DESENVOLVIMENTO DE UM CÓDIGO EM PYTHON PARA GERAÇÃO DE MATRIZES DE CORRELAÇÃO DE PEARSON COM LAÇOS A PARTIR DE "N" VARIÁVEIS TOMADAS DUAS A DUAS

Jonathas Vinícius Gonzaga Alves Araujo
Instituto Militar de Engenharia- IME
Praça General Tibúrcio, 80, SE/9 – Engenharia de Computação
jonathasvgaa@gmail.com

Marcos dos Santos
Instituto Militar de Engenharia- IME
Praça General Tibúrcio, 80, SE/9 – Engenharia de Computação
marcosdossantos_doutorado_uff@yahoo.com.br

Carlos Francisco Simões Gomes
Universidade Federal Fluminense – UFF
Rua Passo da Pátria, nº 156, Escola de Engenharia, São Domingos, Niterói/RJ
cfsg1@bol.com.br

RESUMO

O coeficiente de correlação de Pearson mede a correlação linear entre duas variáveis, devendo esta correlação estar compreendida no intervalo de -1 a 1, sendo -1 fortemente correlacionadas negativamente e 1 fortemente correlacionadas positivamente. A ferramenta desenvolvida tem o propósito de calcular os coeficientes de correlação de Pearson de "n" variáveis tomadas duas a duas. Com isso, todos os coeficientes podem ser dispostos em uma matriz quadrada de ordem "n". Também é possível estabelecer um valor de corte para o coeficiente, de maneira a filtrar apenas as correlações "fortes". Além disso, o código também calcula a correlação de uma variável qualquer com ela própria descasada no indexador em uma unidade, a fim de se verificar a inércia desta variável sobre ela mesma. Consequentemente, a diagonal principal da Matriz de Pearson deixa de ser populada apenas por valores 1. O programa foi desenvolvido na linguagem de programação Python, pois ela dá a possibilidade de desenvolvimento de uma plataforma web, ou até mesmo a criação de um aplicativo para smartphones. Para facilitar o entendimento da ferramenta, lançou-se mão de um exemplo ilustrativo que foi resolvido passo a passo.

Palavra-chave: Algoritmo em Python; Matriz de Correlação de Pearson; Apoio à Decisão.

ABSTRACT

Pearson's correlation coefficient measures the linear correlation between two variables, which should be within the range of -1 to 1, with -1 strongly negatively correlated and 1 strongly correlated positively. The developed tool has the purpose of calculating Pearson's correlation coefficients of "n" variables taken two by two. With this, all coefficients can be arranged in a square matrix of order "n". It is also possible to set a cutoff value for the coefficient to filter out only the "strong" correlations. In addition, the code also calculates the correlation of any variable with itself mismatched in the indexer by one unit to verify the inertia of this variable on itself. Consequently, the main diagonal of the Pearson Matrix is no longer populated by values only. The program was developed in the Python programming language, as it gives the possibility of developing a web platform, or even creating a smartphone application. To make the tool easier to understand, we used an illustrative example that was solved step by step.

Keywords: Python algorithm; Pearson Correlation Matrix; Decision Support.

Como Citar:
ARAUJO, J. V. G. A.; SANTOS, M.; REIS, GOMES, C. F. S. Desenvolvimento de um código em Python para geração de matrizes de correlação de Pearson com laços a partir de "n" variáveis tomadas duas a duas. *In*: SIMPÓSIO DE PESQUISA OPERACIONAL E LOGÍSTICA DA MARINHA, 19., 2019, Rio de Janeiro, RJ. **Anais** […]. Rio de Janeiro: Centro de Análises de Sistemas Navais, 2019.

1. INTRODUÇÃO

O Coeficiente de Correlação de Pearson também é chamado de "coeficiente de correlação produto-momento" ou simplesmente de "ρ de Pearson". Segundo Miot (2018), é um teste estatístico que explora a intensidade e o sentido do comportamento mútuo entre variáveis. Este coeficiente pode assumir apenas valores entre -1 e 1.

A correlação indica a interdependência entre duas variáveis. O cálculo do Coeficiente de Correlação de Pearson serve para detectar o grau de correlação entre as variáveis quando não se é facilmente compreendida sua interdependência.

Como mostra a Figura 1, o coeficiente 0 (zero) representa uma correlação neutra e separa a correlação negativa da positiva, quanto mais o coeficiente se aproxima de -1, mais forte é a correlação negativa, como também, quanto mais se aproxima de 1, mais forte é a correlação, mas neste caso, positiva.

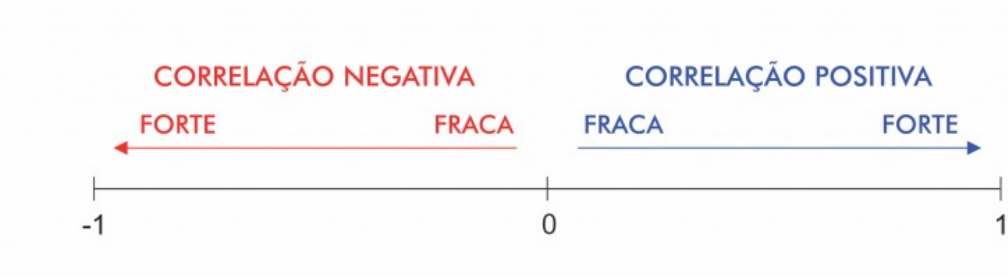

Figura 1- Intervalo de Correlação de Pearson. Fonte: Autores (2019)

Na Figura 2, por meio dos diagramas de dispersão, pode-se observar os tipos de correlação, respectivamente, a correlação negativa, a correlação nula e a correlação positiva.

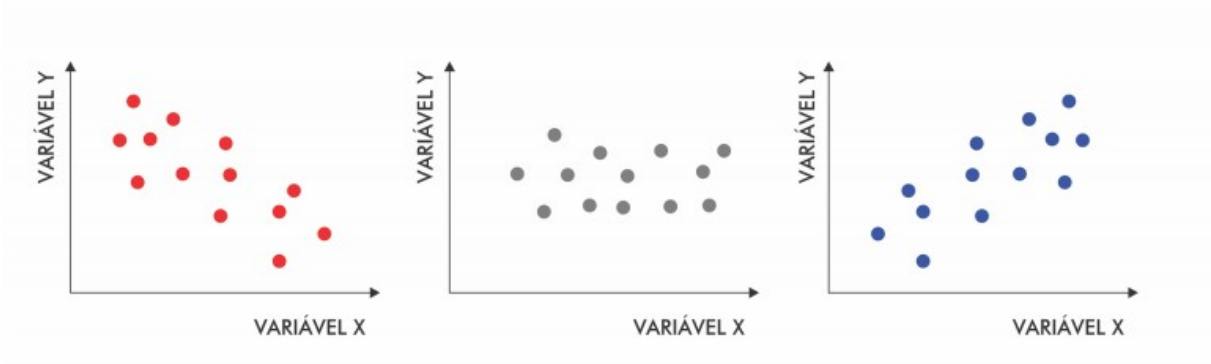

Figura 2- Diagramas de dispersão. Fonte: Autores (2019)

O coeficiente de Correlação de Pearson é calculado a partir da expressão a seguir.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}} \quad (1)$$

Diversas possibilidades de interpretação da correlação são aceitas. Pode ser considerado a inclinação da reta que representa a correlação entre as variáveis, ou as séries de valores como vetores, e o ρ, geometricamente, representaria o cosseno do ângulo formado entre os vetores.

Na Figura 3, pode-se observar uma interpretação mais comum a respeito do valor do coeficiente.

Valor de ρ (+ ou −)	Interpretação
0.00 a 0.19	Uma correlação bem fraca
0.20 a 0.39	Uma correlação fraca
0.40 a 0.69	Uma correlação moderada
0.70 a 0.89	Uma correlação forte
0.90 a 1.00	Uma correlação muito forte

Figura 3- Interpretação do coeficiente de correlação de Pearson. Fonte: Autores (2019)

A proposta apresentada é a de que o *software* faça automaticamente a correlação de todas as variáveis entre si por meio de uma matriz, atribuindo o valor a elas. Há também a possibilidade de o usuário fazer o cálculo do coeficiente com uma mesma variável e/ou especificar o valor de corte desejado para que a ferramenta refina o resultado. O valor de

corte atribuído, por exemplo, se for 0,7 serão escolhidos os valores entre 0,7 e 1 juntamente com os valores entre -0,7 e -1.

2. FUNDAMENTAÇÃO TEÓRICA

Bourgard (2017) aplicou a correlação de Pearson em uma análise em períodos de crises financeiras. Seu resultado também se apresentou provando a eficácia da ferramenta. Um dos resultados foi a descoberta que a medida que aumenta o volume de exportações e importações, assim como o aumento do consumo de energia das indústrias, o PIB também aumenta.

Já Pianuccia (2019) considerou a correlação entre as variáveis da amostra, com valores de coeficientes de Pearson, a fim de evitar problemas de multicolinearidade no uso de árvore de decisão para previsão de geração de viagens como alternativa ao método de classificação cruzada.

Um estudo bem recente realizado por Paiva, Pedrosa e Galvão (2019), empregou o teste de Pearson para verificar a correlação estatística entre a variável dependente (taxa de detecção de AIDS) e as covariáveis (indicadores socioeconômicos). O teste de Shapiro-Wilk foi utilizado para medir a normalidade da variável dependente. Para todos os testes do estudo, considerou-se alfa abaixo de 0,05 como necessário para a rejeição da hipótese nula, sendo essa a independência dos valores das taxas de AIDS em relação aos indicadores socioeconômicos da região.

Para Filho et al. (2014), em particular, uma correlação de valor zero significa que as variáveis são ortogonais entre si (ausência de correlação). Uma correlação positiva indica que quando x aumenta, y também aumenta, ou seja, valores altos de x estão associados a valores altos de y.

Para Araújo, Caran e Souza (2016), o cálculo de Pearson indica que, à medida que uma variável for alterada, a outra variável terá um comportamento linear e previsível.

Toebe *et al*. (2015) realizaram uma pesquisa em 2014 cujo objetivo era analisar o desempenho financeiro da empresa Natura por meio de da correlação entre os investimentos sociais e ambientais e a receita líquida do período de 2009 a 2013, e para isso foi utilizado o coeficiente de correlação de Pearson (r). A partir disso foi resultado que os coeficientes de correlação foi de 0,857 e 0,986 e 0,856, nos permitindo visualizar uma correlação positiva entre as variáveis, provando que a empresa tem investido em ações da sociedade e do meio ambiente. Lira e Neto (2006) apontam que a significância do coeficiente de correlação estimado é verificada por meio de um teste de hipóteses.

Petreça, Benedetti e Silva (2011) acreditam que a correlação obtida por meio de do coeficiente de Pearson, além de ser a mais conhecida, é linear. Para o coeficiente ser significativo, ele precisa ter elevada magnitude, ou seja, ser próximo de 1. Pontes (2010) afirma que o coeficiente de correlação linear de Pearson nem sempre é adequado, principalmente quando uma ou todas as variáveis são medidas em uma escala ordinal.

O coeficiente de correlação pode ser influenciado pela existência nos dados de alguns valores, sendo eles muito maiores ou menores que os outros, por isso deve-se dar uma atenção especial a este detalhe, acredita Martins (2014).

3. DESENVOLVIMENTO DA FERRAMENTA

A pesquisa foi realizada pelos pesquisadores do Instituto Militar de Engenharia (IME) e da Universidade Federal Fluminense (UFF), nos laboratórios do IME. O *software* foi desenvolvido na linguagem de programação Python, pois é suportado pela maioria das

plataformas presentes no mercado atualmente e também porque os recursos de extensibilidade do Python permitem integrar JAVA e NET.

4. APLICAÇÃO

Para exemplificar de forma lúdica, será utilizada a Tabela 1 a seguir com as duas variáveis, Massa Muscular (Y) e Idade (X) com seus respectivos valores.

Massa Muscular (Y)	Idade (X)
82.0	71.0
91.0	64.0
100.0	43.0
68.0	67.0
87.0	56.0
73.0	73.0
78.0	68.0
80.0	56.0
65.0	76.0

Tabela 1- Valores das variáveis. Fonte: Autores (2019)

A partir dos valores da Tabela 1, a próxima etapa é inseri-los no *software*, como mostra a Figura 4.

Figura 4- Entrada de dados no software. Fonte: Autores (2019)

Após a entrada de dados, dá-se prosseguimento ao programa e ele imprime os resultados obtidos em forma de uma lista e uma matriz, como ilustra a Figura 5.

```
=============== Lista dos Valores dos Coeficientes de Acordo com as Variáveis ===============

Item Massa Muscular comparado ao item Massa Muscular = 1.0
Item Massa Muscular comparado ao item Idade = -0.806

Item Idade comparado ao item Massa Muscular = -0.806
Item Idade comparado ao item Idade = 1.0


=================================================================================================

================== Matriz de Correlação de Pearson ==================

                Massa Muscular  Idade
Massa Muscular           1.000 -0.806
Idade                   -0.806  1.000

=====================================================================
```

Figura 5-Matriz de Correlação de Pearson. Fonte: Autores (2019)

Assim, são apresentados os coeficientes que satisfizeram ao valor de corte inserido no programa. No exemplo em lide, atribuiu-se o valor de corte 0,8.

Em seguida o programa pergunta se o usuário pretende fazer o cálculo do coeficiente de correlação de Pearson da variável com ela mesma. Esta opção só será solicitada se o tamanho da amostra for maior que 3. Porque para fazer este cálculo, o programa exclui o primeiro valor da amostra na primeira coluna e o último valor da amostra na segunda coluna e roda o programa novamente. A Figura 6 mostra essas etapas do cálculo.

```
============ Interações que Satisfizeram a Nota de Corte ============

 Massa Muscular com Idade = -0.806
 Idade com Massa Muscular = -0.806

=================================================================================

Deseja Efetuar o Cálculo do Coeficiente da Variável com Ela Mesma? Sim ou Não? _
```

Figura 6- Coeficientes que atenderam ao valor de corte. Fonte: Autores (2019)

Se o usuário decidir não continuar com o procedimento, o programa é finalizado. Na Figura 7 é apresentado o resultado da correlação de uma variável com ela própria, como por exemplo massa muscular com massa muscular, resultando numa correlação de -0,144.

```
=============== Lista dos Valores dos Coeficientes de Acordo com as Variáveis ===============

Item Massa Muscular comparado ao item Massa Muscular = -0.144
Item Massa Muscular comparado ao item Idade = -0.806

Item Idade comparado ao item Massa Muscular = -0.806
Item Idade comparado ao item Idade = -0.355

================== Matriz de Correlação de Pearson ==================

                Massa Muscular  Idade
Massa Muscular          -0.144 -0.806
Idade                   -0.806 -0.355
======================================================================

========== Lista com os Valores Aceitos Pela Nota de Corte ==========

 Massa Muscular com Idade = -0.806
 Idade com Massa Muscular = -0.806

======================================================================

Obrigado!

Programa Finalizado!
```

Figura 7- Matriz de Correlação de Pearson incluindo de cada variável com ela própria. Fonte: Autores (2019).

Como observado, os valores de coeficientes entre uma mesma variável foram alterados podendo ser visualizados com uma maior exatidão para onde a reta está direcionada, geometricamente falando. Logo após, o programa apresenta uma nova lista com os novos valores filtrados pelo valor de corte previamente estabelecido, e é finalizado.

A outra inovação do *software* é poder ser feito o cálculo com diversas variáveis ao mesmo tempo. No exemplo anterior foi feito somente entre as variáveis Massa Corporal e Idade, agora será incluído mais uma, a variável Altura e será executado novamente o programa. A Tabela 2 apresenta os valores.

Massa Muscular (Y)	Idade (X)	Altura (Z)
82.0	71.0	172.0
91.0	64.0	154.0
100.0	43.0	185.0
68.0	67.0	164.0
87.0	56.0	172.0
73.0	73.0	169.0
78.0	68.0	174.0
80.0	56.0	181.0
65.0	76.0	170.0

Tabela 2- Novos valores inseridos ao exemplo. Fonte: Autores (2019)

Inseridos os novos valores no programa, o resultado é apresentado na Figura 8.

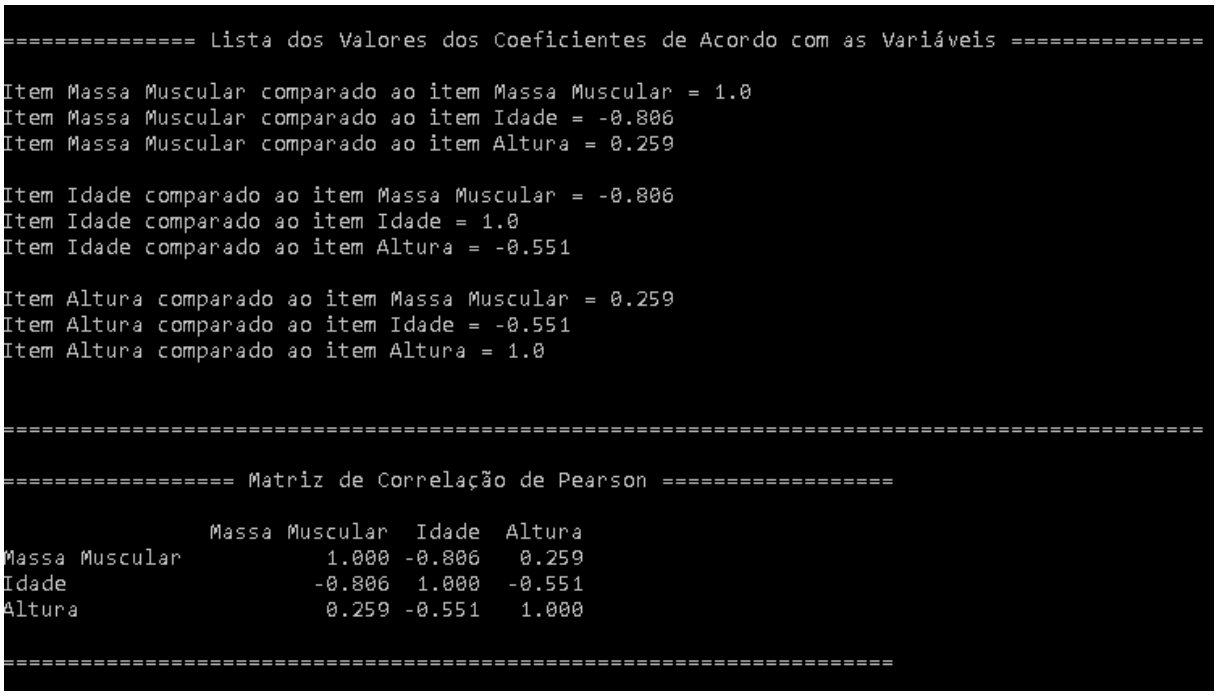

Figura 8- Coeficientes de correlação de Pearson entre as variáveis duas a duas. Fonte: Autores (2019)

Como pode ser observado, o programa calculou os coeficientes entre todas as variáveis duas a duas e apresentou uma matriz com os respectivos resultados. Logo após, o *software* informa os valores que satisfizeram ao valor de corte previamente estabelecido que foi mantido em 0,8. Os valores podem ser observados na Figura 9.

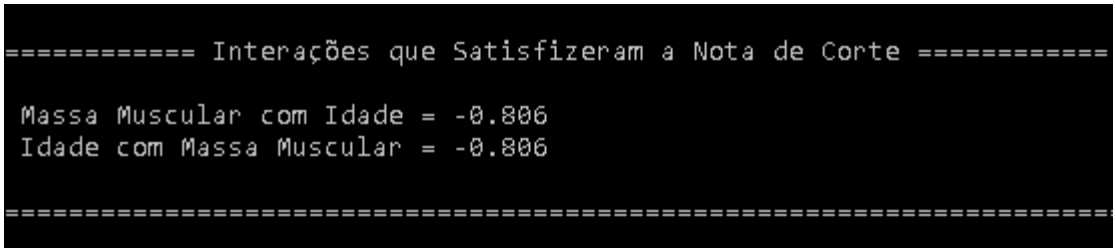

Figura 9- Coeficientes que atenderam ao valor de corte. Fonte: Autores (2019)

Continuando, o programa pergunta se o usuário deseja prosseguir com o cálculo do coeficiente das variáveis com elas próprias. Se a resposta for positiva, o programa irá calcular os novos valores, conforme apresenta a Figura 10, juntamente com uma nova Matriz de Correlação de Pearson.

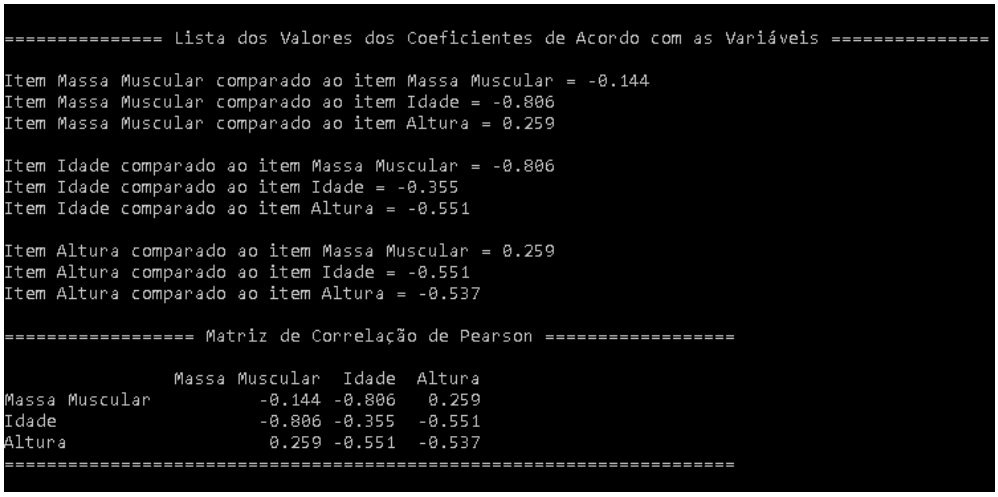

Figura 10- Matriz de Correlação de Pearson incluindo de cada variável com ela própria. Fonte: Autores (2019).

O *software* apresenta os novos coeficientes aceitos pelo valor de corte, agora incluindo os valores dos coeficientes das variáveis com elas próprias. Encerra-se assim o programa, como pode ser observado na Figura 11.

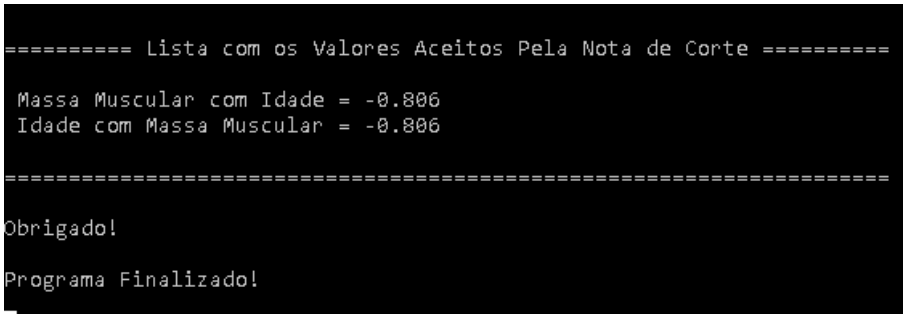

Figura 11- Valores aceitos pelo valor de corte. Fonte: Autores (2019)

Como pode ser observado, não houve alteração neste exemplo, pelo fato de mesmo com o novo cálculo, os valores dos coeficientes não serem aprovados pelo valor de corte.

5. CONSIDERAÇÕES FINAIS

Acredita-se que a pesquisa apresentada alcançou o seu objetivo na medida que gerou um produto para a sociedade, na forma de um *software*. Tal ferramenta pode ser utilizada tanto no mundo corporativo, quanto nas lides acadêmicas. Além disso, possui um caráter eminentemente multidisciplinar, podendo ser aplicada nos mais diversos campos do conhecimento, tais como: Economia, Medicina, Biologia, Matemática, Estatística, Computação, Direito, entre tantas outras. Os pesquisadores interessados no *software* podem entrar em contato diretamente com os autores do artigo.

6. REFERÊNCIAS BIBLIOGRÁFICAS

[1] MIOT, H. A. Correlation analysis in clinical and experimental studies. **J Vasc Bras**. 17(4):275-279. Out.-Dez. 2018.

[2] BOURGARD, B.; GOMES, C. F. S. As variáveis econômicas no Brasil e o PIB: uma análise em períodos de crises financeiras por meio de da correlação de Pearson. **Almanaque Multidisciplinar de Pesquisa**. Vo l 1. No 2. 76-98. 2017.

[3] PIANUCCIA, M. N.; PITOMBOA, C. S. Uso de árvore de decisão para previsão de geração de viagens como alternativa ao método de classificação cruzada. **Revista de Engenharia Civil**. 56:5-13. 2019.

[4] PAIVA, I. S. S.; PEDROSA, N. L.; GALVÃO, M. T. G. Spatial analysis of AIDS and the social determinants of health. **Rev Bras Epidemiol**. 22: E190032. 2019.

[5] FILHO, D. B. F.; ROCHA, E. C.; JÚNIOR, J. A. S.; PARANHOS, R.; NEVES, J. A.; SILVA, M. B. Desvendando os Mistérios do Coeficiente de Correlação de Pearson: O retorno. **Leviathan | Cadernos de Pesquisa Política**. N. 8, pp.66-95, 2014.

[6] ARAÚJO, R. F.; CARAN. G. M.; SOUZA. I. V. P. Orientação temática e coeficiente de correlação para análise comparativa entre dados altmétricos e citações: uma análise da **revista DataGramaZero**. Porto Alegre, v. 22, n. 3, p. 184-200, set/dez. 2016.

[7] TOEBE. M.; FILHO. A. C.; LOPES. S. J.; BURIN. C.; SILVEIRA. T. R.; CASAROTTO, G. Dimensionamento amostral para estimação de coeficientes de correlação em híbridos de milho, safras e níveis de precisão. **Bragantia**, Campinas, v.74, n. 1, p.16-24, 2015.

[8] LIRA, S. A.; NETO, A. C. Coeficientes de correlação para variáveis ordinais e dicotômicas derivados do coeficiente linear de Pearson. **RECIE**, Uberlândia, v. 15, n. 1/2, p. 45-53, jan.-dez. 2006.

[9] PETREÇA, D. R.; BENEDETTI, T. R. B.; SILVA. D. A. S. Validation of the flexibility component of the AAHPERD functional fitness assessment in Brazilian older adults. **Rev Bras Cineantropom Desempenho Hum** 13(6):455-460. 2011.

[10] SILVA, N. E. F.; SOUZA, S. M. A. Finanças e sustentabilidade: Análise da correlação entre a receita líquida e os investimentos sociais e ambientais da natureza do período de 2009 a 2013. **Anais do III SINGEP e II S2IS** – São Paulo – SP – Brasil – 11/2014.

[11] PONTES, A. C. F. **Ensino da Correlação de postos no Ensino Médio**. Universidade Federal do Acre, 2010.

[12] MARTINS. M. E. G. Coeficiente de correlação amostral. **Revista de Ciência Elementar**, 2(02):0069. 2014.