired from residence construction data, with statistical regression (Tsanas and Xifara, 2012). This can be useful for designing air conditioning systems efficiently, avoiding unnecessary expenses.

10. PARECER DO ORIENTADOR

O aluno Mauro Sérgio dos Santos Moura tem demonstrado proatividade e dedicação durante a execução das tarefas da iniciação científica. Tem melhorado sua interação com os demais alunos. Conseguiu publicar um trabalho no Encontro Nacional de Modelagem Computacional 2019, o que demonstra a qualidade técnica do trabalho, que foi avaliado por três revisores da área de Modelagem Computacional. Tem conseguido excelentes resultados em suas pesquisas, indo além do que foi solicitado no plano de trabalho, com os modelos avançados de redes profundas convolucionais e u-net. Apresenta potencial para atividades científicas.

Santarém, 30 de agosto de 2019

Assinatura do orientador

Mauro Sérgio dos Santos Moura

Assinatura do bolsista