

APLICAÇÃO ANALYTICS COM ESTUDO DE CORRELAÇÃO: FATORES QUE TEM RELAÇÃO COM O SURGIMENTO DE DOENÇAS CARDÍACAS

Alyne Grieger¹
Bruno Eduardo Fernandes Hofmeister²
Eduardo Neves³
João Vicente Miranda⁴
Lucas Comicholli Pinheiro⁵
Luiz Fernando da Costa Junior⁶
Luiz Gustavo de Miranda Gomes⁷
Macauly Gonçalves Trevisiol⁸
Felipe Muller da Cruz Naves⁹
Isaak Assunção¹⁰
Rejane Costa¹¹

RESUMO

Em 8 de maio de 2019 os alunos de todos os semestres de 2019 do curso de Análise e Desenvolvimento de Sistemas participaram do III Aula Aberta. No evento estes foram organizados em equipes com no máximo 10 integrantes, para cada equipe foi entregue um problema de ordem quantitativa. Os professores indicaram um Dataset (conjunto de dados disponibilizados para pesquisa) para cada grupo com dois objetivos a serem alcançados. O primeiro, desenvolver uma rotina automatizada para captura, tratamento e carga dos dados de um site da web um Banco de Dados Não Relacional; o segundo, extrair os dados, armazená-los em uma planilha e realizar um estudo estatístico de correlação para identificar quais variáveis apresentam relação direta com a variável alvo determinada no problema apresentado. Para desenvolver a rotina automatizada utilizamos a linguagem de programação Python 3, com as bibliotecas SELENIUM e PANDAS, para armazenar os dados foi utilizado o Banco de Dados MongoDB. O problema apresentado para nosso grupo tem origem nos dados extraídos no site <https://www.kaggle.com/ronitf/heart-disease-uci> sob o título de *Heart Disease UCI* tendo como alvo a identificação das variáveis que tem correlação com a presença de doença cardíaca nos pacientes. A base de dados foi obtida em pesquisa na Cleveland Clinic Foundation. A variável alvo refere-se à presença de doença cardíaca no paciente. É inteiro valorizado de 0 (sem presença) a 4. Na rotina desenvolvida foi possível identificar o potencial para a automatização dos processos de captura, extração e análise de dados como forma de identificar soluções para problemas que tenham origem dados disponíveis na internet. A partir da implementação desta rotina e do conhecimento das tecnologias aplicadas os integrantes da equipe identificaram diversas outras aplicações. Após execução da rotina que capturou e armazenou os dados foram identificadas e extraídas as variáveis independentes A, B, C e D do dataset original e a variável F como dependente. Nos testes realizados as variáveis que apresentaram correlação significativa foram $A = 0,8767$ e $B = 0,98788$.

Palavras-chave: Correlação. Analytics. Estatística. Doença cardíaca. Python.

¹ Titulação: Graduanda em ADS. Instituição atual: Faculdade Cesusc. E-mail: alyne.grieger@outlook.com

² Titulação: Graduando em ADS. Instituição atual: Faculdade Cesusc. E-mail: brunohofmeister11@gmail.com

³ Titulação: Graduando em ADS. Instituição atual: Faculdade Cesusc. E-mail: eduardocn1999@gmail.com

⁴ Titulação: Graduando em ADS. Instituição atual: Faculdade Cesusc. E-mail: jvmarroni99@gmail.com

⁵ Titulação: Graduando em ADS. Instituição atual: Faculdade Cesusc. E-mail: lucascpinheiro.studies@gmail.com

⁶ Titulação: Graduando em ADS. Instituição atual: Faculdade Cesusc. E-mail: luizjr20@outlook.com

⁷ Titulação: Graduando em ADS. Instituição atual: Faculdade Cesusc. E-mail: sncguga@gmail.com

⁸ Titulação: Graduando em ADS. Instituição atual: Faculdade Cesusc. E-mail: macaulygt2105@gmail.com

⁹ Titulação: Graduando em ADS. Instituição atual: Faculdade Cesusc. E-mail: felipe.naves99@gmail.com

¹⁰ Titulação: Graduando em ADS. Instituição atual: Faculdade Cesusc. E-mail: isaakassuncao@gmail.com

¹¹ Titulação: Professora Mestre em Engenharia de Produção. Instituição atual: Faculdade Cesusc. E-mail: rejanecestafloripa@gmail.com