Ciência de dados - Dados abertos da UFRN: ITP pré e pós pandemia

Abmael Dantas Gomes, Addan Felipe Neri Andrade
e Jeová Henrique Linhares

recapitulando:

O tema proposto escolhido por nós foi a análise pré e pós pandemia da disciplina de ITP do curso BTI da UFRN.

pergunta:

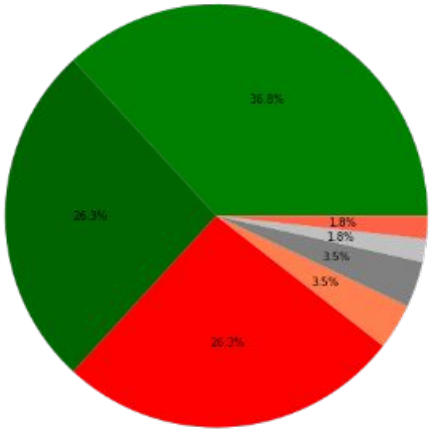
O desempenho dos alunos de ITP foi afetado pela pandemia?

“disclaimer”:

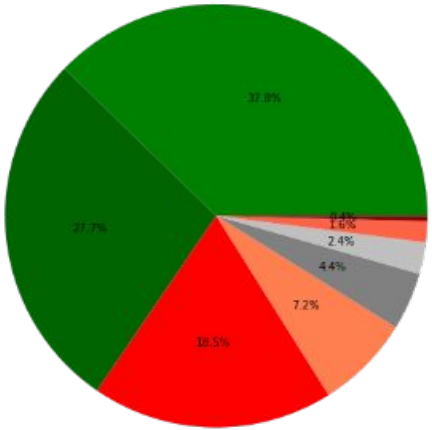
Desde a última apresentação, foi feita a confirmação das bases e dos dados.

Matrículas: 2018.1 e 2018.2

2018.1

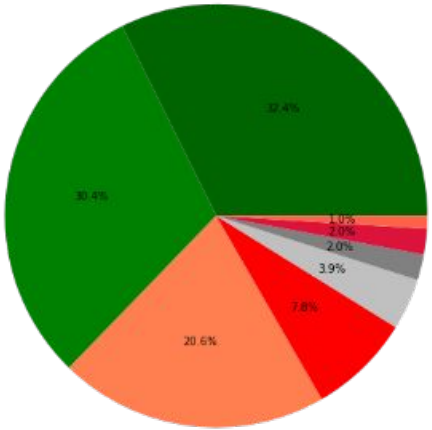


2018.2

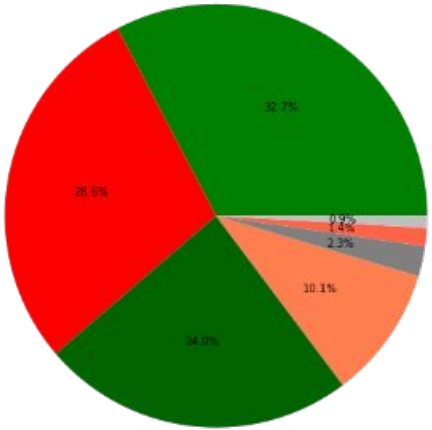


Matrículas: 2019.1 e 2019.2

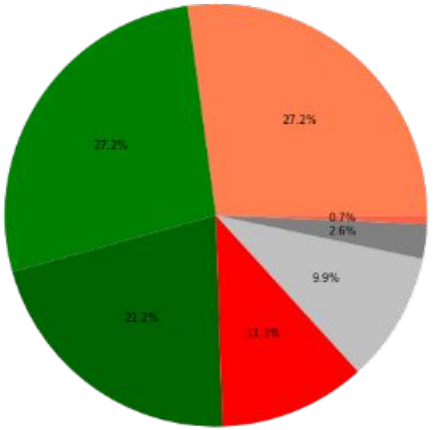
2019.1



2019.2

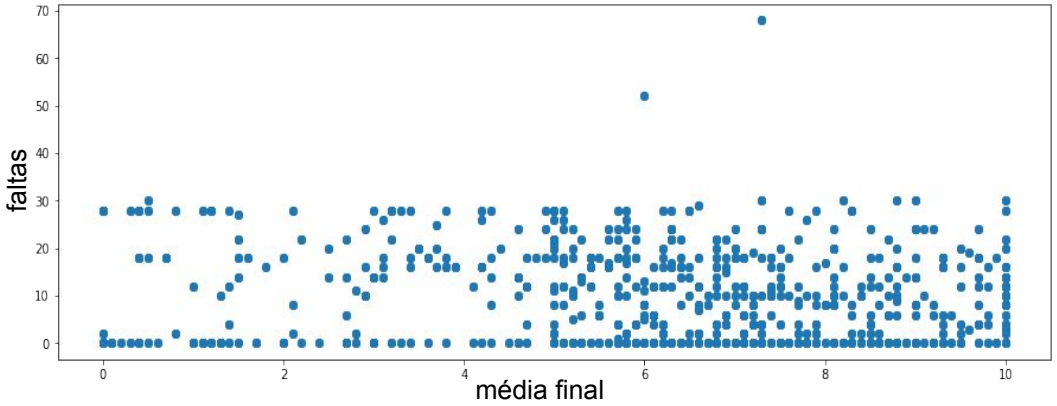


Matrículas: 2020.1

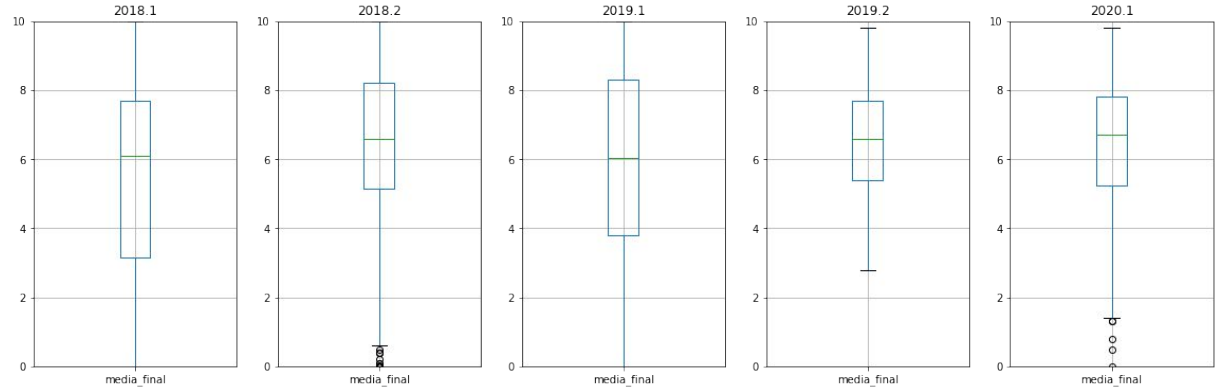


porcentagens relativas às descrições de matrícula, logo, não implica diretamente em resultado de aprovação/reprovação, também, por se tratar de dados que precedem as limpezas e filtragens que fizemos, ou seja, é uma análise bruta das matrículas presentes nos dados abertos

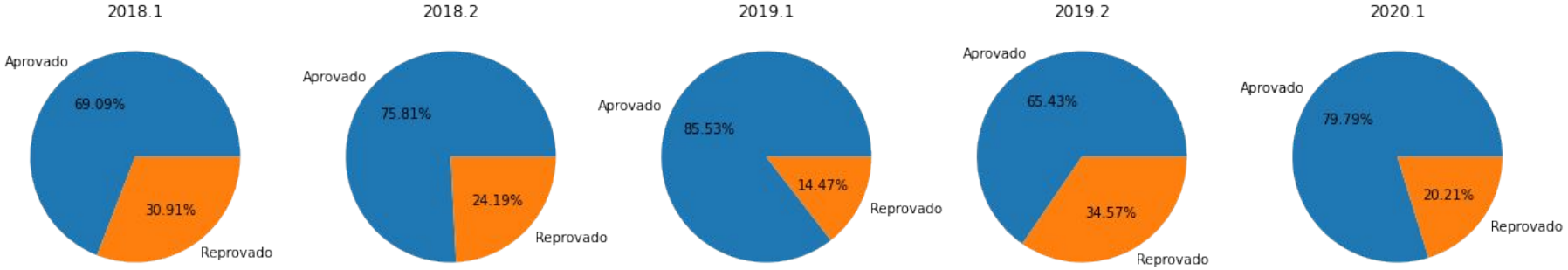
Análise exploratória (zoom se precisar)



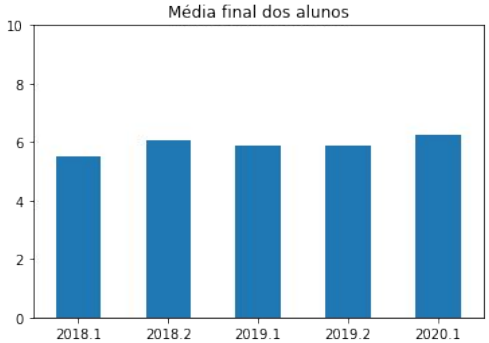
boxplot médias finais:

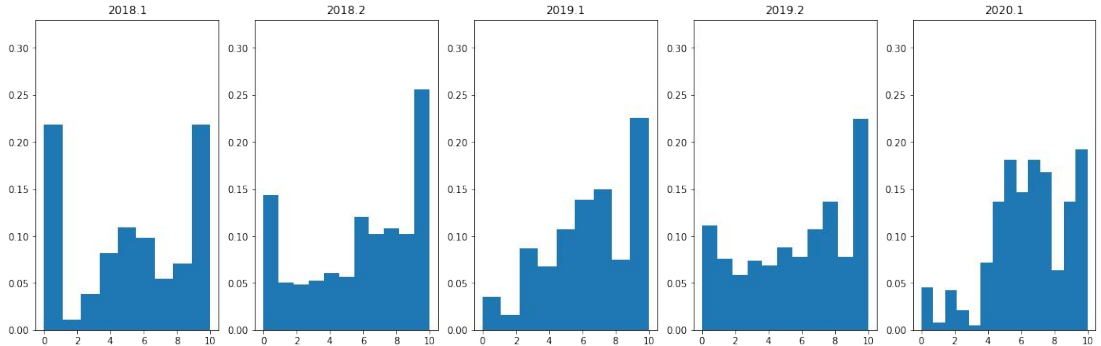


Análise exploratória (zoom se precisar)



aprovação por semestre





distribuição das notas

Atualização da leitura e tipos da base

```
1 base_final.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1965 entries, 152699 to 1699991
Data columns (total 16 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   discente                    1965 non-null   object
 1   unidade                     1965 non-null   uint16
 2   nota                        1965 non-null   float64
 3   media_final                 1965 non-null   float64
 4   aprovado                    1965 non-null   bool
 5   reposicao                   1965 non-null   bool
 6   reprovadoporfalta           1965 non-null   bool
 7   numero_total_faltas         1965 non-null   uint16
 8   descricao                   1965 non-null   object
 9   nivel_ensino                1965 non-null   object
 10  id_turma                    1965 non-null   uint16
 11  codigo                      1965 non-null   object
 12  disciplina                  1965 non-null   object
 13  docente                     1965 non-null   object
 14  semestre                    1965 non-null   object
 15  reprovadomatematicamente    1965 non-null   bool
dtypes: bool(4), float64(2), object(7), uint16(3)
memory usage: 172.7+ KB
```

```
1  # http://dados.ufrn.br/dataset/turmas
2
3  path = '/content/drive/'+"MyDrive"+'/DataScience/Dados/';
4
5  # turmas dos 3 semestres
6  turmas_url_2018_02 = path+'turmas-2018.2.csv'
7  turmas_url_2019_02 = path+'turmas-2019.2.csv'
8  turmas_url_2018_01 = path+'turmas-2018.1.csv'
9  turmas_url_2019_01 = path+'turmas-2019.1.csv'
10 turmas_url_2020_01 = path+'turmas-2020.6.csv'
11
12 # matrículas em componentes dos 3 semestres
13 notas_url_2018_02 = path+'matricula-componente-20182.csv'
14 notas_url_2019_02 = path+'matricula-componente-20192.csv'
15 notas_url_2018_01 = path+'matricula-componente-20181.csv'
16 notas_url_2019_01 = path+'matricula-componente-20191.csv'
17 notas_url_2020_01 = path+'matricula-componente-20206.csv'
18
19 # componentes curriculares presenciais
20 disciplinas_url = path+'componentes-curriculares-presenciais.csv'
21
22 # docentes da ufrn
23 docentes_url = path+'docentes.csv'
```

Recortes:

discente	unidade	nota	media_final	aprovado	reposicao	reprovadoporfalta	numero_total_faltas	descricao	nivel_ensino	id_turma	codigo	disciplina	docente	semestre
6de8a838aa97afbc920c26f719aeab50	1	4.5	7.2	true	false	false	10	APROVADO	GRADUAÇÃO	7902	DIM0118.0	INTRODUÇÃO ÀS TÉCNICAS DE PROGRAMAÇÃO	RAFAEL BESERRA GOMES	2018.1
6de8a838aa97afbc920c26f719aeab50	2	10.0	7.2	true	false	false	10	APROVADO	GRADUAÇÃO	7902	DIM0118.0	INTRODUÇÃO ÀS TÉCNICAS DE PROGRAMAÇÃO	RAFAEL BESERRA GOMES	2018.1
6de8a838aa97afbc920c26f719aeab50	3	7.0	7.2	true	false	false	10	APROVADO	GRADUAÇÃO	7902	DIM0118.0	INTRODUÇÃO ÀS TÉCNICAS DE PROGRAMAÇÃO	RAFAEL BESERRA GOMES	2018.1
d638dffa503c3470874f0e5d8a8c502b	1	4.5	6.4	true	false	false	14	APROVADO POR NOTA	GRADUAÇÃO	7902	DIM0118.0	INTRODUÇÃO ÀS TÉCNICAS DE PROGRAMAÇÃO	RAFAEL BESERRA GOMES	2018.1
d638dffa503c3470874f0e5d8a8c502b	2	7.6	6.4	true	false	false	14	APROVADO POR NOTA	GRADUAÇÃO	7902	DIM0118.0	INTRODUÇÃO ÀS TÉCNICAS DE PROGRAMAÇÃO	RAFAEL BESERRA GOMES	2018.1

existem registros das disciplinas de todos os cursos, na base geral, pois os recortes por curso, disciplina, nível de ensino são feitos depois (para termos o potencial de explorar e analisar outros cursos e/ou disciplinas), mas na base_final existem apenas os registros de ITP

“Agrupamento” e “noções” adicionadas:

```python
def aprovado(row):
    if row["descricao"] == "APROVADO" or row["descricao"] == "APROVADO POR NOTA":
        return True
    else:
        return False
```

```python
def reprovadoporfalta(row):
    if row["descricao"] == "REPROVADO POR FALTAS" or row["descricao"] == "REPROVADO POR NOTA E FALTA" or row["descricao"] == "REPROVADO POR MÉDIA E POR FALTAS":
        return True
    else:
        return False
```

Refinamento e melhora da base:

Outro desafio encontrado durante o desenvolvimento do projeto foi o de tornar a base mais fiel à realidade do curso/disciplina fazendo a remoção exclusivamente de alunos matematicamente incapazes de serem APRN na disciplina de acordo com o regulamento.

Art. 108. O estudante que realiza avaliação de reposição é considerado aprovado, quanto à avaliação de aprendizagem, se satisfaz um dos seguintes critérios:

I – tem média final igual ou superior a 7,0 (sete); ou

II – tem média final igual ou superior a 5,0 (cinco), com rendimento acadêmico igual ou superior a 3,0 (três) na avaliação de reposição.

Parágrafo único. O estudante que realiza avaliação de reposição e não atinge os critérios de aprovação definidos neste artigo é considerado reprovado.

Refinamento e melhora da base:

```
def matematicamentereprovado(row):
  if row["aprovado"] == 0:
    if row["media_final"] < 5:
      if row["reprovadoporfalta"] == True:
        return True;
        # ele tem que sair
      else:
        return False;
    else:
      return False;
  else:
    return False;
```

Após os rotular as entradas na base com a coluna ‘reprovadomatematicamente’, os excluídos ainda podem ser analisados, pois são separados em uma base própria:

```
1 # discentes excluidos pela regra
2 excluidos[excluidos.duplicated(subset = [
```

```
1696
```

Fizemos a verificação se:

- Reprovado
- Média < 5
- Reprovado por falta (vide a definição)
- Então removemos da base_final

Base normalizada (0-1.0):

id_turma	discente	docente	unidade	nota	media_final	numero_total_faltas	aprovado	reposicao	reprovadoporfalta	docente_nome
7902	0.062525	1.0	0.0	0.45	0.72	0.147059	1	0	0	RAFAEL
7902	0.062525	1.0	0.5	1.00	0.72	0.147059	1	0	0	RAFAEL
7902	0.062525	1.0	1.0	0.70	0.72	0.147059	1	0	0	RAFAEL
7902	0.243308	1.0	0.0	0.45	0.64	0.205882	1	0	0	RAFAEL
7902	0.243308	1.0	0.5	0.76	0.64	0.205882	1	0	0	RAFAEL

Como ‘discente’ e ‘docente’ eram campos tipo object (string), para possivelmente usá-los em algum algoritmo de aprendizado, transformamos ambos para o valor original aplicado a um hash int e depois transformamos o hash int para um valor float normalizado (0-1.0).

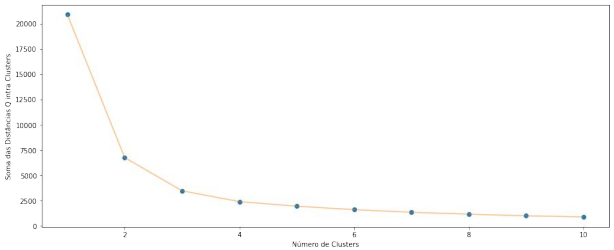
média ou nota 10 = 1.0, média ou nota 5 = 0.5, unidade 1 = 0.0, unidade 2 = 0.5, unidade 3 = 1.0

	nota	media_final	numero_total_faltas	aprovado
count	1965.000000	1965.000000	1965.000000	1965.000000
mean	0.591318	0.608412	0.142232	0.746565
std	0.315708	0.258489	0.143274	0.435089
min	0.000000	0.000000	0.000000	0.000000
25%	0.370000	0.500000	0.000000	0.000000
50%	0.640000	0.640000	0.117647	1.000000
75%	0.860000	0.790000	0.264706	1.000000
max	1.000000	1.000000	1.000000	1.000000

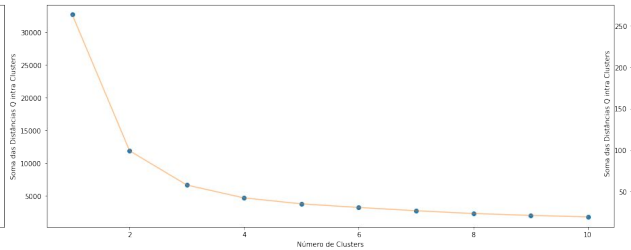
K-means:

O K-means é um algoritmo do tipo não supervisionado, ou seja, que não trabalha com dados rotulados O objetivo desse algoritmo é encontrar similaridades entre os dados e agrupá-los conforme o número de cluster passado pelo argumento K.

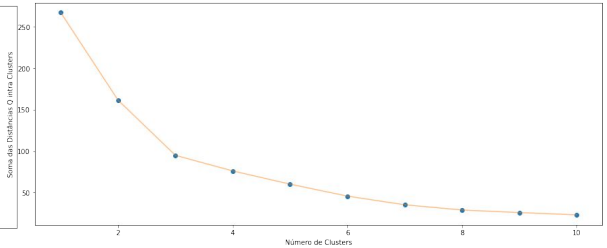
Para definir o número de clusters usamos técnica do “cotovelo”, abaixo, como exemplo, estão alguns gráficos elbow que fizemos:



elbow-nota_x_unidade

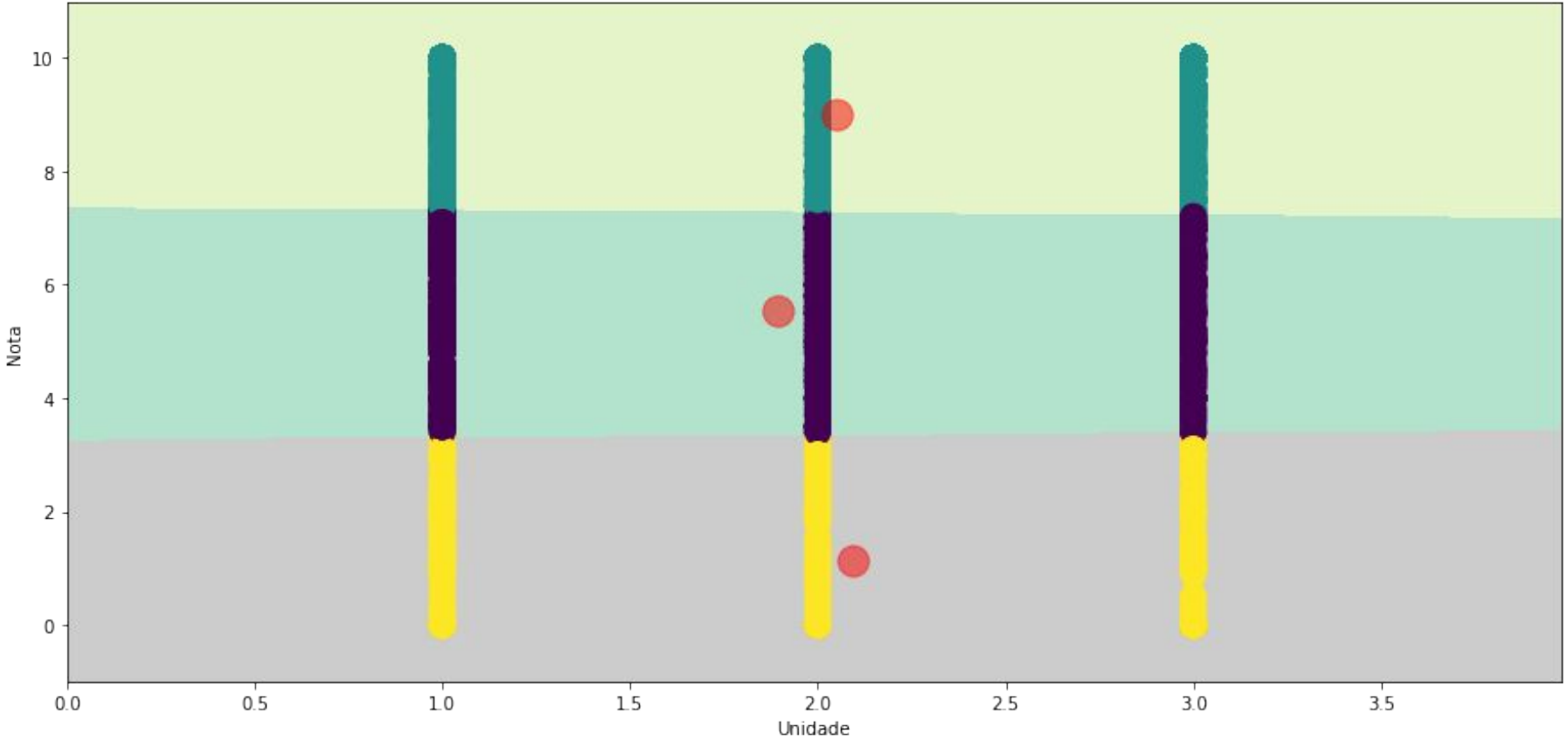


elbow-nota_x_media_final

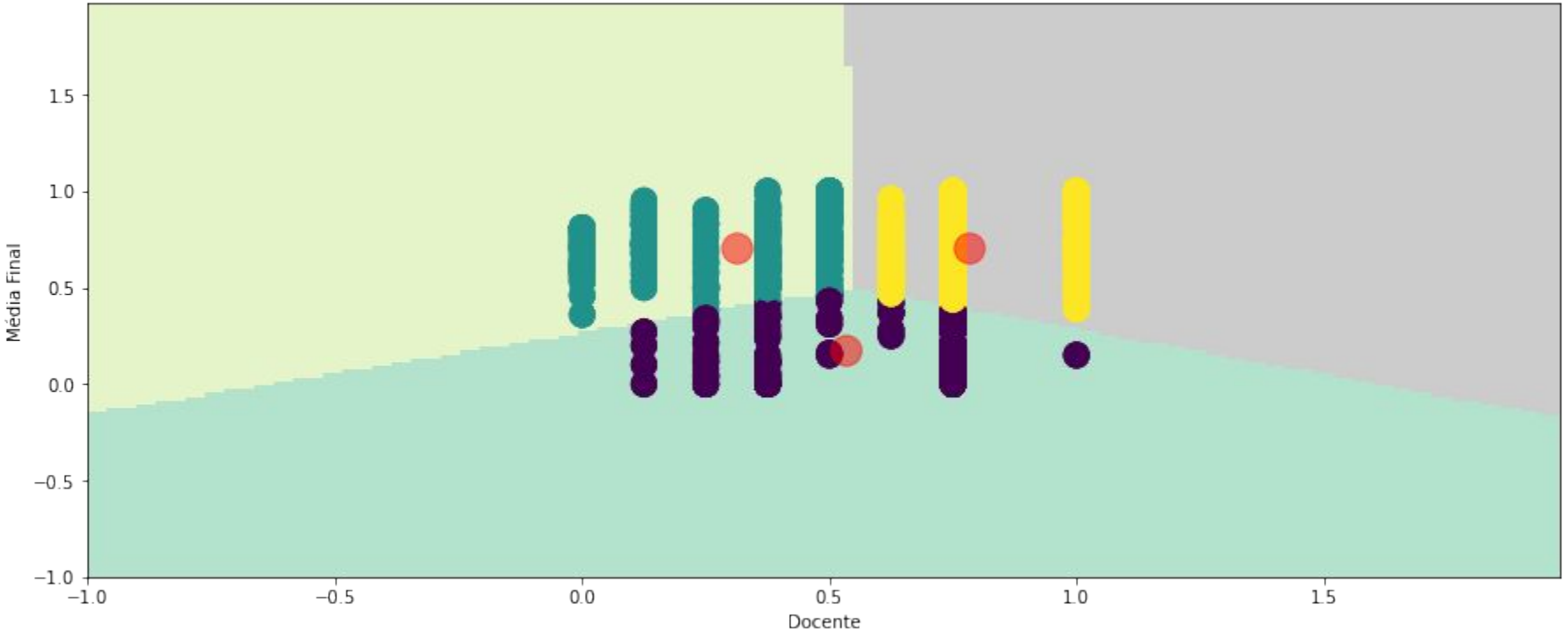


elbow-media_final_normalizada_x_docente

Unidade x nota (geral)

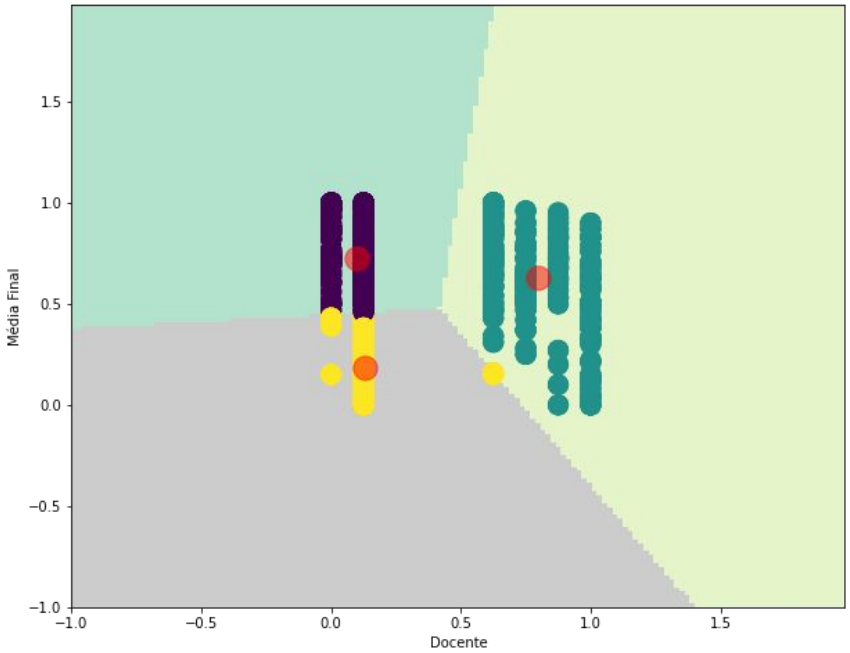


Média final normalizada x docente em float hash (geral)

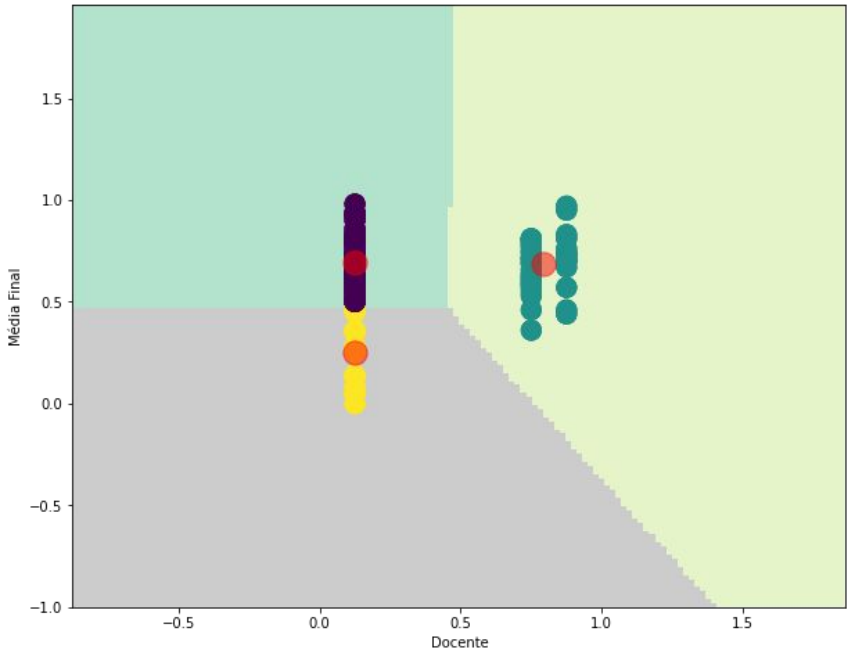


Média final normalizada x docente em float hash

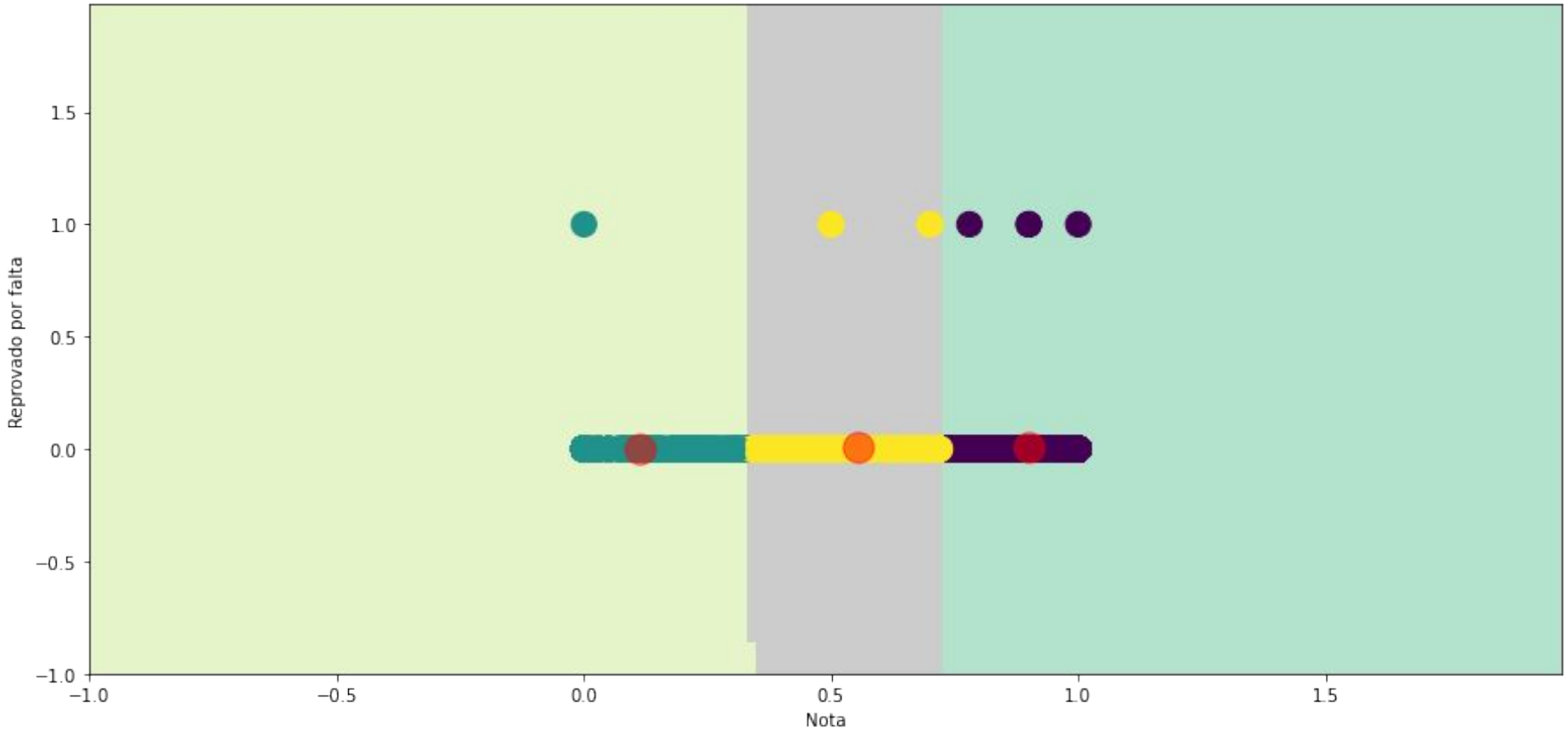
2018.½,2019.½



2020.1

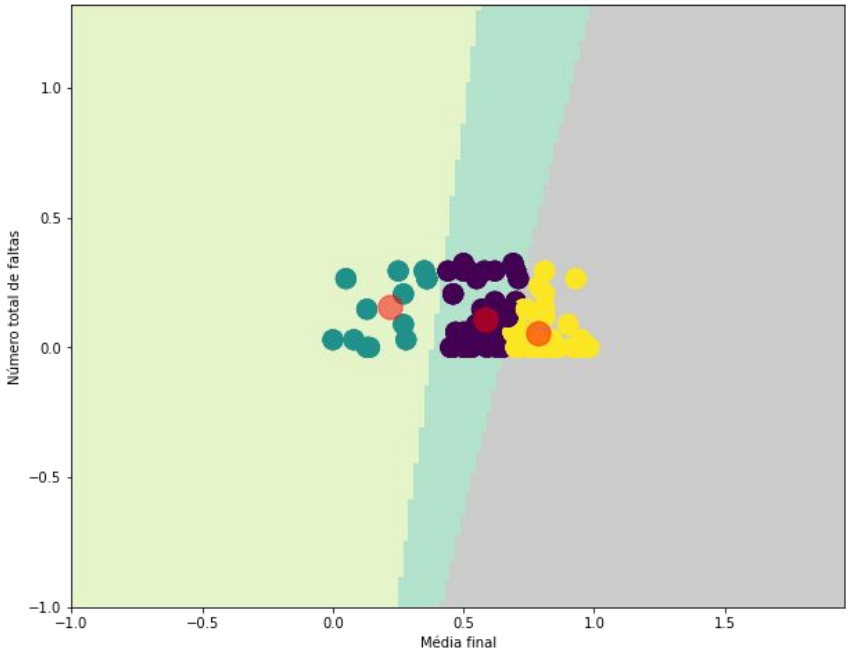


Nota normalizada x reprovado por falta (geral)

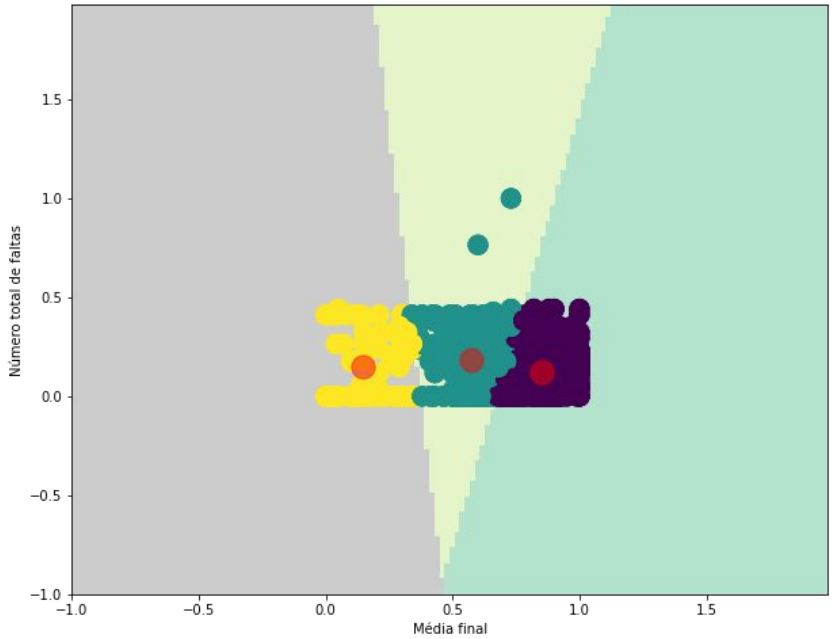


Número total de faltas x média final

2018.½,2019.½



2020.1



Obrigado pela atenção

https://github.com/abmaeld/data-science-itp-analisys