



08 a 11 de Outubro de 2018
Instituto Federal Fluminense
Búzios - RJ

ANÁLISE DE ARQUITETURAS DE *DEEP LEARNING* APLICADO A UM BENCHMARK DE CLASSIFICAÇÃO

Henrique Matheus Ferreira da Silva¹ – ferreirasilva_matheus@hotmail.com

Max Tatsuhiko Mitsuya^{1,2} – maxmitsuya@yahoo.com.br

Clayton André Maia dos Santos^{1,3} – maia.csmayer@gmail.com

Anderson Alvarenga de Moura Meneses^{1,2,3} – anderson.meneses@ufopa.edu.br

¹ Universidade Federal do Oeste do Pará, Laboratório de Inteligência Computacional – Santarém, PA, Brasil

² Universidade Federal do Oeste do Pará, Programa de Pós-Graduação em Recursos Naturais da Amazônia – Santarém, PA, Brasil

³ Universidade Federal do Oeste do Pará, Programa de Pós-Graduação em Sociedade, Natureza e Desenvolvimento – Santarém, PA, Brasil

Resumo. *O Aprendizado Profundo (Deep Learning) é uma subárea de Aprendizado de Máquina (Machine Learning), que investiga técnicas para simular o comportamento do cérebro humano. No presente trabalho apresenta-se uma análise de arquiteturas de Deep Learning na solução de um problema de Classificação. Para tanto, utilizou-se a base de dados wine quality dataset (Cortez et al., 2009). Foram analisadas 9 arquiteturas de Deep Learning, com variações no número de neurônios e camadas ocultas, de forma a comparar seus desempenhos usando as métricas: Acurácia, Precisão, Recall e F1 Score. Posteriormente foi utilizado o teste de Kruskal-Wallis para avaliar se há diferença estatisticamente significativa entre as arquiteturas com a finalidade de reduzir o custo computacional para futuras aplicações. Verificou-se que a diferença entre os desempenhos não é estatisticamente significativa (com $\alpha = 5\%$) para nenhuma das métricas utilizadas para este conjunto de dados, sendo assim possível usar a arquitetura menos complexa sem comprometer os resultados obtidos, reduzindo o custo computacional.*

Palavras-chave: *Deep Learning, Inteligência Artificial, Aprendizado Supervisionado, Redes Neurais*

1. INTRODUÇÃO

Os avanços nas tecnologias da informação tornaram possível coletar, armazenar e processar conjuntos de dados massivos e complexos. Todos esses dados contêm informações valiosas, como tendências e padrões, que podem ser usados para melhorar a tomada de decisão.

Na literatura, é possível observar o uso de diferentes técnicas de Aprendizado de Máquina (do inglês *Machine Learning*), tais como Redes Neurais Artificiais - RNAs (Sun et al., 1997), e Máquinas de Vetores de Suporte - SVM do inglês *Support Vector Machines* (Beltran & Vicencio, 2008) (Cortez et al., 2009).

O Aprendizado Profundo (do inglês, *Deep Learning*) (LeCun et al, 2015) é uma subárea de *Machine Learning* que investiga técnicas para simular o comportamento do cérebro humano em tarefas como reconhecimento de padrões, reconhecimento visual e processamento de linguagem natural usando redes neurais com a adição da ferramenta *CUDA da Nvidia*.

Neste trabalho é proposta uma análise de arquiteturas de *Deep Learning* – para predição de *benchmarks* de *machine Learning*. Para tanto foi utilizada a base de dados *Wine Quality Data Set*, disponível no *website Machine Learning Repository*, da Universidade da Califórnia em Irvine (<https://archive.ics.uci.edu/ml/datasets/wine+quality>).

A predição do *wine quality* consiste na determinação do tipo de vinho, se ele é branco (*White*) ou tinto (*Red*) a partir de propriedades físico-químicas (12 propriedades) deliberado pelo autor (Cortez et al., 2009), a base de dados é amplamente utilizada como *benchmark* para algoritmos de aprendizado supervisionado, tendo a sua importância no fato de ser um conjunto de dados advindo do mundo real.

Serão implementadas diferentes arquiteturas de *Deep Learning Multilayer Perceptron* com o intuito de saber quais são as melhores configurações da rede, a fim de verificar se existe uma diferença estatística significativa para futuras aplicações. A rede foi implementada em *Python*, usando a *Integrated Development Environment (IDE) Spyder*. O diferencial do trabalho está nas avaliações das arquiteturas, usando o cálculo das métricas *precisão (precision)*, *recall*, *f1-score* e *acurácia (accuracy)* em detrimento da simples análise da acurácia somente. Foi usado o teste estatístico de *Kruskall Wallis* para verificar se há diferença estatística entre as arquiteturas

O restante do trabalho está organizado da seguinte forma: a seção 2 apresenta o referencial teórico, destacando os modelos de *Machine Learning* aplicados para resolução do problema. Na seção 3 é apresentada a metodologia e o desenho experimental. Os resultados experimentais e análises são discutidos na seção 4. Por fim, a seção 5 discute a conclusão do trabalho.

2. REFERENCIAL TEÓRICO

Esta Seção apresenta considerações de ordem geral sobre *Machine Learning*, RNAs e arquiteturas de *Deep Learning*.

2.1 *Machine Learning*

Entidades inteligentes destacam-se pela capacidade de adequação a novos ambientes e de resolução de novos problemas. Um computador pode ser orientado a interpretar informações recebidas de uma forma que melhore gradualmente seu desempenho (Rich & Knight, 1991). Essa é a base na qual a área de pesquisa em *Machine Learning*, subárea da inteligência artificial, está fundamentada, isto é, *Machine Learning* estuda métodos computacionais capazes de obter um determinado conhecimento específico através de experiências.

Os algoritmos de *Machine Learning* possuem normalmente foco em métodos estatísticos e sua aplicação prática inclui tarefas de reconhecimento de padrões diversos, tais como pesquisas na web, filtragem de conteúdo em redes sociais, recomendações de websites de comércio eletrônico, e está cada vez mais presente em produtos de consumo como câmeras e

smartphones. Portanto, esses algoritmos podem “aprender” a definir padrões das classes envolvidas em um problema, a partir de exemplos reais obtidos do ambiente (Lecun, 2015).

Machine Learning pode ser dividida em aprendizado supervisionado e não-supervisionado (Haykin, 2008). No primeiro, o conhecimento é representado por um conjunto de exemplos de pares de entrada-saída que são transmitidos em uma sequência de instruções que o computador seguirá para alcançar o efeito desejado. Enquanto que, no Aprendizado não-supervisionado, não há exemplos rotulados da função a ser aprendida. Por este motivo, problemas de aprendizado supervisionado são denominados de Problemas de Classificação e problemas de aprendizado não-supervisionado são chamados de Problemas de Agrupamento.

Dada a necessidade de maior poder de processamento para grande volume de dados, as técnicas convencionais de aprendizado de máquina tornaram-se limitadas em sua capacidade de processar dados naturais em sua forma bruta. Cada vez mais, essas aplicações fazem uso de um conjunto de técnicas chamada *Deep Learning*, que serão tratados na Seção 2.3.

2.2 Redes Neurais Artificiais

Uma RNA (Haykin, 2008) é representada na forma de um grafo direcionado, onde os *nodos* são neurônios que representam as unidades de processamento, os arcos representam as conexões e as setas dos arcos indicam a direção normal do fluxo do sinal. Os neurônios são organizados em camadas, de modo que cada neurônio na camada i se conecta a cada neurônio na camada $i + 1$. As conexões são ponderadas e é através da modificação desses pesos que as RNAs aprendem. As RNAs possuem uma camada de entrada e uma camada de saída. Quaisquer camadas entre elas são chamadas de camadas ocultas. O *forward pass* de uma RNA é onde a informação sai da camada de entrada, através de quaisquer camadas ocultas, até a saída. O aprendizado (atualização os pesos) acontece durante o *backwards pass*. A Figura 1 apresenta um exemplo de uma rede com quatro camadas possuindo duas camadas ocultas (HLs – do inglês *Hidden Layers*). Essas redes de camadas múltiplas são chamados de *Perceptrons* Multicamadas (*Multilayer Perceptrons - MLPs*)

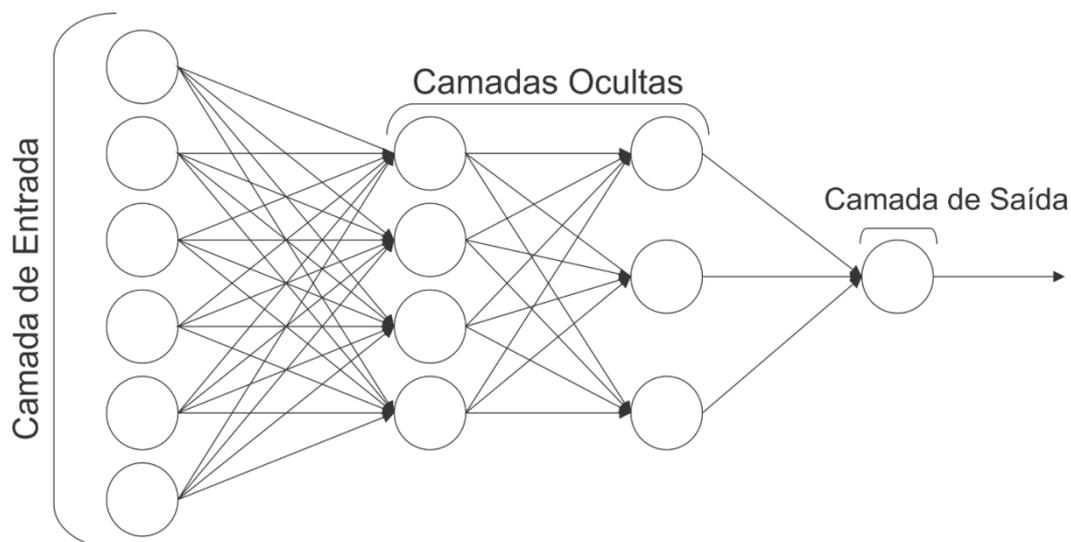


Figura 1- Estrutura de uma RNA do tipo MLP. Adaptado de Nielsen (2015).

2.3 Deep Learning

Deep Learning é uma forma de *Machine Learning* que permite que os computadores aprendam com a experiência e compreendam o mundo em termos de uma hierarquia de conceitos. Tendo sua base fundamentada no conceito de RNA ou sistemas computacionais que imitam a maneira como o cérebro humano funciona (Goodfellow et al., 2016). O *Deep Learning* tem sido responsável por avanços recentes em visão computacional, reconhecimento de fala, processamento de linguagem natural e reconhecimento de áudio.

O processo de aprendizado ocorre a partir das camadas de neurônios utilizadas para processamento dos dados. A informação é passada através de cada camada, com a saída da camada anterior fornecendo entrada para a próxima camada. A primeira camada em uma rede é chamada de camada de entrada, enquanto a última é chamada de camada de saída. Todas as camadas entre as duas são referidas como HLs. Cada camada é tipicamente um algoritmo simples e uniforme contendo um tipo de função de ativação. A Figura 2 apresenta uma breve diferença entre a arquitetura simples de rede neural e uma arquitetura de *Deep Learning*.

Atualmente, o processamento de *Big Data* e a evolução da Inteligência podem atuar em conjunto com o *Deep Learning*. Neste sentido, redes neurais baseadas no conceito de *Deep Learning* (Redes Neurais Profundas) podem receber quantidades continuamente crescentes de dados para aumentar a eficiência dos processos de treinamento. Quanto maior o volume de dados, mais eficiente é esse processo.

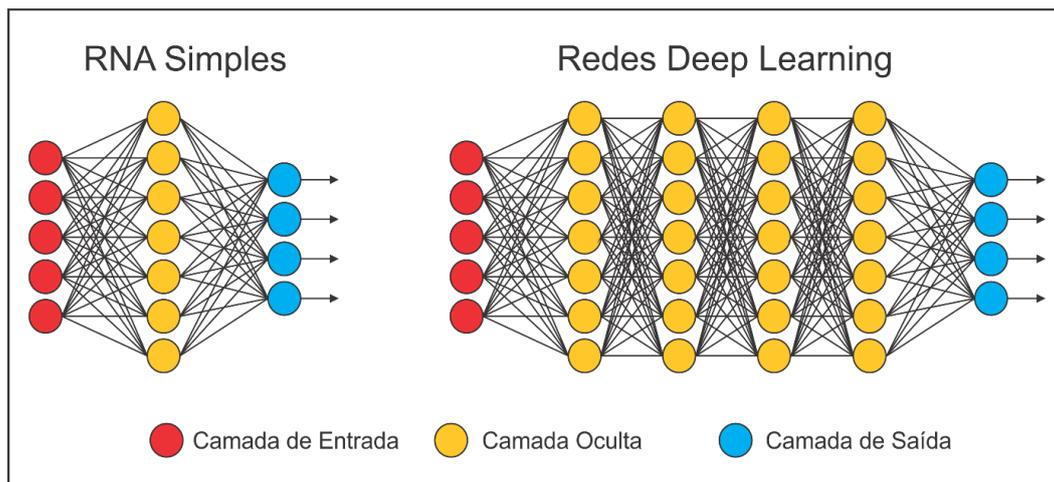


Figura 2 – Estrutura de RNA Simples e *Deep Learning*. Adaptado de Edwards (2018).

Neste trabalho, serão utilizadas redes do tipo *Deep Feed Forward* (DFF). A maioria dos resultados experimentais recentes com arquitetura profunda é obtida com modelos que podem ser transformados em RNAs supervisionadas profundas, mas com esquemas de inicialização ou treinamento diferentes das RNAs clássicas de *Feed Forward* (Glorot & Bengio, 2010).

3. PROCEDIMENTO METODOLÓGICO

Esta Seção apresentará informações sobre a classificação do *dataset* escolhido, as arquiteturas utilizadas, as métricas usadas e as análises estatísticas.

3.1 Base de dados *wine quality*

Neste trabalho foi utilizado a base de dados *wine quality* (Cortez et al., 2009). Esta base de dados é composta por dois conjuntos de dados relacionados às características físico-químicas de vinhos branco (*white*) e tinto (*red*). O conjunto de dados é formado por 1599 amostras de vinho tinto e 4898 de vinho branco, possuindo variáveis de entrada representadas pelas características físico-químico dos respectivos vinhos, e como variável de saída a a predição do vinho.

Neste trabalho, foram combinados os dois conjuntos de dados de vinhos de forma aleatória, totalizando 6497 instâncias. Neste sentido, temos uma base de dados na qual as entradas são representadas pelas variáveis físico-químicas e como saída temos a classificação do vinho em tinto (1) ou branco (0), tornando-se um problema de classificação binária.

Para implementação das redes *Deep Learning*, o conjunto de dados utilizado foi dividido em 80% para treinamento e 20% para teste, dados divididos de forma aleatória. Os resultados das métricas apresentadas no presente trabalho foram avaliadas com o conjunto de teste.

3.2 Arquiteturas das Redes *Deep Learning*

Para o desenvolvimento deste trabalho foram elaboradas diferentes arquiteturas de redes *Deep Learning* no intuito de avaliar a performance das mesmas. As arquiteturas possuem funções de ativação do tipo ReLU – do inglês *Rectified Linear Unit* nas HLs, e do tipo *Sigmoid* na camada de saída. Para todas as arquiteturas o treinamento se deu em 200 épocas.

Elas diferem quanto ao número de neurônios na HL, bem como em relação à quantidade destas camadas, conforme Tabela 1.

Tabela 1. Arquiteturas das redes implementadas.

Arquitetura	Nº Neurônios na HL 1	Nº Neurônios na HL 2	Nº Neurônios na HL 3
ARQ 1	20	-	-
ARQ 2	20	20	-
ARQ 3	20	20	20
ARQ 4	40	-	-
ARQ 5	40	40	-
ARQ 6	40	40	40
ARQ 7	60	-	-
ARQ 8	60	60	-
ARQ 9	60	60	60

3.3 Métricas de desempenho

Foram realizados 10 testes para cada arquitetura implementada. Para quantificar a performance de predição da cada arquitetura testada, foram utilizados quatro métricas, sendo elas: *Acurácia* (A_{cc}), *Precisão* (P), *Recall* e *F1 Score*. Tais métricas podem ser obtidas calculando

$$A_{cc} = \frac{VP_i + VN_i}{N_i} \quad (1)$$

$$P = \frac{VP_i}{VP_i + FP_i} \quad (2)$$

$$Recall = \frac{VP_i}{VP_i + FN_i} \quad (3)$$

$$F1\ Score = \frac{2 \times P \times Recall}{P + Recall} \quad (4)$$

onde, i é a i -ésima entrada; VP_i , VN_i , FP_i e FN_i , representam os verdadeiros positivos e negativos e falsos positivos e negativos, respectivamente; N_i é o total de amostras, dados obtidos a partir dos valores classificados pela rede, comparando com o “gabarito” dos dados.

3.4 Análise estatística

Para análise estatística, foram construídos *boxplots* para análise exploratória visual dos resultados obtidos para as métricas (*Acurácia*, *Precisão*, *Recall* e *F1 Score*). Para a construção dos *boxplots* foi utilizada a linguagem *Python*, com o auxílio das bibliotecas *pandas* e *matplotlib.pyplot*. O teste usado para verificar se há diferença estatisticamente significativa entre os resultados das métricas para cada arquitetura foi o teste não-paramétrico de *Kruskal-Wallis*, com um nível de significância de 0,05, utilizando os procedimentos *UNIVARIATE* e *NPARIWAY* do software *SAS / STAT® 9.2 University Edition*.

4. RESULTADOS E DISCUSSÃO

Usando 10 testes como base, calculou-se uma média para cada arquitetura para analisar se há possíveis diferenças estatisticamente significativas pelas distintas inicializações dos pesos da rede. Classificou-se o conjunto de dados de teste comparando com o resultado já conhecido do *dataset*, formando os respectivos *bloxplots* das métricas especificadas para cada uma das 9 arquiteturas. As Figuras 3, 4, 5 e 6 apresentam a análise exploratória visual para cada uma das métricas.

A Figura 3 apresenta uma a métrica acurácia. Pode-se observar outliers nas arquiteturas 3,4,5 e 6.

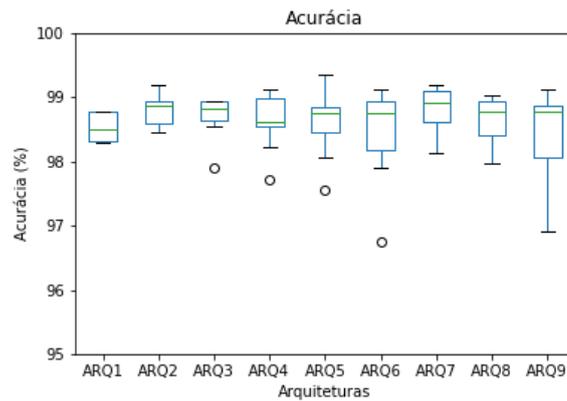


Figura 3: *Boxplot* da acurácia.

A Figura 4 analisa a precisão, que é a proporção do número de classificações corretas para o número total de instâncias positivas previstas, ou seja, se a rede classificou como vinho tinto realmente era vinho tinto e se a rede classificou como vinho branco o que realmente era branco.

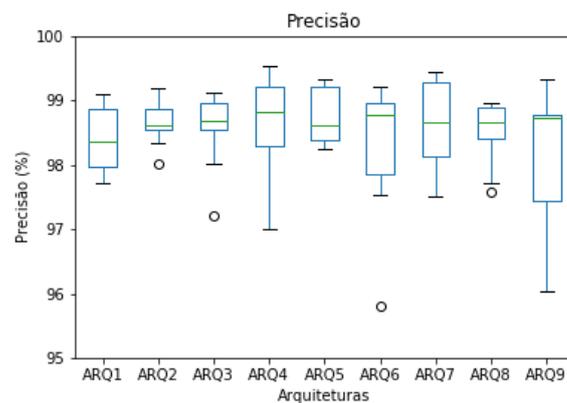


Figura 4: *Boxplot* da precisão.

Na Figura 5 é ilustrado o *Recall*, que é a proporção do número de classificações corretas para o número total de instâncias positivas reais

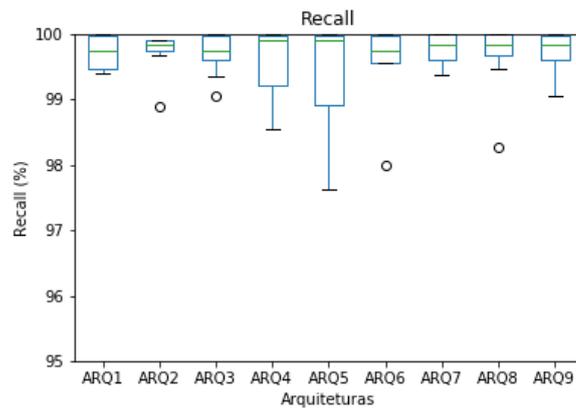


Figura 5: *Boxplot do Recall.*

A Figura 6 retrata a *F1-score* que é a média harmônica entre precisão e recall, de modo a trazer um número único indicador da qualidade geral do seu modelo e trabalha bem até com conjuntos de dados de classes desproporcionais.

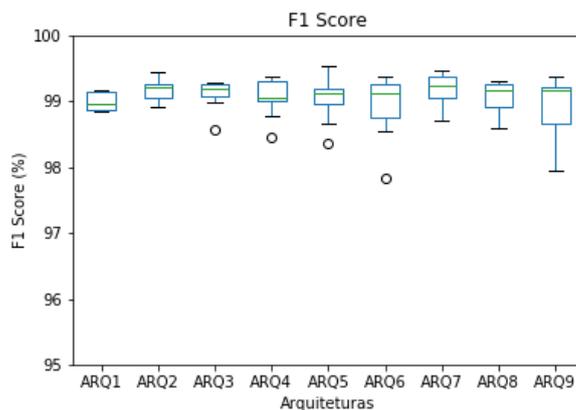


Figura 6: *Boxplot do F1 Score.*

Analisando as Fig. 3, 4, 5 e 6 nota-se que todas as arquiteturas testadas apresentaram métricas superiores a 95. No entanto, pelos resultados dos testes de *Kruskal Wallis* fora revelado que não houve diferença estatisticamente significativa entre as arquiteturas, conforme descrito na Tabela 2.

Tabela 2 - Resultados do teste de *Kruskal-Wallis*

Kruskal-Wallis	Acurácia	Precisão	Recall	F1 Score
χ^2	7,58	2,68	9,924	6,9389
p-value	0,4751	0,95	0,9983	0,5432

De acordo com os valores de *p-value*, para todas as métricas analisadas, seu valor foi maior que o nível de significância 0,05 em todos os casos. Dessa forma, para o problema de classificação binária analisado, arquitetura de rede com uma HL e 20 neurônios (ARQ 1) apresenta desempenho estatisticamente igual quando utilizado uma rede com 3 HLs e 60 neurônios (ARQ 9). Sendo assim, pudemos observar que a análise visual dos *bloxplots* por si só não é suficiente para identificar se uma arquitetura é melhor que a outra, sendo necessário utilizar um teste estatístico como o de *Kruskall Wallis*.

5. CONCLUSÕES

Pelo teste estatístico de *Kruskal Wallis* pode-se aferir que a diferença de desempenho entre as arquiteturas analisadas não é estatisticamente significativa (para um nível de significância de 5%). Neste sentido, considerando o problema de classificação binária avaliado, não há perdas significativas de desempenho ao se utilizar arquiteturas com menos neurônios e camadas, resultando num relevante ganho computacional, visto que com o aumento da complexidade, maior é o custo computacional. Os resultados apresentados neste trabalho possibilitaram investigar, com rigor estatístico, o desempenho de arquiteturas de Redes *Deep Learning* na solução de problemas de classificação binária, contribuindo para construção de conhecimento e servindo de base para estudos relacionados ao conteúdo de *Deep Learning*. Como trabalhos futuros, propõe-se aplicar a metodologia a outros *benchmarks*, e com outras arquiteturas de redes.

Agradecimentos.

H.M.F.S. agradece à Universidade Federal do Oeste do Pará (Bolsa PIBIC/UFOPA). M.T.M agradece ao apoio da CAPES e UFOPA. A.A.M.M. agradece à UFOPA (Edital PROPPIT - Apoio a Grupos de Pesquisa).

REFERÊNCIAS

- Beltran, D. M.; Vicencio, S. B. (2008), “*Chilean Wine Classification Using Volatile Organic Compounds Data Obtained With a Fast GC Analyzer*”. IEEE Transactions on Instrumentation and Measurement, vol 57, 2421-2436.
- Bishop, C. M. (2006), “*Pattern Recognition and Machine Learning*”, *Information Science and Statistics*. Springer-Verlag, Berlin.
- Cortez, P.; Cerdeira, A.; Almeida, F.; Matos, T.; Reis, J. (2009), “*Modeling wine preferences by data mining from physicochemical properties*”, Decision Support Systems, vol 47, 547-553.
- de Freitas, G.C.S.; Peixoto, F.C.; Vianna Jr, A.S. (2008), Simulation of a thermal battery using Phoenix. Journal of Power Sources, 179, 424-429. B .
- Ebeler, S., (1999), “*Flavor Chemistry: Thirty Years of Progress*”, Klumer Academic Publishers, 409-422.
- Edwards, C. (2018), “*Deep Learning hunts for signals among the noise*”. Communications of the ACM, vol 61 (6), 13-14.
- Glorot, X.; Bengio, Y. (2010), “*Understanding the difficulty of training Deep feedforward neural networks*”, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, PMLR 9:249-256.
- Goodfellow, I.; Bengio, Y.; Courville, A. (2016), “*Deep Learning*”. MIT Press.
- Gupta, Y. (2018), “*Selection of important features and predicting wine quality using machine Learning techniques*”, Procedia Computer Science, vol 125, 305-312.
- Haykin, S. (2008), “*Neural Networks and Learning Machines*” 3rd Edition, Prentice Hall.
- Lecun, Y., Bengio, Y., Hinton, G. (2015), “*Deep Learning*”. Nature, 521 (7553), 436-444.
- Anais do XXI ENMC – Encontro Nacional de Modelagem Computacional e IX ECTM – Encontro de Ciências e Tecnologia de Materiais.
Búzios, RJ – 08 a 11 Outubro 2018

- Legin, A., Rudnitskaya, A., Lvova, L., Vlasov, Y., Di Natale, C., D'Amico, A., (2003), “*Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception*”, *Analytica Chimica Acta*, 484 (1), 33-44.
- Nielsen, M. A. (2015), “*Neural Networks and Deep Learning*”. Determination Press.
- Rich, E.; Knight, K. (1991), “*Artificial Intelligence*”, McGraw-Hill.
- Sun, L.X.; Danzer, K.; Thiel, G. (1997), “*Classification of wine samples by means of artificial neural networks and discrimination analytical methods*”, *Fresenius Journal of Analytical Chemistry*, vol 359 (2), 143–149.
- Witten, I. H.; Frank, E. (2005), “*Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*”, 2º ed., Morgan Kaufmann, San Francisco.
- Gulli, Antonio;Pal, Sujit (2017) , *Deep Learning with Keras*

DEEP LEARNING ARCHITECTURE ANALYSIS APPLIED TO A BENCHMARK OF CLASSIFICATION

Abstract. *Deep Learning is a Machine Learning subarea that investigates techniques for simulating the behavior of the human brain. In the present work we present an analysis of architectures of Deep Learning in the solution of a Classification problem. For this purpose, the wine quality dataset database was used (Cortez et al., 2009). We analyzed 9 architectures of Deep Learning, with variations in the number of neurons and hidden layers, in order to compare their performances using the metrics: Accuracy, Accuracy, Recall and F1 Score. Subsequently, the Kruskal-Wallis test was used to evaluate if there is a statistically significant difference between the architectures in order to reduce the computational cost for future applications. It was verified that the difference between the performances is not statistically significant (with $\alpha = 5\%$) for none of the metrics used for this dataset, so it is possible to use the less complex architecture without compromising the results obtained, reducing the computational cost.*

Keywords: *Wine dataset, Deep Learning, artificial intelligence, metrics machine Learning and Kruskall Wallis*