UNIFOR – UNIVERSIDADE DE FORTALEZA

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

MATEMÁTICA PARA COMPUTAÇÃO

TRABALHO AV1

Fortaleza - 2022

EMANUEL VIDAL RODRIGUES DO MONTE E SILVA – 2225424

RAFAEL SABOIA DUNBAHEVITZ – 2224942

MATEMÁTICA PARA COMPUTAÇÃO

TRABALHO AV1

Trabalho apresentado para composição de nota da AV1 da disciplina de Matemática para Computação do curso de Análise e Desenvolvimento de Sistemas da Universidade de Fortaleza

Professor: Jose Rubens Rodrigues de Sousa

Fortaleza – 2022

SUMÁRIO

1 INTRODUÇÃO

Neste trabalho, fizemos a análise de 3 conjuntos de pessoas, fornecidos pelo professor da disciplina, para identificar suas características e conseguir responder as perguntas definidas no tópico "*Trabalho da AV1*" no Unifor EAD (também conhecido como AVA).

O referente trabalho, trata de uma dinâmica extensivamente utilizada no mercado de trabalho. Aqui, nos referimos ao cientista de dados, profissional que trabalha de diversas formas, com informações, sejam elas acumuladas ou não. E são denominados como Big Data, os trabalhos em que possuem grandes volumes de aspectos, que podem compor ou não direta ou indiretamente um banco de dados.

A obtenção de tais informações pode ocorrer através de diversas formas, como o apresentado no problema em questão, através de uma coleta, que conforme os dados contidos na pesquisa, observa-se equívocos de escrita, em que aparenta ser resultado um trabalho manual, mas, também existem formas automatizadas de obtenção destes dados e até trabalhos em larga escala, atividade conhecida como Data Mining. Aqui, os aspectos são irregularmente distribuídos em alguma população/base de dados/coleção de objetos, e para que possam ser utilizados no trabalho, devem ser coletados de forma extensiva e completa, podendo ocorrer a necessidade de tratamento ou não.

Mas, você deve estar se perguntando, "Para que servem tantos dados, por que trabalhar com eles?". Bom, como poderemos ver a seguir, o trabalho do cientista desta área em questão, a ciência conhecida como Data Science, tem como o principal objetivo o melhoramento da tomada de decisões. Observamos em Provost e Fawcett (2013):

> "[...] o objetivo final da *data Science* é melhorar o processo de tomada de decisão, tendo em vista que este é diretamente de interesse do negócio. [...] A *data Science* é diferente de outros métodos de processamento de dados e tem dominado o mercado e atenção dos negócios. Estes, utilizam do DDD (*data-driven decision-making),* que significa que tomam decisões baseadas na análise dos dados e não baseado na intuição." (tradução livre, Provost e Fawcett, 2013)

Como ferramenta de análise deste trabalho e da disciplina em curso, utilizamos de forma extensa o Diagrama de Venn, proposto por John Venn em 1880, como forma de melhor entender os agrupamentos de dados, os conjuntos, de forma visual. Lá atrás, a única forma de visualização que foi aceita amplamente entre os pesquisadores da época, os chamados Círculos de Euler foram a inspiração inicial para construção do diagrama (Venn, 1880).

2 METODOLOGIA

Toda metodologia, resultados, respostas da questões e conclusões podem ser acessadas através do link para o site criado: *https://trab-mat.vercel.app/*

Para conseguirmos fazer a análise, primeiro utilizamos da conversão dos arquivos fornecidos, que se encontravam inicialmente em formato ".csv" (*comma-separated values*), para um formato que fosse mais fácil a manipulação, no caso, escolhemos trabalhar com o software Microsoft Excel, por isso, para que todas as linhas e colunas ficassem bem alinhadas, utilizamos um conversor online para converter os arquivos para o formato ". xlxs" (*Microsoft Excel Open XML Spreadsheet*).

Também foi preciso fazer um pré-tratamento nos dados, antes de tentar efetuar qualquer análise. Os dados foram fornecidos com diversos erros de digitação, input, datas incorretas, colunas trocadas de lugar, utilização de caracteres inválidos. De acordo com o site Pós-Graduando (2016):

> "Anotar os dados de forma imprecisa: Por descuido ou mau planejamento, pode-se acabar com informações menos precisas que o necessário. O exemplo comum é da pesquisa em que se pergunta a idade do pesquisado, e não sua data de nascimento. No primeiro caso, há maiores chances de respostas incorretas, por arredondamento, esquecimento ou mentira do pesquisado."

Assim, para seguir com o tratamento dos dados, foi definido um algoritmo (conjunto de instruções para realização de uma tarefa com intuito de resolver um problema) pela equipe, para tratar de tais problemas. Este algoritmo consiste dos seguintes passos:

- Removendo linhas com primeiros nomes totalmente ilegíveis;
- Corrigir os primeiros nomes e sobrenomes que fazem algum sentido, porém estão apenas com os caracteres embaralhados;
- Se apenas um dos sobrenomes é igual a um do pai ou mãe, corrigi-lo e descartar os demais. Se ambos os sobrenomes estiverem presentes, corrigir ambos;
- Se o primeiro nome for acompanhando de 1 ou mais iniciais de sobrenomes (em maiúsculo), mantém-se como está;
- Se o primeiro nome estiver correto e for acompanhado de uma inicial de sobrenome, porém existindo outro sobrenome de forma completa, remove-se o que tem somente a inicial;
- Se estiver presente apenas o nome sem sobrenome, mantém-se como está;
- Inverter a coluna nome da mãe - nome do pai, já que seus dados estão trocados;
- Padronizar coluna do sexo, utilizando M e F como padrão.

Com isso, conseguiu-se neutralizar, quase que por completo, os efeitos negativos dos dados errados. Além disso, a coluna de "Data de Nascimento" e a coluna de "Data da Dengue", ambas possuem datas com discrepâncias, como, datas de nascimento acima de 2022, ou muito antigas, e datas da dengue acima de 2022 ou mais antigas que a data de nascimento. Nestes

casos, foram ignorados tais discrepâncias, já que os dados considerados mais importantes pela equipe são os nomes, para que seja efetuada a correta análise dos problemas propostos. Apesar da correção feita com as colunas de “Nome do Pai” e “Nome da Mãe”, tais não foram utilizadas para a análise feita.

Contudo, mesmo após o tratamento dos dados, a equipe não alcançou um resultado que julgou bom, e foi decidida uma nova abordagem. Assim, já que os integrantes estão no curso de Análise e Desenvolvimento de Sistemas, fez muito sentido construir algo com as habilidades que são pertinentes à tal área, como utilização da linguagem de programação *Javascript*. Assim, decidiu-se montar um *website*, com um intuito mais de representar uma *dashboard*, com a análise dos dados, os Diagramas de Venn e apresentação das listas de resultados.

Foi utilizado, no *Javascript*, a coleta dos dados em formato *json* (que são os dados obtidos após a limpeza das planilhas, convertidos para o formato *JavaScript Object Notation*, utilizando uma ferramenta *online https://kinoar.github.io/xlsx-to-json/)* para conseguirmos fazer procuras, iterações, manipulações em *arrays* e objetos. Esta coleta foi feita com a função *fetch*. Após, para cada questão apresentada, os dados obtidos são trabalhados em diferentes objetos, um com o conjunto dos alunos, outro com o conjunto das pessoas que tiveram dengue e o conjunto das pessoas que pegaram ônibus.

Com isso, dependendo das especificidades de cada questão, foram utilizados diferentes métodos. Para questões que envolvem a intersecção dos três conjuntos, foram utilizados os *methods* .reduce e .*filter*, assim criando um terceiro grupo, com as pessoas pertinentes a todos os três conjuntos. Para questões que precisamos fazer a “remoção”(por exemplo: A - B) de um conjunto de outro, foram utilizados os *methods .filter* e .*includes*. E por último, para conseguir as intersecções entre apenas 2 conjuntos, foram utilizados os *methods filter* e *indexOf*.

Na elaboração dos Diagramas de Venn, estes, já representam de forma correta os conjuntos com os dados do arquivo *json*, sem precisar que realizemos alguma função mais complexa. Para tal, foi utilizada a *library Chart.js*(*https://www.chartjs.org/*). Com isso, precisamos apenas fazer a configuração de alguns parâmetros para exibição do gráfico e realizar sua geração.

Para gerar as listas finais, chamadas de “relatórios”, que consistem nas listas de nomes e outros parâmetros exigidos diferentemente por cada questão, foi utilizado um algoritmo de busca, com complexidade de tempo quadrática $O(N^2)$, que mesmo devagar em algumas utilizações, serviu de forma integral dado o baixo número de iterações necessárias. Após

realizada a conferência dos dados que precisam ser exibidos, eles são gerados em uma nova janela *html*.

3 RESULTADOS

Abaixo, podemos ver um exemplo de como os resultados serão apresentados no site criado. Um total de 4 Diagramas de Venn foram criados, que são suficientes para abordar todas as questões solicitadas. As listas "respostas" das questões, podem ser acessadas através do botão "gerar lista" que se encontra em cada item de 1(um) à 10(dez).

Exemplo:

"*1) Relatório Educação: Informar nome, data de nascimento e id dos cidadãos de XPTO que frequentaram a escola, menos os cidadãos que tiveram dengue*."

Como podemos observar no Diagrama de Venn, apresentado na figura X, quando consideramos apenas os conjuntos Alunos(A) e Dengue(D), podemos identificar os seguintes dados: X frequentaram a escola, Y tiveram dengue, e X – Y foram Z, assim, como o estudado na matéria, o que pede-se na questão 1, é A – D, que resultou em Z pessoas. Seus nomes, datas de nascimento e ids podem ser encontrados na tabela gerada no site (*https://trab-mat.vercel.app/)*

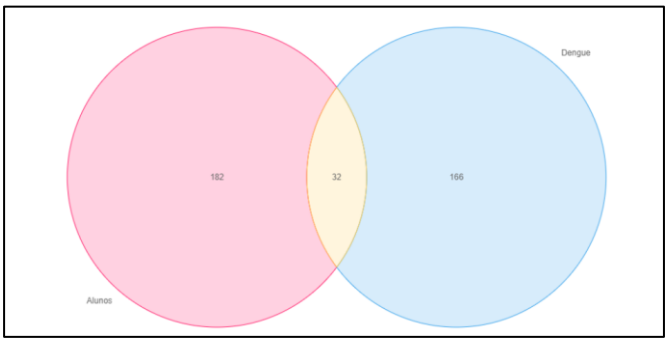

Figura 1 Diagrama de Venn

4 CONCLUSÃO

Ao observar os grupos formados pela proposta do docente, podemos formar as seguintes suposições:

Foi constatado que o número de pessoas que frequentam a escola e tiveram dengue, foi expressivamente maior quando comparado aos outros conjuntos. Então, à título deste trabalho, podemos formular a hipótese de que na escola há um possível foco de dengue, o qual está contagiando as pessoas que por lá transitam.

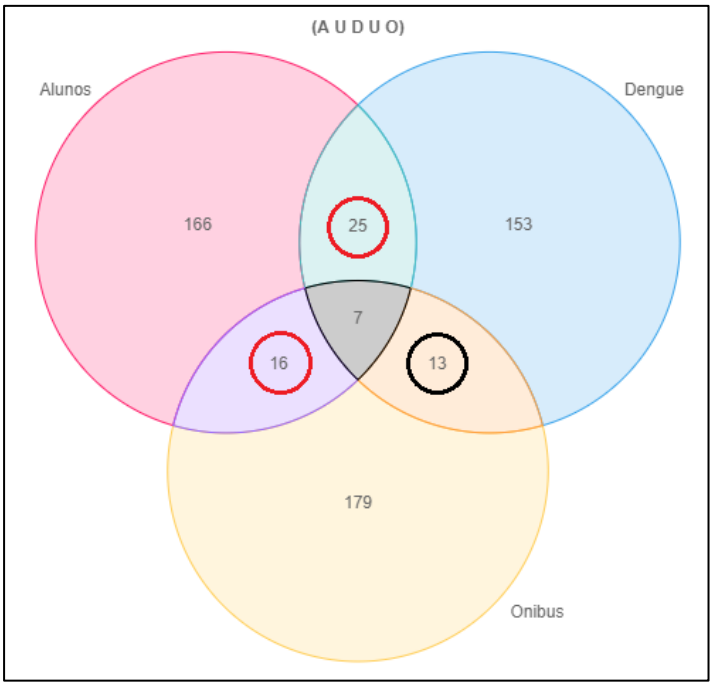

Figura 2 Diagrama de Venn referente à Conclusão: Hipótese de possível foco de dengue na escola

Além disso, observou-se uma baixa quantidade de pessoas que utilizam ônibus, como meio de transporte para se deslocar até a escola, quando comparado ao total do conjunto. Sendo assim, uma pauta plausível para uma discussão com a prefeitura de XPTO, sobre uma possível deficiência na infraestrutura do transporte público dificultando o acesso ao modal ônibus.

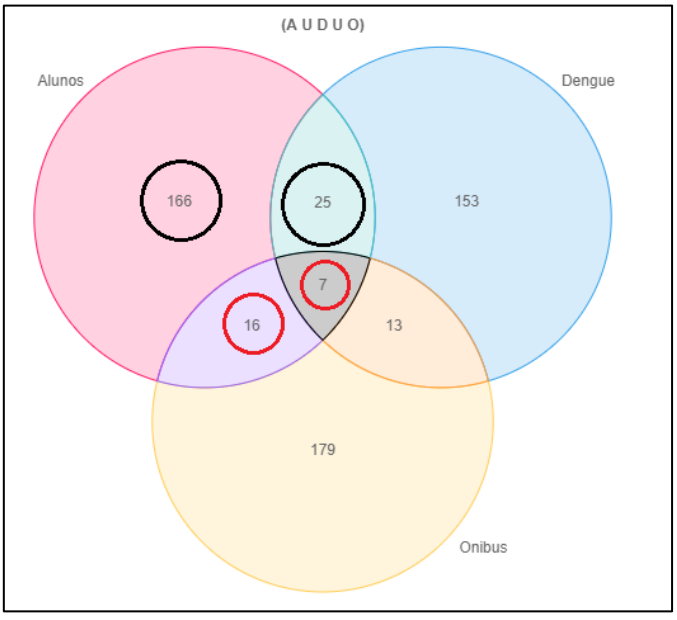

Figura 3 Diagrama de Venn referente à discussão de melhoria do transporte urbano.

5 REFERÊNCIAS BIBLIOGRÁFICAS

Nove (9) erros comuns na coleta de dados de uma pesquisa científica. Pós-graudando, 22 de jul. de 2016. Disponível em: <https://posgraduando.com/erros-coleta-dados/>. Acesso em: 31 de ago. de 2022.

J. Venn M.A. (1880) I. On the diagrammatic and mechanical representation of propositions and reasonings , Philosophical Magazine Series 5, 10:59, 1-18.