Apoena Revista Eletrônica

ANÁLISE DE REGRESSÕES LINEARES E REDES NEURAIS ARTIFICIAIS: Um estudo comparativo de técnicas de inteligência artificial aplicadas na previsãode preço das ações da Nvidia

Gustavo D. Dos santos[1], Edcarlos S. Pereira[1], José Vicente C. Santos[2]Gilson Amorim Carvalho[2]

Resumo

A inteligência artificial (IA) demonstra ser a principal tendencia dos últimos anos e é progressivo nos mais diversos ramos, uma delas é a regressão, usado para estimar o valor de algo baseado em uma série de outros dados históricos. Dito isso, objetivo geral desta pesquisa é aplicar e comparar técnicas de regressão linear e redes neurais artificiais. Os objetivos específicos são: Apresentar a estrutura básica do modelo de regressão utilizando redes neurais artificiais e regressões lineares, assim como um estudo comparativo com as técnicas utilizadas. A metodologia adotada será uma revisão da literatura de forma quantitativa, uma vez que foi utilizada a coleta de dados e seus tratamentos. Este trabalho utilizou-se de uma base de dados com informações sobre o preço das ações da Nvidia. Os resultados se provaram eficientes, sendo avaliados através dos métodos de desempenho R^2, RSME e MAE, o modelo de regressão Ridge obteve o melhor desempenho em comparação aos outros modelos.

Palavras-chave: Inteligência artificial; redes neurais artificiais; Regressão linear; Nvidia;

Introdução

A adoção de sistemas de inteligência artificial (IA) é um processo progressivo nos mais diversos ramos, com o aumento dos processos de eficiência e da precisão dos serviços associados. Na saúde, vislumbra-se também esse fenômeno, onde 14 mil imagens de lesões da pele catalogados por dermatologistas foram armazenados em uma rede neural. Com o uso de um sistema de reconhecimento de padrões (Pattern Recognition), a rede neural foi capaz de reconhecer mais de três tipos de lesões. O sistema obteve uma precisão de 72%, comparado com a precisão de 66% de dermatologistas especializados (MUKHERJEE, 2017).

A inteligência artificial é o principal assunto quando se fala em tecnologias utilizadas na evolução e aperfeiçoamento dos veículos autônomos.(GONÇALVES,2011) enfatiza que “Do ponto de vista histórico, a robótica e a IA deram início ao desenvolvimento dos veículos autônomos”. Porém, a adaptação dessas tecnologias em veículos traz diversos desafios, como a migração dos carros de
de países com pouca infraestrutura.

As vantagens da adoção da tecnologia podem proporcionar uma maior velocidade e também precisão na qualidade na execução de trabalhos maçantes e repetitivos, temfeito com que cada vez mais empresas invistam em sua utilização (HUESO, 2017).

[1] Bacharelado em Ciência da Computação – Centro Universidade Jorge Amado

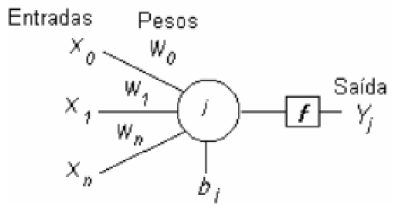
Desse modo, o objetivo desse artigo é aplicar e comparar técnicas de redes neurais artificiais e regressões lineares com o propósito de analisar o melhor resultado de previsão através dos dados estatísticos. O artigo está dividido da seguinte forma. A primeira seção é apresentada a fundamentação teórica, na qual haverá uma revisão das pesquisas e dos temas abordados. Na segunda seção teremos a apresentação dos dados e analise da base de dados utilizando as técnicas de regressão. Na quarta seção temos o resultado e análise comparativa dos modelos abordados. Por fim, na última seção teremos a conclusão e observações do projeto.

Origem e Avanços da Inteligência artificial

As redes são compostas por neurônios e estes são organizados hierarquicamente em camadas. A primeira camada é denominada camada de entrada, a qual recebe informações das variáveis independentes. A camada final é a camada de saída e as intermediárias são denominadas camadas ocultas. As conexões no percurso da rede são ponderadas por pesos, então, os neurônios da próxima camada recebem a soma dos sinais da camada anterior ponderada pelos pesos, esta informação passa então por uma função de ativação (COOPER, 2010).

Pelo fato dos modelos RNAs passarem por ajustes repetitivos, eles possuem capacidade de aprendizagem, com tamanha capacidade adaptativa, diferentes modelos foram desenvolvidos sendo modificado para novos tipos de aprendizagens e arquiteturas (BRAGA et al., 2000).

Segundo AMBRÓSIO (2002), a arquitetura da rede em geral é relacionada a função desta mesma, as quais são compostas por diversos elementos processadores, estando dispostas a uma ou várias camadas. Um elemento processador, ou neurônio, são integradores de sinais, cuja a função é integrar sinais de outros neurônios assim como sinais de entrada da rede, refletir os dados através de pesos designados e repassa-los para outros neurônios ou para a saída da rede, como vista na figura abaixo.



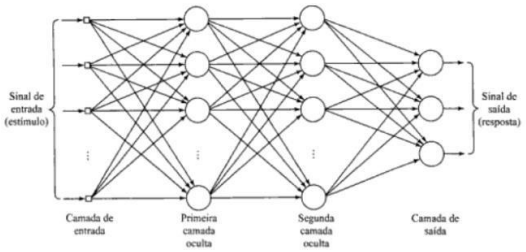
(AMBROSIO, 2002)
Figura 1 - Um neurônio artificial

Rede Neural artificial – Perceptron de múltiplas camadas

O multilayer perceptron (MLP) tem sido aplicado com sucesso na resolução de diversos problemas complexos, através de um algoritmo conhecido como propagação de erro (error back-propagation), podendo ser vista como uma generalização de um algoritmo de filtragem adaptativo conhecido como mínimo quadrado médio (LMS). Essa aprendizagem baseada em back-propagation consiste em dois passos dentre as diversas

camadas da rede, a propagação, onde o vetor de entrada é aplicado aos nós sensoriais da rede, se propagando através da mesma, e a retropropagação, onde os pesossinápticos são ajustados de acordo com uma regra de correção de erro (HYAKIN, 2001).

O multilayer perceptron é uma rede feedfoward, onde cada camada se conecta à próxima camada tendo a mesma direção, partindo da camada de entrada até a camada de saída. As camadas são classificadas da seguinte forma: Camada de entrada, onde os padrões são apresentados a rede neural, cada neurônio nessa camada deve receber uma variável independente que influencia o resultado da rede neural. Camadas escondidas, onde é feita a maior parte do processamento, as redes multilayer podem possuir uma ou mais camadas escondidas. Camadas de saída, onde o resultado será apresentado, como mostrado na figura 2 (TAVARES, 2018).



(HYAKIN, 2001)

Figura 2 - Multilayer perceptron com duas camadas ocultas

Regressão Linear

A regressão linear tem o objetivo de apresentar através de uma modelagem matemática, verifica-se que uma relação existente entre duas variáveis, a partir de n observações dessas variáveis. Dentre elas, a variável que está sendo calculada é chamada de variável dependente. A variável ou as variáveis que estão sendo usadas paracalcular a variável dependente são chamadas de variáveis independentes. Por exemplo, ao utilizarmos dados com a temperatura, umidade e pressão do ar, que seriam asvariáveis independentes Y, poderíamos prever a velocidade do vento, que seria a variável dependente X (MONTGOMERY, 2009).

$$\hat{Y} = a + bX$$

Sendo que o valor a é intersecção de Y, ou seja, o ponto onde a reta ajustada corta o eixo da variável Y, b é a inclinação e Ŷ é o valor estimado de Y para um dado valor de X, mais de uma variável independente. Para esse modelo, é chamada de regressão linear múltipla. Sendo a regressão múltipla um modelo com n variáveis independentes (MONTGOMERY, 2009).

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n + \epsilon_1$$

Onde α é intersecção de Y. β_1 é a inclinação de Y em relação à variável X_1, mantendo constantes as variáveis $X_1, X_2, \ldots, X_n$. Assim como β_1, β_2 e β_n também possuem inclinação de Y em relação à variável X_2 e X_n enquanto ε é o erro aleatório em Y, para a observação i.

Regressão Ridge

O modelo de Regressão Ridge tem como objetivo analisar e tratar de problemas com dados de regressão linear múltipla. Quando a coleta de dados é obtida usando um corte de amostras inadequados, o erro de multicolinearidade ocorre. A multicolinearidade é um problema recorrente em regressões e existe como uma relação quase linear entre as variáveis independentes, os efeitos podem criar avaliações imprecisas, com isso em mente, Hoerl \& Kennard (1970) desenvolveram o modelo de regressão Ridge para tratar dados de regressão linear múltipla que sofre multicolinearidade (KHALAF, 2022; HOERL, 2020).

Médias Moveis

Dentre as técnicas utilizadas na análise de tendencias, a que possui maior versatilidade e utilidade para os analistas são as medias moveis. Pela simplicidade de cálculo e resultados satisfatórios, ela é uma das técnicas mais antigas utilizadas pelos analistas técnicos. Sendo um estimador capaz de suavizar os dados de preços e seguir a tendência apresentada por um determinado gráfico (INFOMONEY, 2005; NORONHA, 1987).

A média móvel aritmética (MMA), apesar do seu cálculo mais simples, é utilizada até hoje agindo como uma extensão de uma media, quando aplicado, ela calcula a média dos valores de um período de tempo aplicando-os em um gráfico. Este indicador normalmente é aplicado sobre os preços de fechamento do dia (NORONHA, 1987; SENT, 2017).

$$MMA = \frac{V^1 + V^2 + V^3 + ... + V^n}{n}$$

Sendo MMA, a média móvel aritmética, V^n a quantidade de períodos que se queira calcular e 'n' sendo o número de informações aplicadas para a média.

Avaliação dos métodos de desempenhos

Porém, existem muitas aplicações de regressão que envolvem situações em que aprecipitação foram aplicados para avaliar os testes estatísticos, inclusive com o coeficiente de determinação (R^2) que passa as ser interpretado como a proporção do desvio na variável dependente que é previsível a partir da variável independente, sendo usado para verificar a relação entre os dados estimados e os dados observados. O maior valor encontrado entre 0 e 1, através da comparação dos valores observados e estimadosindica o melhor ajuste do modelo (RUEZZENE et al., 2021).

Onde: R^2 é o coeficiente de determinação (%), $\hat{Y}_i$ é o valor observado davariável dependente, $\bar{Y}$ é a média da variável dependente e Y_i é o valor estimado da variável dependente.

$$R^2 = \frac{\Sigma_{i-1}^n (\hat{Y}_i - \overline{Y})^2}{\Sigma_{i-1}^n (Y_i - \overline{Y})^2} = 1 - \frac{\Sigma_{i-1}^n \epsilon_i^2/(n-1)}{\Sigma_{i-1}^n (Y_i - \overline{Y})^2/(n-1)}$$

Erro absoluto médio (MAE) é uma medida de erro usada para representar a diferença média entre os valores estimados e observados. Diferente do R^2, o MAE deve ser mais próximo ou igual a 0 (OLIVEIRA et al., 2015).

$$MAE = \frac{\Sigma_{j=1}^{n} \mid O_j - X_j \mid}{n}$$

Sendo MAE o erro absoluto médio, X_j são os valores estimados pelo modelo, O_j são os valores observados no modelo e n é o número de observações.

Erro quadrático médio (RMSE) assim como o MAE, também é uma medida de erro, porem o RMSE obtém informações sobre o grau de dispersão obtido na comparação com os valores estimados. Assim o RMSE é sensível aos valores extremos em comparação com o MAE (OLIVEIRA et al., 2015).

$$RMSE = \sqrt{\frac{\Sigma_{j=1}^{n}(O_j - X_j)^2}{n}}$$

Sendo RMSE o erro quadrático médio, X_j são os valores estimados pelo modelo, O_j são os valores observados no modelo e n é o número de observações.

PROJETO

O estudo foi feito utilizando a linguagem Python, e para a importação dos dados, foi utilizado a biblioteca Yfinance, onde é disponibilizado informações sobre as cotações de ações, taxas de câmbios, criptomoedas e as oscilações do mercado em (TESKE, 2018) de acordo com a listagem da Associação Nacional de Corretores de Títulos de Cotações Automáticas (NASDAQ). Para a aplicação da regressão linear, regressão Ridge e rede neural MLP foram utilizadas a biblioteca Pycaret, assim como a biblioteca Matplotlib para a criação de gráficos e a biblioteca Numpy para funções matemáticas e matrizes.

Sobre a periodicidade dos dados, as informações são diárias e contemplam de janeiro de 2010 até janeiro de 2020, constituindo uma base de dados com 2518 linhas e 5 colunas, sendo elas o dia exato em que foi aberto o ativo, o valor do ativo durante a abertura, o maior valor que o ativo alcançou no dia, o menor valor que o ativo alcançou no dia, o valor final desse ativo no dia e o número de ativos que foram comprados e vendidos em determinado dia, como visto na figura 3.

(Elaborado pelos Autores)

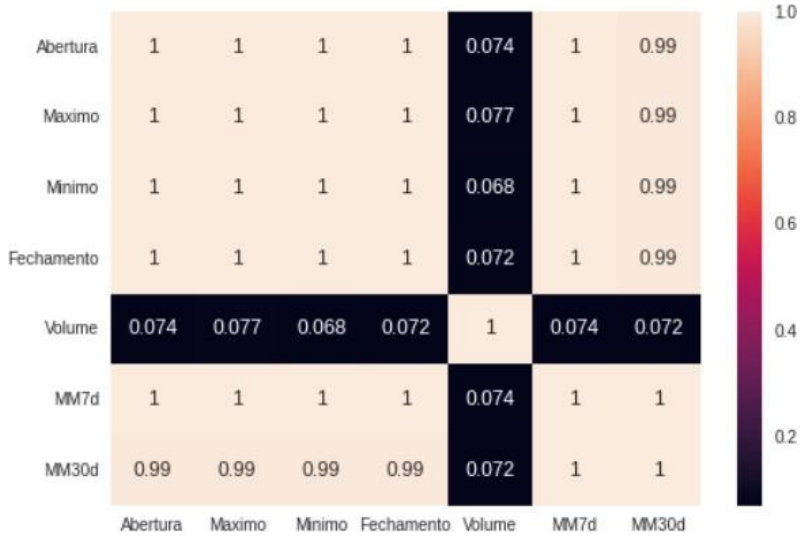
Date	Abertura	Maximo	Minimo	Fechamento	Volume
2010-01-04 00:00:00-05:00	4.247985	4.273230	4.156187	4.243395	80020400
2010-01-05 00:00:00-05:00	4.227330	4.351258	4.227330	4.305359	72864800
2010-01-06 00:00:00-05:00	4.303066	4.342080	4.261756	4.332900	64916800
2010-01-07 00:00:00-05:00	4.309949	4.328308	4.215855	4.247984	54779200
2010-01-08 00:00:00-05:00	4.213563	4.287002	4.188318	4.257167	47816800
...	...	...	...	...	...
2019-12-27 00:00:00-05:00	59.758432	59.892999	58.764118	59.028271	25464400
2019-12-30 00:00:00-05:00	58.808969	58.861299	57.580406	57.894402	25805600
2019-12-31 00:00:00-05:00	57.540540	58.731722	57.348658	58.637028	23100400
2020-01-02 00:00:00-05:00	59.496769	59.785843	58.990891	59.785843	23753600
2020-01-03 00:00:00-05:00	58.587185	59.267503	58.337984	58.828911	20538400

2518 rows × 5 columns

Figura 3 - Ações da Nvidia

Na figura é possível ver o gráfico das ações da Nvidia onde foi aplicado uma divisão dos dados, constituindo em 90% dos dados em treinamento e 10% dos dados para validação, a divisão ocorre em janeiro de 2019.

Na figura 5 foi analisada as correlações entre as variáveis dependentes, foi adicionado ao modelo medias moveis de 7 e 30 dias aplicado ao fechamento do dia. Os preços de abertura, mínimo, máximo, fechamento e medias moveis de 7 e 30 dias possuem forte correlações entre si, por outro lado o volume possui uma fraca relação com as outras variáveis independentes.



(Elaborado pelos Autores)
Figura 5 - Gráfico de correlações

RESULTADOS

Foram utilizadas 3 modelos de inteligência artificial: Regressão linear múltipla, regressão Ridge e perceptron de múltiplas camadas. Os modelos de regressão linear dupla e Ridge receberam 6 variáveis independentes, enquanto para a rede neural perceptron, o treinamento foi executado com 100 camadas ocultas e com no máximo, 500 iterações. Os modelos passaram por 10 testes para o cálculo de medias e acurácia,

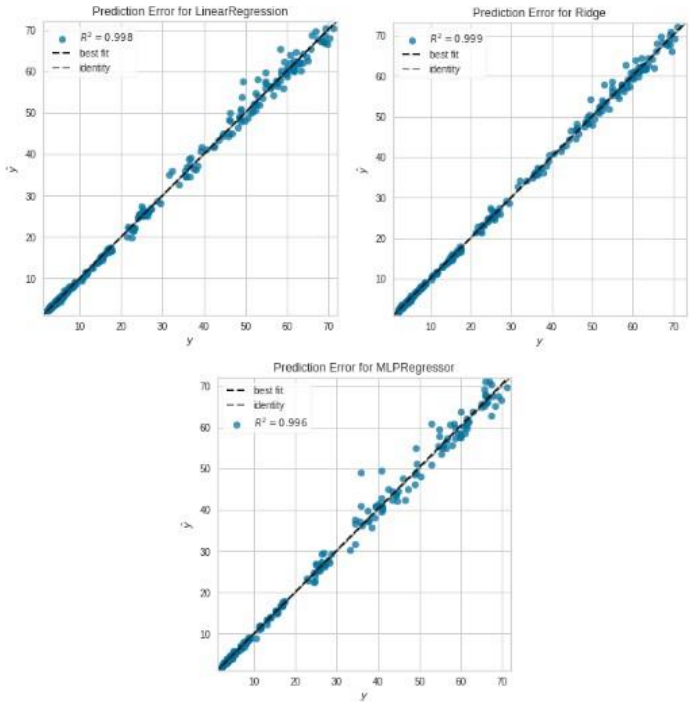
R^2											
Teste	0	1	2	3	4	5	6	7	8	9	Medias
LR	0.9978	0.9980	0.9973	0.9969	0.9959	0.9961	0.9974	0.9974	0.9980	0.9973	0.9972
Ridge	0.9986	0.9993	0.9987	0.9987	0.9972	0.9986	0.9982	0.9990	0.9990	0.9994	0.9987
MLP	0.9916	0.9926	0.9886	0.9937	0.9922	0.9937	0.9955	0.9959	0.9964	0.9932	0.9969

MAE											
Teste	0	1	2	3	4	5	6	7	8	9	Medias
LR	0.4773	0.5207	0.6258	0.6387	0.5315	0.5329	0.4203	0.6900	0.5591	0.5902	0.5587
Ridge	0.3687	0.2121	0.2519	0.3082	0.3845	0.3330	0.3448	0.2956	0.2936	0.2354	0.3028
MLP	0.7649	0.3993	0.6719	0.4802	0.5319	0.3931	0.3422	0.5117	0.4806	0.4237	0.5000

RMSE											
Teste	0	1	2	3	4	5	6	7	8	9	Medias
LR	0.9333	0.7662	0.9600	1.0168	1.3051	1.1957	0.9805	1.0387	0.8877	0.9968	1.0081
Ridge	0.7385	0.4485	0.6527	0.6692	1.0766	0.7228	0.8230	0.6368	0.6344	0.4609	0.6863
MLP	1.0044	0.8679	1.3863	1.3592	1.2869	0.7692	0.6926	1.1047	1.0457	0.9121	1.0429

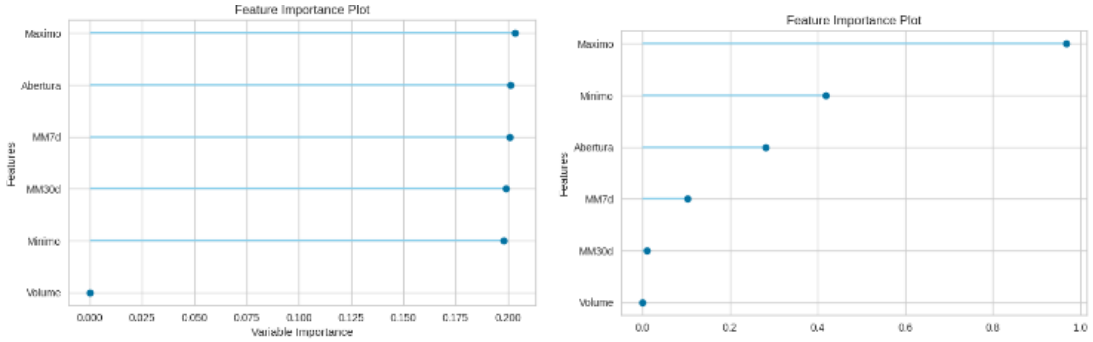
(Elaborado pelos Autores)
Figura 6 - Avaliação de desempenho ao longo de 10 testes

Os resultados oferecidos pelo modelo nos permitem observar o ajuste do conjunto de dados conforme o coeficiente de determinação(R^2). Esse gráfico compara o modelo gerado (best fit) com um modelo ideal (identity), onde no eixo x estão os valores reais e no eixo y, os valores preditos. Como as retas estão alinhadas, temos uma evidência do bom desempenho e ajuste dos dados, sendo positivo a funcionalidade do modelo.



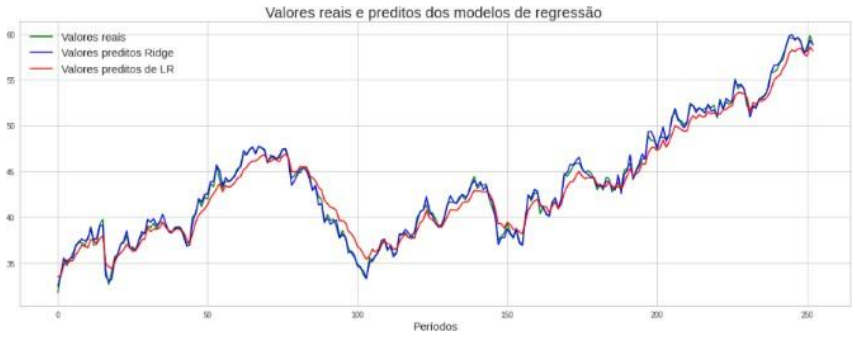
(Elaborado pelos Autores)
Figura 7 - Erro de Predição dos modelos

A feature importance é o gráfico onde podemos observar as variáveis organizadas pelo nível de importância entre os modelos. Para o modelo de regressão linear múltipla podemos ver que as variáveis máximo, mínimo, abertura e medias moveis de 7 a 30 dias possuem forte importância durante o aprendizado do modelo, enquanto a regressão Ridge, a variável máxima teve grande importância durante os testes, seguido por mínimo, abertura, medias moveis de 7 e 30 dias. Tanto a regressão linear múltipla quanto a regressão Ridge não utilizaram a variável volume, não tendo importância no modelo.



(Elaborado pelos Autores)
Figura 8 - As variáveis mais utilizadas no modelo

Na figura 9 podemos observar o gráfico de previsão dos modelos de regressão linear múltipla e Ridge. Dentro dos valores reais, o modelo Ridge é o que melhor se enquadra no modelo, mesmo seguindo a tendencia, o modelo de regressão linear múltipla possui desempenho inferior.



(Elaborado pelos Autores)
Figura 9 - Gráfico de regressão de LR e Ridge

Para a rede neural, podemos ver os resultados do modelo utilizando dados de treinamento e validação em R^2. Com os valores próximos, as suspeitas de sobre ajuste (overfitting) são baixas. Com os resíduos se concentrando no 0, o modelo apresenta bom desempenho.

(Elaborado pelos Autores)
Figura 10 - Resíduos de treinamento e validação

Podemos observar o comportamento do perceptron de múltipla camada conforme a série de tempo. Assim como a regressão linear múltipla, a rede neural teve um desempenho menor que o modelo Ridge no ajuste dos dados. Seu MAE e RMSE tiveram um grande aumento após a entrada dos dados de validação, o que diminuiu o desempenho do modelo.

(Elaborado pelos Autores)
Figura 11 - Gráfico de regressão MLP

E por fim, os resultados do R², MAE e RMSE para os dados de validação.

Modelos	RMSE	MAE	R²
	Validação	Validação	Validação
LR	0.9333	0.7428	0.9785
Ridge	0.3774	0.2891	0.9965
MLP	1.7390	1.3721	0.9252

(Elaborado pelos Autores)
Figura 12 - Avaliação de desempenho com dados de validação

CONCLUSÃO

Este estudo tem como objetivo a aplicação e comparação das técnicas de Regressão linear múltipla, regressão Ridge e perceptron de múltiplas camadas com o propósito de adquirir o melhor resultado de previsão das ações da Nvidia. Os resultados obtidos dos modelos foram aceitáveis, apesar das diversas técnicas complexas que podem ser aplicadas aos modelos de regressão.

Dentre os modelos com melhor resultado de RMSE, MAE e R^2, a regressão Ridge obteve o melhor desempenho ao longo dos testes, seguido pela regressão linear múltipla e perceptron de múltiplas camadas, que não se ajustou bem ao comportamento dos dados, como visto na figura 12.

Baseado no gráfico de feature importance do modelo de regressão linear múltipla (Figura 8), é possível notar um nível elevado de correlação entre as variáveis independentes, provocando uma super inflação dos valores estimados. Isso pode causar um menor desempenho do modelo, assim indicando multicolinearidade.Tal efeito não ocorre com o modelo de regressão Ridge, um método muito útil na presença de multicolinearidade, que apesar de não excluir nenhuma variável independente, é um meio efetivo para combate-lo.

Por fim, para estudos futuros, propõe-se analises e aplicações com outros modelos de inteligência artificial assim como aumento dos dados e novas variáveis independentes. Com a utilização da biblioteca pycaret, o pré processamento dos dados e comparações dos modelos designados torna o uso de inteligências artificiais para regressão um campo vasto de oportunidades a ser explorado.

1. REFERENCIAS

AMBRÓSIO, Paulo Eduardo. **Redes Neurais Artificiais no apoio ao diagnostico diferencial de lesões intersticiais pulmonares**. São Paulo, 2002.

BRAGA, A. P.; CARVALHO, A. P. L.; LUDEMIR, T. B. **Redes Neurais Artificiais**: Teoria e Aplicações. 2. ed. Rio de Janeiro. Ed. LTC, 2000.

COOPER, J. C. B. **Artificial neural networks versus multivariate statistics**: An application from economics. Journal of Applied Statistics v.26, pp. 909-921, 2010.

GONÇALVES, L. F. S. **Desenvolvimento de sistema de navegação autônoma por GNSS**. 2011. 192 f. Dissertação (Mestrado em Engenharia de Transportes) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2011.

HAYKIN, S. **Redes Neurais**: princípios e práticas. Tradução de Paulo Martins Engel. Porto Alegre. Ed. Bookman, 2001.

HOERL, R. (2020). **Ridge Regression**: A Historical Context. Technometrics. 62. 420-425. 10.1080/00401706.2020.1742207.

HUESO, L. C. **Big data e inteligencia artificial. Una aproximación a su tratamiento jurídico desde los derechos fundamentales**. Dilemata, ISSN-e 1989-7022, n.24, 2017, p. 131-150.

INFOMONEY COM BLOOMBERG. **Médias Móveis**: saiba como funcionam e como utilizar este indicador. InfoMoney – Em Educação/Guias Disponível em:

http://www.infomoney.com.br/educacao/guias/noticia/365152/medias-moveis-saiba-comofuncionam-como-utilizar-este-indicador. 2017.

KHALAF, G. **Improving the Ordinary Least Squares Estimator by Ridge Regression**. Open Access Library Journal, 9, 1-8. doi: 10.4236/oalib.1108738, 2022.

KHASHEI, M.; BIJARI, M. **An artificial neural network model for timeseries forecasting**. Expert Systems with Applications, v.37, p. 479-489, 2010.

MONTGOMERY, D. C., RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros**. Rio de Janeiro: LTC. 4a edição, 2009.

MUKHERJEE S. **A.I. Versus M.D**: what happens when diagnosis is automated? The New Yorker [on line] 2010 april 3. [capturado 3 mai. 2017] Disponível em: http://www.newyorker.com/magazine/2017/04/03/ai-versus-md

NORONHA, M. **Análise técnica**: teorias, ferramentas, estratégias. São Paulo: BM&F, 1987

OLIVEIRA B. P. J., CECÍLIO A. R., PRUSKI F. F., ZANATTI S. S. **Espacialização da erosividade das chuvas no Brasil a partir de séries sintéticas de precipitação**. Revista Brasileira de Ciências Agrárias, vol. 10, núm. 4, 2015, pp. 558-563, 2015.

ROSENBLATT, F. **The perceptron**: A probabilistic model for information storage and organization in the brain. Psychological Review, v.65, n.6, p. 386-408, 1958.

RUEZZENE, B. C., MIRANDA, B. R., TECH B. R. A., MAUAD F. F. **Preenchimento de falhas em dados de precipitação através de métodos tradicionais e por inteligência artificial**. Revista Brasileira De Climatologia, 29, 177–204. Recuperado de https://ojs.ufgd.edu.br/index.php/rbclima/article/view/15184, 2021.

RUSSEL, S., NORVING, P. **Inteligência artificial**, São Paulo, Campus, 2004.

SENT. D. L. E. **Análise técnica de ações**: decisão de investimento com base nas médias móveis. Universidade Tecnológica Federal do Paraná. Disponivel em:http://repositorio.utfpr.edu.br/jspui/handle/1/14197, 2017.

TAVARES J. T. S. **Sistema Automático de Negociação para a Bolsa de Valores Utilizando Redes Neurais Multilayer Perceptron e Regressão Linear**. Feira de Santana, 2018

TESKE, D. **NVIDIA Corporation**: A Strategic Audit. Honors Theses, University of Nebraska-Lincoln, 2018.Disponivel em: https://digitalcommons.unl.edu/honorstheses/18