

Trabalho de Data Science

Universidade do País das Maravilhas

Tópicos em Engenharia de Dados

Docente: Bruno Cipriano

Discentes: João Eleutério

MCiD | 2023/24

Índice

[Índice de Figuras 3](#_Toc168743736)

[Introdução 4](#_Toc168743737)

[Tecnologias usadas 5](#_Toc168743738)

[XAMPP 5](#_Toc168743739)

[PhpMyAdmin 5](#_Toc168743740)

[Jupyter Notebook 5](#_Toc168743741)

[Python 5](#_Toc168743742)

[Dificuldades e Problemas encontrados 6](#_Toc168743743)

[Procedimentos e resultados Dados 7](#_Toc168743744)

[Referências 11](#_Toc168743745)

[Bibliografia 12](#_Toc168743746)

# Índice de Figuras

[Figura 1- Diagrama ilustrativo do projeto 7](#_Toc167288836)

[Figura 2 - Correlação entre média final com a média de entrada 9](#_Toc167288837)

[Figura 3 - Scatterplot da relação entre média final com a média de entrada 9](#_Toc167288838)

[Figura 4 - Correlação entre média final com a os cursos de origem 10](#_Toc167288839)

[Figura 6 - Relação entre Média Final e a Quantidade de alunos de Curso Profissional 10](#_Toc167288840)

[Figura 5 - Relação entre Média Final e a Quantidade de alunos de Curso de Ensino Geral 10](#_Toc167288841)

# Introdução

Após concluir o mestrado em Ciência de Dados, a empresa Dados Complicados contratou-me para atuar como "Engenheiro de Dados Júnior". Fui designado a um projeto multidisciplinar relacionado a Ciência de Dados e Educação, destinado ao cliente Universidade do País das Maravilhas. Foi fornecido um *dataframe* contendo dados dos alunos, incluindo número de aluno, nome, sobrenome, curso, média de entrada, se haviam concluído o curso e a média final. O Cientista de Dados informou que deveria realizar uma Análise Exploratória de Dados (EDA) e alertou que os dados poderiam conter erros, solicitando que eu comunicasse os mesmos ao cliente.

Um email foi enviado solicitando a criação de um *dataframe* para análises de correlações sobre dados de alunos de uma Universidade. Sendo que, naquele momento, apenas foi fornecido um arquivo .sql com a tabela "aluno" e um CSV com dados complementares. Esta tabela não estava normalizada, sendo necessários procedimentos de engenharia reversa para entender sua estrutura, sendo que, também, não havia informação de como o CSV se conecta à mesma. É suposto que a tabela contenha dados de alunos de apenas três cursos - Informática, Psicologia e Comunicação. Caso existisse algum erro ou problema nos dados, era pedido que estes fossem reportados.

O projeto foi dividido em duas partes em que a primeira consistia só no ficheiro .sql, a inserção na base de dados e encontrar tipos de dados “estranhos” ou inválidos que resultassem na separação de dados em ficheiros para responder à pergunta “Existe alguma relação entre a média de entrada na licenciatura e a média de conclusão da licenciatura?”. Já na segunda parte do projeto foram, então, inseridos os dados do .csv para responder à pergunta “Existe alguma relação entre a origem dos alunos (curso profissional ou curso do ensino geral) e a média final da licenciatura?”.

# Tecnologias usadas

No projeto foram utilizadas várias tecnologias como:

## XAMPP

O XAMPP é uma distribuição compacta e leve do Apache que inclui as tecnologias mais utilizadas para o desenvolvimento web num único pacote. O seu conteúdo, tamanho reduzido e portabilidade tornam-no uma ferramenta perfeita para estudantes que criam e testam aplicações em PHP e MySQL (Dvorski, 2007).

## PhpMyAdmin

O phpMyAdmin é provavelmente a ferramenta de administração mais frequente para o MySQL, pois permite criar, modificar e excluir bases de dados e tabelas, inserir, alterar e remover registros de dados, além de importar e exportar bases de dados inteiras, entre outros (Kofler, 2005).

## Jupyter Notebook

O Jupyter Notebook é uma plataforma interativa que facilita a criação e o compartilhamento de documentos que incluem código executável, visualizações e texto explicativo. É amplamente adotado para análise de dados, *machine learning,* computação científica e ensino. Suporta diversas linguagens de programação, sendo que o Python é o mais utilizado.

## Python

Python é uma linguagem de programação de alto nível utilizada pela sua simplicidade e eficiência. Foi desenvolvida para melhorar a legibilidade do código, utilizando uma sintaxe clara e intuitiva, o que permite aos programadores escreverem menos código em comparação com outras linguagens, como C++ ou Java.

# Dificuldades e Problemas encontrados

Durante o projeto, várias dificuldades foram encontradas. No início, houve desafios para entender o funcionamento do XAMPP, DBeaver e phpMyAdmin, já que essas ferramentas nunca haviam sido utilizadas antes. Após aprender a conectar o DBeaver ao phpMyAdmin e usar o XAMPP, surgiu um problema ao replicar a tabela com o dump recebido. A solução adotada foi inserir as colunas visíveis no dump (nr\_aluno, nome, apelido, curso, media\_inicio, curso\_concluido e media\_final) todas como varchar(255) - nullable (permitindo até 255 caracteres ou null) para identificar os tipos de dados.

Outro desafio foi identificar dados "estranhos" ou inválidos. Inicialmente, foram verificadas as condições padrão, como a impossibilidade de valores null, médias não podendo ser menores que 0 ou maiores que 20, e cursos restritos a Informática, Psicologia e Comunicação. Além disso, foram considerados casos em que o curso foi concluído, mas a média final era inferior a 10. Embora essa abordagem inicial tenha identificado 28 dados incomuns, o cliente indicou que havia mais. Semanas depois, foi descoberto que a query estava errada e que era necessário excluir os dados de alunos que ainda não haviam concluído o curso, resultando então na identificação de 43 dados.

Outro problema, relacionado à programação, foi a conexão com a base de dados (realizada antes de receber as instruções corretas) e a necessidade de pesquisar como criar alguns gráficos.

Outro problema, mas mais direcionado para programação foi a conexão com a BD (feito antes de explicarem como se faz), e alguns gráficos precisei de ir pesquisar como se faziam.

# Procedimentos e resultados Dados

Uma imagem com diagrama, captura de ecrã, file, Esquema

Descrição gerada automaticamenteFoi pedido que se transformasse os dados em dois ficheiros como mostra Figura 1 - Diagrama ilustrativo do projeto :

Figura 1- Diagrama ilustrativo do projeto

No início do projeto foi feita a configuração da conexão da base de dados com o PhpMyAdmin através do XAMPP, sendo que depois foi feita a criação da tabela e a inserção de dados na bsae de dados no PhpMyAdmin. Quando os dados estavam na base de dados foram feitas queries para descobrir valores “estranhos” ou inválidos. Primeiramente foi analisada uma visão geral dos dados como:

* select distinct nr\_aluno from aluno
* select distinct nome from aluno
* etc (o mesmo para o resto das colunas)

Foi verificado, com isto, nomes em inglês e apelidos com parenteses, mas o cliente avisou que eram válidos. Também existiam muitos cursos que não os indicados no email. Deu para observar que medias tinham valores abaixo de 0 e acima de 20, curso\_concluido só tinha valores 0 ou 1. Nas restantes colunas não foram percetíveis outros erros ou informações pertinentes para esta etapa.

Verificou-se se existiam valores nulls:

* select \* from aluno where nr\_aluno is null or nome is null or apelido is null or curso is null or media\_entrada is null or curso\_concluido is null;

Nesta query não foram feitas todas as colunas porque, caso o aluno ainda não tenha acabado o curso, não tem nota final. Por isso, é válido haver nulls nessa coluna. Com isto descobriu-se que existiam valores a null em curso.

Depois disso, pensou-se que caso tenha média de entrada menor que 10 é um inválido porque assim não conseguia entrar no curso:

* select \* from aluno where media\_entrada < 10

Assim percebeu-se que havia alunos com media\_entrada menor que 10.

Pensando neste racicionio de poder entrar ou poder acabar o curso, pode-se criar uma query que vê se o aluno terminou o curso e tem menos que 10 ou null, ou se o aluno não acabou o curso e tem mais de 10 (é válido um aluno não ter acabado o curso e ter nota final, pois pode ter chumbado):

* select \* from aluno where (curso\_concluido = 0 and media\_final > 10) or (curso\_concluido = 1 and media\_final < 10) or (curso\_concluido = 1 and media\_final is null)

Todos os alunos que apareceram tinham dados inválido de media\_final menor que 0 ou null.

Depois de ter falado com o cliente, o mesmo decidiu não querer alunos que não tenham acabado a licenciatura também. Então foi feita a seguinte query:

* select \* from aluno where curso\_concluido = 0

Fazendo a query para mostrar os dados “estranhos” /inválido fica:

* select \* from aluno where media\_entrada < 10 or media\_entrada > 20 or media\_final < 10 or media\_final > 20 or media\_final is null or curso\_concluido = 0 or curso is null or curso not in ('Informática','Psicologia','Comunicação');

Assim podemos observar 43 dados “estranhos” ou inválidos.

Passado isto, passando para o jupyter foi feita a conexão com a base de dados e criou-se/escreveu-se num ficheiro de csv para os dados que são válidos, e num ficheiro de texto estes dados “estranhos” /inválidos. O cliente também quis que se colocasse os alunos que não terminaram o curso num ficheiro de texto também.

Feito isto somente com os dados válidos realizou-se a EDA e descobriu-se que não existem valores null,

Na coluna nome, o top 5 são: Rita – 114 alunos, Avelino – 78 alunos, Rodrigo – 78 alunos, Jorge – 76 alunos e Pedro - 75 alunos e, no bottom 5 são: Thomas Anderson – 1 aluno, Luke – 1 aluno, Anakin Skywalker – 1 aluno, Beatriz – 36 alunos, e Diana – 43 aluno.

Na coluna apelido, o top 5 são: Morais – 170 alunos, Castelo -167 alunos, Sampaio – 155 alunos, Pires – 150 alunos, Cintra – 148 alunos e, no bottom 5 são: (Neo) – 1 aluno, Skywalker - 1 aluno, (Lord Valder) – 1 aluno, Pereira – 125 alunos e Silva - 127 alunos.

No curso, o top 3 (únicos) são: Psicologia – 587 alunos, Informática – 576 alunos, Comunicação – 554 alunos.

Na media\_entrada, o top 5 são: 17 – 182 alunos, 20 – 179 alunos, 10 -169 alunos, 14 – 156 alunos, 18 – 155 alunos e, no bottom 5 são: 16 – 132 alunos, 13 – 145 alunos, 12 – 146 alunos, 15 – 149 alunos, 11 – 151 alunos**.** Em que a média aproximadamente 15.0.

No curso\_concluido, só aparece o valor 1 que significa terminou o curso.

Na media\_final, o top 5 são: 13 – 182 alunos, 20 – 170 alunos, 19 - 169 alunos, 10 – 166 alunos, 12 – 159 alunos e, no bottom 5 são: 15 – 129 alunos, 18 – 130 alunos, 17 – 145 alunos, 16 – 150 alunos, 14 – 158 alunos. Em que a média aproximadamente 14.9.

Após o EDA consegue-se responder à pergunta “Existe alguma relação entre a média de entrada na licenciatura e a média de conclusão da licenciatura?”

Uma imagem com texto, captura de ecrã, diagrama, Retângulo

Descrição gerada automaticamenteUma imagem com texto, captura de ecrã, file, Gráfico

Descrição gerada automaticamenteA média de entrada e média final apresenta uma correlação negativa forte (-0,88, estando próxima de -1). Isto quer dizer, que quanto maior for a média inicial, menor vai ser a média final e, vice-versa. Confirmando a resposta, não havendo 100% dos dados em que se a média inicial elevada a final irá ser baixa ou vice-versa, com a visualização de um scatterplot observou-se 1 ponto separado dos demais. Então, fazendo manualmente a pesquisa de quais são e observar-se que os alunos com número de aluno 2210000, 2210100 e 2290110 têm uma nota de entrada alta e igual a nota final. Neste caso, acontece que as duas são iguais e muito elevadas, por isso, não é proporcional.

Figura 2 - Correlação entre média final com a média de entrada

Figura 3 - Scatterplot da relação entre média final com a média de entrada

Com isto concluído, foi entregue um ficheiro .csv com duas colunas NR\_ALUNO e CURSO\_ORIGEM em que o nr\_aluno correspondia ao nr\_aluno da tabela e o curso\_origem, se o aluno vinha de curso profissional ou de ensino geral.

Fez-se a renomeação das colunas para ficarem com o mesmo nome (estas estão em maiúsculas), pelo menos o nr\_aluno. Depois foi feito um *merge* dos *dataframes* em que o nr\_aluno ligava os dois e, caso não existisse, ficava a NaN.

Com isto, o top 3 (únicos) são: Geral - 1100 alunos, Profissional – 597 alunos e não sabemos de 20 alunos.

Com isto, fez-se a correlação respondendo assim à pergunta “Existe alguma relação entre a origem dos alunos (curso profissional ou curso do ensino geral) e a média final da licenciatura?”

A média final e se é curso profissional apresenta uma correlação positiva fraca (0,34). Isto quer dizer que, quando é curso profissional, maior vai ser a média final e caso seja curso geral, menor vai ser a média final. Confirmando a resposta, não havendo 100% dos dados em que se a média final é elevada e o curso ser sempre profissional, ou se a média final ser baixa e ser sempre curso ensino geral ou vice-versa, com a visualização de um gráfico de barras agrupadas observa-se as quantidades nas diferentes notas dos cursos. Notando assim, que é verdade que quanto mais baixo, mais provável ser curso de ensino geral e quanto maior a média final, mais provável ser curso profissional.

Uma imagem com texto, captura de ecrã, diagrama, número

Descrição gerada automaticamente

Figura 4 - Correlação entre média final com a os cursos de origem

Uma imagem com texto, captura de ecrã, file, Gráfico

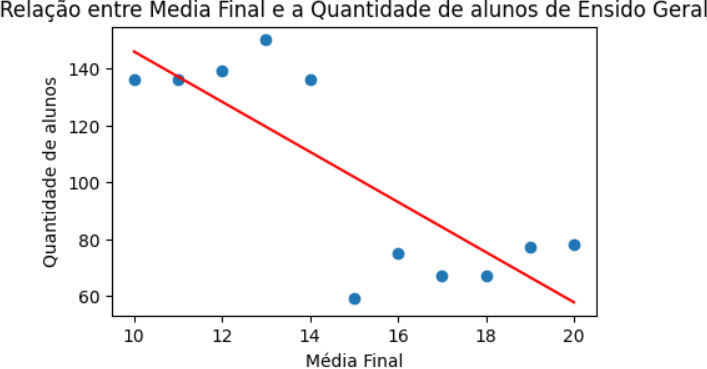
Descrição gerada automaticamente

Figura 6 - Relação entre Média Final e a Quantidade de alunos de Curso Profissional

Figura 5 - Relação entre Média Final e a Quantidade de alunos de Curso de Ensino Geral

Correlações feitas e comprovadas sempre acompanhadas com gráficos com ilustração em que a linha mostra se é positiva (caso a linha esteja a crescer) ou se é negativa (caso a linha esteja a diminuir).

# Referências

<https://jupyter.org/>

<https://www.python.org/>

Dvorski, D. D. (2007). Installing, configuring, and developing with Xampp. *Skills Canada*

*Kofler, M. (2005). phpMyAdmin. The Definitive Guide to MySQL5, 87-116.*

# Bibliografia

EDA - Exploratory Data Analysis – É uma abordagem à análise de conjuntos de dados de modo a resumir as suas características principais.

BD – Base de dados - É um conjunto organizado de informações estruturadas ou de dados geralmente armazenados eletronicamente num sistema informático.

XAMPP - X (para qualquer dos diferentes sistemas operativos), Apache, MariaDB, PHP, Perl.

Nan - Not a Number é usado em contextos numéricos para representar um valor que não é um número válido.