Universidade Federal de Viçosa – *Campus* Florestal

CCF 425 - Introdução à Ciência dos Dados

Documentação Final

Trabalho Prático

Professor: Fabrício A. Silva

Arthur Sales da Silva - 3501
Mateus Coelho - 3488
Gabriel Morais - 3497

Florestal, MG

2021

Sumário

Parte 1

A parte 1 do trabalho foi de escolher a *database* e procurar realizar 20 perguntas para serem respondidas ao longo das partes 2, 3 e 4. Sabendo disso, a o tema escolhido foi ensino superior no Brasil no ano 2019, e a tabela utilizada foi “Microdados Censo da Educação Superior 2019”, que pode ser encontrada no link: https://download.inep.gov.br/microdados/microdados_educacao_superior_2019.zip

As 20 perguntas feitas foram:

1) Quantos homens e quantas mulheres cada área possui?
2) Qual a distribuição de alunos matriculados por faixa etária?
3) Qual a distribuição de docentes em exercício por faixa etária?
4) Quantos campus ofertam determinado curso? (Aqui o pensamento é o usuário entrar com o nome do curso que deseja)
5) Qual a relação entre o índice de desistência e idade para os 3 cursos de maior nível de desistência?
6) Qual o grau de formação mais comum dos docentes? Comparação entre IES particulares e públicas.
7) Quantos candidatos se candidatam para uma vaga em IES em cada Estado do país?
8) Há algum aluno que repete na tabela(fazendo dois cursos por exemplo)? Se sim, qual o máximo de repetições, e quantos alunos repetem?
9) Qual o curso com maior desvio padrão em relação às idades dos matriculados?
10) Qual curso tem a idade média mais baixa? E a mais alta?
11) Qual a probabilidade de um aluno matriculado em curso da área de computação ter algum tipo de deficiência e não desistir do curso? Compare com a probabilidade disso ocorrer no curso com menor índice de desistência.
12) Dado um número x de cursos ofertados em uma IES, essa IES se encontra em uma capital?
13) Qual o curso com mais alunos que fazem projeto de extensão? Ponderado de acordo com número de alunos no curso e número de IES que ofertam o curso
14) Qual a distribuição das cores dos candidatos por meio de ingresso na instituição?
15) Comparação da relação entre estudantes branco x não brancos nas instituições públicas e privadas
16) Qual a concentração de docentes por região? A concentração em capitais é, proporcionalmente, maior?
17) Qual o top 5 cursos mais concorridos? Em quais faculdades a concorrência por esses cursos é maior?
18) Qual a relação de matrículas de alunos portadores de necessidades especiais por subárea?
19) Qual a relação do número de concluintes de cursos de graduação presencial x à distância, por região?
20) Dado um conjunto de características de um indivíduo, qual área é a mais provável desse indivíduo pertencer?

Parte 2

Na parte 2, realizamos parcialmente o tratamento dos dados, excluindo algumas linhas e colunas que não utilizamos futuramente.

Para realizarmos o tratamento de dados, inicialmente realizamos um tratamento inicial, pois o tamanho total dos datasets excedia o limite suportado pelas memórias RAMs dos nossos computadores quando os carregávamos. Neste tratamento inicial derrubamos todas as colunas que não iríamos utilizar para responder às perguntas definidas na parte anterior do trabalho e, também, executamos o comando drop_duplicates em todos datasets, para excluir potenciais linhas duplicadas nos dados. Com isso, conseguimos reduzir consideravelmente o tamanho dos dados e, para conseguir carregá-los de maneira mais eficiente posteriormente, geramos novos arquivos csv com esses tratamentos aplicados.

As colunas que foram retiradas de cada base de dados podem ser encontradas nos arquivos “drop.txt”, que estão na branch main, na pasta “RawToStaged”, a qual também contêm os códigos executados para esse tratamento inicial. É possível encontrar também as colunas que serão utilizadas, ou seja, todas as colunas de uma determinada tabela menos as colunas que foram removidas, nos arquivos “used.txt”, na pasta “UsedColumns”. Após a transformação inicial analisamos as linhas e colunas restantes que poderiam nos causar problemas e, então, utilizamos os códigos encontrados na pasta “StagedToCurated” para realizar um drop nessas linhas, para evitar fazermos um estudo com dados que poderiam gerar ruídos na análise. Para todas as tabelas dropamos todas as linhas que correspondiam a um ano anterior a 2011, pois há algumas colunas com valores nulos para objetos criados antes de 2011 que causariam problemas nas nossas análises.

Para a tabela de alunos, retiramos também as linhas nas quais as informações referentes à cor de pele e deficiência do aluno não estavam presentes, ou que estivessem presentes, mas como “aluno não quis declarar”, também foram removidas as linhas referentes a alunos já falecidos. Na tabela de cursos, as linhas com tipo de grau acadêmico vazio foram removidas, pois indicam cursos com nível acadêmico sequencial ou de formação específica, portanto não são interessantes para a nossa análise. Para a tabela de docentes, excluímos as linhas referentes a docentes falecidos.

Todos os documentos citados acima podem ser encontrados no repositório do trabalho que se encontra no GitHub, neste link: https://github.com/GabrielMoraisReis/TP-CDD/tree/main/Parte2

Parte 3 e 4

Pergunta 1

Para esta pergunta pegamos somente as colunas "CO_CINE_ROTULO", "TP_SEXO" e "TP_SITUACAO" da tabela de aluno, e as colunas "NO_CINE_AREA_GERAL" e "CO_CINE_ROTULO" da tabela Aux Cine Brasil, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Fizemos a análise apenas dos alunos com situação de matrícula cursando ou trancada;
- Foi preciso converter as coluna “CO_CINE_ROTULO” da tabela aluno e “CO_CINE_ROTULO” da tabela Aux Cine Brasil para *str* para ser possível realizar o merge das duas tabelas;
- Utilizamos o FacetGrid da *SeaBorn* para exibir os resultados.

Pergunta 2

Para esta pergunta pegamos somente as colunas "TP_SITUACAO" e "NU_IDADE" da tabela de aluno, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Filtramos por apenas alunos que estão efetivamente cursando no momento algum curso (matrícula trancada também não entra);
- Dividimos os dados em alunos até 70 anos e alunos com mais de 70 anos, para diversificar a análise em dois grupos;
- Utilizamos o HistPlot do SeaBorn para exibir os resultados.

Pergunta 3

Para esta pergunta pegamos somente as colunas "ID_DOCENTE", "TP_SITUACAO" e "NU_IDADE" da tabela de docente, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Alguns professores da base estão em exercício em mais de 5 lugares. Consideramos esses professores como erro da base;
- Utilizamos o HistPlot do SeaBorn para exibir os resultados.

Pergunta 4

Para esta pergunta pegamos somente as colunas "CO_LOCAL_OFERTA" e "CO_CINE_ROTULO" da tabela de curso, e as colunas "CO_CINE_ROTULO" e "NO_CINE_ROTULO" da tabela Aux Cine Brasil, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Realizamos um merge das duas tabelas para ser possível usar o nome dos cursos e não somente o código;
- Optamos por mostrar o resultado de acordo com um *input* de Curso.

Pergunta 5

Para esta pergunta pegamos somente as colunas "TP_SITUACAO", "NU_IDADE", "ID_ALUNO", "IN_INGRESSO_TOTAL" e "CO_CINE_ROTULO" da tabela de aluno, pois somente essas eram necessárias para responder a pergunta.

Observações:

- Desistência: corresponde aos alunos com situação de vínculo igual a “desvinculado do curso” ou “transferido para outro curso da mesma IES”;
- Como estamos tratando apenas o ano de 2019 dos microdados da educação superior não é possível realizar o cálculo da taxa de desistência acumulada, por isso optamos pela de desistência anual;
- Taxa de Desistência Anual (INEP - DEED): percentual do número de estudantes que saíram (desvinculado ou transferido) do curso j no ano t em relação ao número de estudantes ingressantes no curso j do ano T, subtraindo-se o número de estudantes falecidos do curso j no ano t.

Decisões importantes:

- Cálculo da taxa de desistência para cada curso. Será usado para descobrir quais os 3 cursos com maior Tada (Taxa de Desistência Anual);
- Utilizamos o FacetGrid do Seaborn para exibir os resultados.

Pergunta 6

Para esta pergunta pegamos somente as colunas "TP_ESCOLARIDADE" e "TP_CATEGORIA_ADMINISTRATIVA" da tabela de docente, pois somente essas eram necessárias para responder a pergunta.

Observações:

- TP_CATEGORIA_ADMINISTRATIVA 4, 5, 6, 8 e 9 são de IES particulares;
- TP_CATEGORIA_ADMINISTRATIVA 1, 2 e 3 são de IES públicas

Decisões importantes:

- Realizamos um replace dos código para as descrições na coluna “TP_ESCOLARIDADE” para facilitar a visualização e entendimento;
- Utilizamos o histograma da *MatPlotLib* para exibir os resultados.

Pergunta 7

Para esta pergunta pegamos somente as colunas "CO_UF" e "QT_INSCRITO_TOTAL" da tabela de curso, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Realizamos um replace na coluna “CO_UF” para substituir os códigos de UF pelos respectivos nomes dos estados para facilitar visualização e entendimento;
- Utilizamos um plot comum da *Pandas* para exibir os resultados.

Pergunta 8

Para esta pergunta pegamos somente as colunas "ID_ALUNO", "TP_SITUACAO" e "TP_MODALIDADE_ENSINO" da tabela de aluno, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Para responder essa pergunta, realizamos uma contagem dos alunos que repetiam, com a ressalva de que eles deveriam estar ou cursando ou com a matrícula trancada.

Pergunta 9

Para esta pergunta pegamos somente as colunas "CO_CINE_ROTULO" e "NU_IDADE" da tabela de aluno, e as colunas "CO_CINE_ROTULO" e "NO_CINE_ROTULO" da tabela Aux Cine Brasil, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Demos um merge da tabela de alunos com a tabela de cine para pegar o nome dos cursos;

Pergunta 10

Para esta pergunta pegamos somente as colunas "CO_CINE_ROTULO" e "NU_IDADE" da tabela de aluno, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Demos um merge da tabela de alunos com a tabela de cine para pegar o nome dos cursos;

Pergunta 11

Para esta pergunta pegamos somente as colunas "TP_SITUACAO", "IN_INGRESSO_TOTAL", "CO_CINE_ROTULO" e "IN_DEFICIENCIA" da tabela de aluno, e as colunas "CO_CINE_ROTULO", "NO_CINE_AREA_GERAL" e "CO_CINE_AREA_GERAL" da coluna de cine, pois somente essas eram necessárias para responder a pergunta.

Observações:

- Usamos novamente a fórmula para o cálculo da Taxa de Desistência Anual definida na pergunta 5.

Decisões importantes:

- Demos um merge da tabela de alunos com a tabela de cine para pegar o nome dos cursos;

Pergunta 13

Para esta pergunta pegamos somente as colunas "CO_CINE_ROTULO" e "NO_CINE_ROTULO" da tabela de cine, as colunas "ID_ALUNO", "IN_COMPLEMENTAR_EXTENSAO" e "CO_CINE_ROTULO" da tabela de aluno, pois somente essas eram necessárias para responder a pergunta.

Observações:

- Para responder essa pergunta, mostramos qual o curso com maior valor de alunos que participam de projetos de extensão, e também o valor proporcional a quantidade de alunos por curso.

Decisões importantes:

- Foi preciso converter as coluna “CO_CINE_ROTULO” da tabela aluno e “CO_CINE_ROTULO” da tabela Aux Cine Brasil para str para ser possível realizar o merge das duas tabelas;
- O merge das tabelas de aluno e de cine, é feito para adicionar a coluna NO_CINE_ROTULO, que contém o nome dos cursos correspondentes aos seus códigos;
- Utilizamos o *BarPlot* da *SeaBorn* para mostrar os 20 cursos com mais alunos participando de projetos de extensão;

Pergunta 14

Para esta pergunta pegamos somente as colunas "TP_COR_RACA", "IN_INGRESSO_VESTIBULAR", "IN_INGRESSO_ENEM", "IN_INGRESSO_AVALIACAO_SERIADA", "IN_INGRESSO_SELECAO_SIMPLIFICA", "IN_INGRESSO_VAGA_REMANESC", "IN_INGRESSO_VAGA_PROG_ESPECIAL", "IN_INGRESSO_TRANSF_EXOFFICIO", "IN_INGRESSO_DECISAO_JUDICIAL", "IN_INGRESSO_CONVENIO_PECG" e "IN_INGRESSO_EGRESSO" da tabela de aluno, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Utilizamos um *for* para percorrer todos os tipos de ingresso para facilitar a realização desta pergunta;
- Utilizamos o *CountPlot* da *SeaBorn* para exibir os resultados.

Pergunta 15

Para esta pergunta pegamos somente as colunas "TP_COR_RACA" e "TP_CATEGORIA_ADMINISTRATIVA" da tabela de aluno, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Utilizamos do comando *replace* para substituir os valores inteiros da coluna “TP_COR_RACA” para strings que correspondem a raças verdadeiras, para facilitar na visualização dos resultados;
- Filtramos os tipos de categoria administrativa que desejávamos ver, no caso, particulares e públicas;
- Filtramos também por raça, para então apresentar os resultados.

Pergunta 16

Para esta pergunta selecionamos somente a coluna "CO_IES" da tabela de docentes, e as colunas “CO_IES”, “IN_CAPITAL”, “CO_REGIAO” da tabela de IES, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Substituir os valores das colunas “CO_REGIAO” e “IN_CAPITAL”, que eram valores inteiros, por strings com valores respectivos aos inteiros, correspondendo assim à realidade. Exemplo: “CO_REGIAO” assumindo valor 1 quer dizer que o docente leciona no Norte, 2 no Nordeste, e assim por diante.
- Fizemos um *merge* com as duas tabelas para que pudéssemos contar quantos docentes lecionam em dada região, e se ele leciona em uma capital ou no interior.

Pergunta 17

Para esta pergunta pegamos somente as colunas "CO_CINE_ROTULO", "QT_INSCRITO_TOTAL" e "QT_VAGA_TOTAL" da tabela de curso, e as tabelas "CO_CINE_ROTULO" e "NO_CINE_ROTULO" da tabela Aux Cine Brasil, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Realizamos o merge das duas tabelas para que tivéssemos o nome dos cursos e apresentar nos resultados;
- Agrupamos pelo curso para então pegar os dados que nos interessam de cada um.

Pergunta 18

Para esta pergunta pegamos somente as colunas "CO_CINE_ROTULO" e "NO_CINE_AREA_ESPECIFICA" da tabela de Aux Cine Brasil, e as colunas "ID_ALUNO", "IN_DEFICIENCIA" e "CO_CINE_ROTULO" da tabela aluno, pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Filtramos os alunos pela condição de ter uma deficiência;
- Realizamos merge das duas tabelas para termos o nome das subáreas.

Pergunta 19

Para esta pergunta pegamos somente as colunas "QT_CONCLUINTE_TOTAL", "CO_IES", "TP_MODALIDADE_ENSINO", "CO_CINE_ROTULO" e "CO_CURSO" da tabela de curso, e as colunas “CO_IES” e “CO_REGIAO” pois somente essas eram necessárias para responder a pergunta.

Decisões importantes:

- Substituir os valores da coluna “CO_REGIAO”, que eram valores inteiros, por strings com valores respectivos aos inteiros, correspondendo assim à realidade. Exemplo: “CO_REGIAO” assumindo valor 1 quer dizer que o docente leciona no Norte, 2 no Nordeste, e assim por diante;
- Realizamos o merge das tabelas para podermos identificar de qual região pertence a IES;
- Filtramos para apenas cursos presenciais e também para apenas cursos à distância

Pergunta 20

Para esta pergunta pegamos somente as colunas escolhidas pelo usuário da tabela aluno.

Decisões importantes:

- O usuário deverá digitar quais colunas usar, então o nome deve ser idêntico ao da tabela, logo, imprimimos uma lista de colunas que poderiam ser utilizadas;
- Tivemos que utilizar o *drop* das linhas que apresentavam valores iguais a NaN, pois utilizamos o Naive Bayes, e este não aceita valores nulos;
- Escolhemos Naive Bayes pelo fato da nossa base de dados ser muito grande.