**Relatório de análise de dados**

Foi utilizado a ferramenta RStudio, um software de ambiente de desenvolvimento integrado para R, linguagem de programação multiparadigma, dinâmica, fracamente tipada, voltada a análise, manipulação e visualização de dados. E utilizado script python para fazer toda a análise e clustering dos dados.

O dataset para análise foi montado a partir da base de dados do INEP de micro dados sobre o censo de educação superior (<http://portal.inep.gov.br/web/guest/microdados>).

Passo a passo para definir perfil:

* Montar Dataset (RStudio)
* Definir quantidade de clusters (Script Python)
* Executar classificação a partir da função K-MEANS (Script Python)

**Bibliotecas Utilizadas**

**Bibliotecas em R**

* **data.table**

Instalar biblioteca: install.packages(“data.table“)

Carregar biblioteca: library(data.table)

Descrição: Biblioteca que provê ferramentas de manipulação de dados grandes. Algumas delas são agregação rápida de dados grandes (por exemplo, 100 GB de RAM), junções ordenadas rapidamente, adição / modificação / exclusão rápida de colunas por grupo, sem nenhuma cópia, coluna de listas, leitura / gravação amigável e rápida com valores separados por caracteres. Ainda oferece uma sintaxe natural e flexível, para um desenvolvimento mais rápido.

* **bit64**

Instalar biblioteca: install.packages(“bit64“)

Carregar biblioteca: library(bit64)

Descrição: Biblioteca a qual fornece números inteiros atômicos S3 de 64 bits (assinados). Eles são úteis para manipular chaves do banco de dados e contar exatamente em +- 2 ^ 63. A classe inteiro64 pode ser usada em vetores, matrizes, matrizes e data.frames. Os métodos estão disponíveis para coerção de e para lógicas, números inteiros, duplos,

caracteres e fatores, bem como muitas funções elementares. Muitas operações algorítmicas rápidas, como 'match' e 'order', suportam exploração e manipulação de dados interativas e, opcionalmente, aproveitam o cache.

**Bibliotecas em Python**

* **numpy**

Instalar biblioteca: pip install numpy

Carregar biblioteca: import numpy

Descrição: Utilizada normalmente para realizar cálculos em arrays multidimensionais. O Numpy fornece um grande conjunto de funções e operações de biblioteca que ajudam a executar facilmente cálculos numéricos.

* **pandas**

Instalar biblioteca: pip install pandas

Carregar biblioteca: import pandas

Descrição: Biblioteca usada normalmente para manipulação e análise de dados, além de oferecer estruturas e operações para manipular tabelas numéricas e séries temporais.

* **matplotlib**

Instalar biblioteca: pip install matplotlib

Carregar biblioteca: import matplotlib

Descrição: Uma biblioteca para criação de gráficos e visualização de dados em geral, feita para a linguagem de python.

* **sklearn**

Instalar biblioteca: pip install sklearn

Carregar biblioteca: import sklearn

Descrição: É uma biblioteca de aprendizado de máquina de código aberto para a linguagem python.

**Montar Dataset**

Sequência de comandos em R para gerar dataset a partir dos microdados do INEP:

*> setwd("C:/Users/LUCAS MATHEUS/Desktop/TCC/DADOS")*

*(seta diretório onde estão os microdados do INEP)*

*> library(data.table)*

*(carrega biblioteca de manipulação de grandes dados “data.table”)*

*> alunos2018 <- fread("DADOS\_ALUNOS\_2018.csv", select = c("TP\_SITUACAO","CO\_IES","TP\_CATEGORIA\_ADMINISTRATIVA","TP\_ORGANIZACAO\_ACADEMICA","TP\_TURNO",*

*"TP\_GRAU\_ACADEMICO","TP\_COR\_RACA","TP\_SEXO","NU\_IDADE","IN\_INGRESSO\_VESTIBULAR","IN\_INGRESSO\_ENEM","IN\_FINANCIAMENTO\_ESTUDANTIL","TP\_ESCOLA\_CONCLUSAO\_ENS\_MEDIO"))*

*(carrega em “alunos2018” os microdados do INEP de 2018 com esses respectivas colunas)*

*> names(alunos2018)[1:13] <- c("ALUNO\_SITUACAO","ALUNO\_INEP\_COD","INSTITUICAO\_CATEGORIA\_ADMINISTRATIVA","INSTITUICAO\_ORGANIZACAO\_ACADEMICA","CURSO\_TURNO","ALUNO\_GRAU\_ACADEMICO","ALUNO\_COR\_RACA","ALUNO\_SEXO","ALUNO\_IDADE","ALUNO\_INGRESSO\_VESTIBULAR","ALUNO\_INGRESSO\_ENEM","ALUNO\_FINANCIAMENTO\_ESTUDANTIL","ALUNO\_ESCOLA\_CONCLUSAO\_ENS\_MEDIO")*

*(Mudar todos os nome das colunas para a nomenclatura padrão)*

*> alunos2017 <- fread("DADOS\_ALUNOS\_2017.csv", select = c("TP\_SITUACAO","CO\_IES","TP\_CATEGORIA\_ADMINISTRATIVA","TP\_ORGANIZACAO\_ACADEMICA","TP\_TURNO","TP\_GRAU\_ACADEMICO","TP\_COR\_RACA","TP\_SEXO","NU\_IDADE","IN\_INGRESSO\_VESTIBULAR","IN\_INGRESSO\_ENEM","IN\_FINANCIAMENTO\_ESTUDANTIL","TP\_ESCOLA\_CONCLUSAO\_ENS\_MEDIO"))*

*(carrega em “alunos2017” os microdados do INEP de 2017 com esses respectivas colunas)*

*> names(alunos2017)[1:13] <- c("ALUNO\_SITUACAO","ALUNO\_INEP\_COD","INSTITUICAO\_CATEGORIA\_ADMINISTRATIVA","INSTITUICAO\_ORGANIZACAO\_ACADEMICA","CURSO\_TURNO","ALUNO\_GRAU\_ACADEMICO","ALUNO\_COR\_RACA","ALUNO\_SEXO","ALUNO\_IDADE","ALUNO\_INGRESSO\_VESTIBULAR","ALUNO\_INGRESSO\_ENEM","ALUNO\_FINANCIAMENTO\_ESTUDANTIL","ALUNO\_ESCOLA\_CONCLUSAO\_ENS\_MEDIO")*

*(Mudar todos os nome das colunas para a nomenclatura padrão)*

*> alunos <- rbind(alunos2018, alunos2017)*

*(Junta os dados contidos em “alunos2018” com “alunos2017” e coloca eles em “alunos”)*

*> rm(alunos2017)*

*(Remove o data frame “alunos2017” já que este não será mais necessário)*

*> rm(alunos2018)*

*(Remove o data frame “alunos2018” já que este não será mais necessário)*

*> alunos2016 <- fread("DADOS\_ALUNOS\_2016.csv", select = c("CO\_ALUNO\_SITUACAO","CO\_IES","CO\_CATEGORIA\_ADMINISTRATIVA","CO\_ORGANIZACAO\_ACADEMICA","CO\_TURNO\_ALUNO","CO\_GRAU\_ACADEMICO","CO\_COR\_RACA\_ALUNO","IN\_SEXO\_ALUNO","NU\_IDADE\_ALUNO","IN\_ING\_VESTIBULAR","IN\_ING\_ENEM","IN\_FINANC\_ESTUDANTIL","CO\_TIPO\_ESCOLA\_ENS\_MEDIO"))*

*(carrega em “alunos2016” os microdados do INEP de 2016 com esses respectivas colunas)*

*> names(alunos2016)[1:13] <- c("ALUNO\_SITUACAO","ALUNO\_INEP\_COD","INSTITUICAO\_CATEGORIA\_ADMINISTRATIVA","INSTITUICAO\_ORGANIZACAO\_ACADEMICA","CURSO\_TURNO","ALUNO\_GRAU\_ACADEMICO","ALUNO\_COR\_RACA","ALUNO\_SEXO","ALUNO\_IDADE","ALUNO\_INGRESSO\_VESTIBULAR","ALUNO\_INGRESSO\_ENEM","ALUNO\_FINANCIAMENTO\_ESTUDANTIL","ALUNO\_ESCOLA\_CONCLUSAO\_ENS\_MEDIO")*

*(Mudar todos os nome das colunas para a nomenclatura padrão)*

*> alunos <- rbind(alunos, alunos2016)*

*(Junta os dados contidos em “alunos” com “alunos2016” e coloca eles em “alunos”)*

*> rm(alunos2016)*

*(Remove o data frame “alunos2016” já que este não será mais necessário)*

*> alunos2015 <- fread("DADOS\_ALUNOS\_2015.csv", select = c("CO\_ALUNO\_SITUACAO","CO\_IES","CO\_CATEGORIA\_ADMINISTRATIVA","CO\_ORGANIZACAO\_ACADEMICA","CO\_TURNO\_ALUNO","CO\_GRAU\_ACADEMICO","CO\_COR\_RACA\_ALUNO","IN\_SEXO\_ALUNO","NU\_IDADE\_ALUNO","IN\_ING\_VESTIBULAR","IN\_ING\_ENEM","IN\_FINANC\_ESTUDANTIL","CO\_TIPO\_ESCOLA\_ENS\_MEDIO"))*

*(carrega em “alunos2015” os microdados do INEP de 2015 com esses respectivas colunas)*

*> names(alunos2015)[1:13] <- c("ALUNO\_SITUACAO","ALUNO\_INEP\_COD","INSTITUICAO\_CATEGORIA\_ADMINISTRATIVA","INSTITUICAO\_ORGANIZACAO\_ACADEMICA","CURSO\_TURNO","ALUNO\_GRAU\_ACADEMICO","ALUNO\_COR\_RACA","ALUNO\_SEXO","ALUNO\_IDADE","ALUNO\_INGRESSO\_VESTIBULAR","ALUNO\_INGRESSO\_ENEM","ALUNO\_FINANCIAMENTO\_ESTUDANTIL","ALUNO\_ESCOLA\_CONCLUSAO\_ENS\_MEDIO")*

*(Mudar todos os nome das colunas para a nomenclatura padrão)*

*> alunos <- rbind(alunos, alunos2015)*

*(Junta os dados contidos em “alunos” com “alunos2015” e coloca eles em “alunos”)*

*> rm(alunos2015)*

*(Remove o data frame “alunos2015” já que este não será mais necessário)*

*> alunos2014 <- fread("DADOS\_ALUNOS\_2014.csv", select = c("CO\_ALUNO\_SITUACAO","CO\_IES","CO\_CATEGORIA\_ADMINISTRATIVA","CO\_ORGANIZACAO\_ACADEMICA","CO\_TURNO\_ALUNO","CO\_GRAU\_ACADEMICO","CO\_COR\_RACA\_ALUNO","IN\_SEXO\_ALUNO","NU\_IDADE\_ALUNO","IN\_ING\_VESTIBULAR","IN\_ING\_ENEM","IN\_FINANC\_ESTUDANTIL","CO\_TIPO\_ESCOLA\_ENS\_MEDIO"))*

*(carrega em “alunos2014” os microdados do INEP de 2014 com esses respectivas colunas)*

*> names(alunos2014)[1:13] <- c("ALUNO\_SITUACAO","ALUNO\_INEP\_COD","INSTITUICAO\_CATEGORIA\_ADMINISTRATIVA","INSTITUICAO\_ORGANIZACAO\_ACADEMICA","CURSO\_TURNO","ALUNO\_GRAU\_ACADEMICO","ALUNO\_COR\_RACA","ALUNO\_SEXO","ALUNO\_IDADE","ALUNO\_INGRESSO\_VESTIBULAR","ALUNO\_INGRESSO\_ENEM","ALUNO\_FINANCIAMENTO\_ESTUDANTIL","ALUNO\_ESCOLA\_CONCLUSAO\_ENS\_MEDIO")*

*(Mudar todos os nome das colunas para a nomenclatura padrão)*

*> alunos <- rbind(alunos, alunos2014)*

*(Junta os dados contidos em “alunos” com “alunos2014” e coloca eles em “alunos”)*

*> rm(alunos2014)*

*(Remove o data frame “alunos2014” já que este não será mais necessário)*

*> saveRDS(alunos, file = "DatasetAlunos")*

*(Salvar o dateset gerado, o qual contém todos os dados de 2014 a 2018 como “DatasetAlunos”)*

Ao final do processo é gerado um dataset com todos os dados de 2014 a 2018 de microdados do censo da educação superior do INEP. Com isso é dada a tabela com as respectivas colunas:

|  |  |  |  |
| --- | --- | --- | --- |
| Colunas | Descrição das Variáveis | Tipo | Descrição das Categorias |
| ALUNO\_SITUACAO | Tipo de situação de vínculo do aluno no curso | Num | 2.Cursando  3. Matrícula trancada  4. Desvinculado do curso  5. Transferido para outro curso da mesma IES  6. Formado 7. Falecido |
| ALUNO\_INEP\_COD | Código único de identificação da IES | Num | - |
| INSTITUICAO\_CATEGORIA\_ADMINISTRATIVA | Tipo da Categoria Administrativa da IES | Num | 1. Pública Federal  2. Pública Estadual  3. Pública Municipal  4. Privada com fins lucrativos  5. Privada sem fins lucrativos  6. Privada - Particular em sentido estrito  7. Especial  8. Privada comunitária  9. Privada confessional |
| INSTITUICAO\_ORGANIZACAO\_ACADEMICA | Tipo da Organização Acadêmica da IES | Num | 1. Universidade  2. Centro Universitário  3. Faculdade  4. Instituto Federal de Educação, Ciência e Tecnologia  5. Centro Federal de Educação Tecnológica |
| CURSO\_TURNO | Tipo do turno do curso ao qual o aluno está vinculado | Num | 1. Matutino  2. Vespertino  3. Noturno  4. Integral  0. Não identificado  (.) Não aplicável (Ead) |
| ALUNO\_GRAU\_ACADEMICO | Tipo do grau acadêmico conferido ao diplomado pelo curso | Num | 1. Bacharelado  2. Licenciatura  3. Tecnológico  4. Bacharelado e Licenciatura  (.) Não aplicável (cursos com nivel acadêmico igual a sequencial de formação específica ou cursos com área básica de curso) |
| ALUNO\_COR\_RACA | Tipo da cor/raça do aluno | Num | 0. Aluno não quis declarar cor/raça  1. Branca  2. Preta  3. Parda  4. Amarela  5. Indígena  9. Não dispõe da informação (Não resposta) |
| ALUNO\_SEXO | Informa o sexo do aluno | Num | 1. Feminino  2. Masculino |
| ALUNO\_IDADE | Idade que o aluno completa no ano de referência do Censo | Num | - |
| ALUNO\_INGRESSO\_VESTIBULAR | Informa se o aluno ingressou no curso por vestibular. | Num | 0. Não  1. Sim |
| ALUNO\_INGRESSO\_ENEM | Informa se o aluno ingressou no curso pelo Enem. | Num | 0. Não  1. Sim |
| ALUNO\_FINANCIAMENTO\_ESTUDANTIL | Informa se o aluno utiliza financiamento estudantil | Num | 0. Não  1. Sim |
| ALUNO\_ESCOLA\_CONCLUSAO\_ENS\_MEDIO | Tipo de escola que o aluno concluiu ensino médio | Num | 1. Pública  2. Privada  9. Não dispõe da informação (Não resposta) |

**Seleção de dataset dos alunos**

Seleção dos dataset em R dos alunos que evadem instituições de ensino superior

*> setwd("C:/Users/LUCAS MATHEUS/Desktop/TCC/DADOS")*

*(seta diretório onde está o dataset “DatasetAlunos”)*

*> DatasetAlunos <- readRDS(“DatasetAlunos”)*

*(Carrega dataset no data frame “DatasetAlunos”)*

*> DatasetAlunosEvasao <- subset(DatasetAlunos, ALUNO\_SITUACAO == 3 | ALUNO\_SITUACAO == 4, select=c(ALUNO\_SITUACAO,ALUNO\_INEP\_COD,INSTITUICAO\_CATEGORIA\_ADMINISTRATIVA,INSTITUICAO\_ORGANIZACAO\_ACADEMICA,CURSO\_TURNO,ALUNO\_GRAU\_ACADEMICO,ALUNO\_COR\_RACA,ALUNO\_SEXO,ALUNO\_IDADE,ALUNO\_INGRESSO\_VESTIBULAR,ALUNO\_INGRESSO\_ENEM,ALUNO\_FINANCIAMENTO\_ESTUDANTIL,ALUNO\_ESCOLA\_CONCLUSAO\_ENS\_MEDIO))*

*(Seleciona no data frame “DatasetAlunosEvasao” os alunos do data frame “DatasetAlunos” que estão com matrícula trancada ou desvinculados no curso)*

*> DatasetAlunosEvasao[is.na(DatasetAlunosEvasao)] <- 0*

*(Faz um varedura no dataset “DatasetAlunosEvasao” substituindo valores nulos por 0)*

*> saveRDS(DatasetAlunosEvasao, file = "DatasetAlunosEvasao")*

*(Salvar o dateset gerado como “DatasetAlunosEvasao”, o qual contém todos os alunos que evadiram o ensino superior)*

*> write.csv(DatasetAlunos, "DatasetAlunosEvasao.csv", row.names = FALSE)*

*(Salva o dataset gerado como “DatasetAlunosEvasao.csv”, o qual contém todos os alunos que evadiram o ensino superior)*

**Calculo de numero de clusters ideal e execução da análise Kmeans**

Calcular número ideal de clusters através de um script python, utilizando a técnica do cotovelo

import numpy as np

import pandas as pd

from matplotlib import pyplot as plt

from sklearn.cluster import KMeans

Data=pd.read\_csv('C:/Users/LUCASMATHEUS/Desktop/TCC/DADOS/DatasetAlunosEvasao.csv')

X = np.array(Data)

wcss = []

for i in range(1, 11):

kmeans = KMeans(*n\_clusters*=i, *init*='k-means++', *random\_state*=0)

kmeans.fit(X)

wcss.append(kmeans.inertia\_)

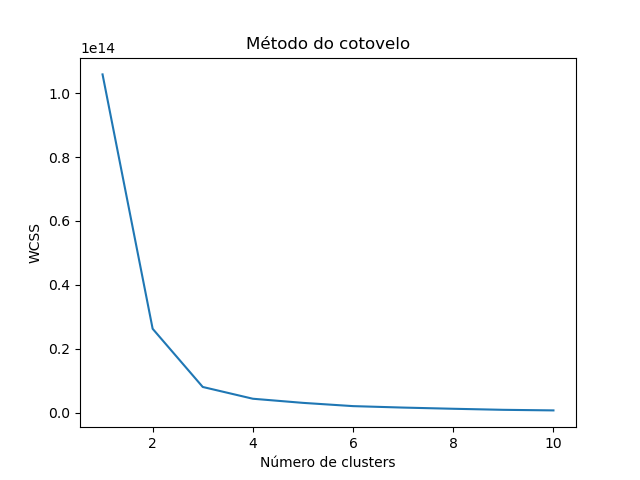
plt.plot(range(1, 11), wcss)

plt.title('Método do cotovelo')

plt.xlabel('Número de clusters')

plt.ylabel('WCSS')

plt.show()



**Executar função Kmeans e exibir gráficos de visualização**

É executado através de um script python a função kmeans no dataset e exibido em gráficos o resultado do clustering

import numpy as np

import pandas as pd

import seaborn as sb

from matplotlib import pyplot as plt

from sklearn.cluster import KMeans

Data = pd.read\_csv('C:/Users/LUCAS MATHEUS/Desktop/TCC/DADOS/DatasetAlunosEvasao.csv')

X = np.array(Data)

kmeans = KMeans(*n\_clusters*=4, *init*='k-means++', *random\_state*=0)

labels = kmeans.fit\_predict(X)

Data['K\_CLASSES'] = kmeans.labels\_

plt.scatter(X[labels==0, 0], X[labels==0, 1], *s*=100, *c*='yellow', *label* ='Cluster 1')

plt.scatter(X[labels==1, 0], X[labels==1, 1], *s*=100, *c*='blue', *label* ='Cluster 2')

plt.scatter(X[labels==2, 0], X[labels==2, 1], *s*=100, *c*='green', *label* ='Cluster 3')

plt.scatter(X[labels==3, 0], X[labels==3, 1], *s*=100, *c*='cyan', *label* ='Cluster 4')

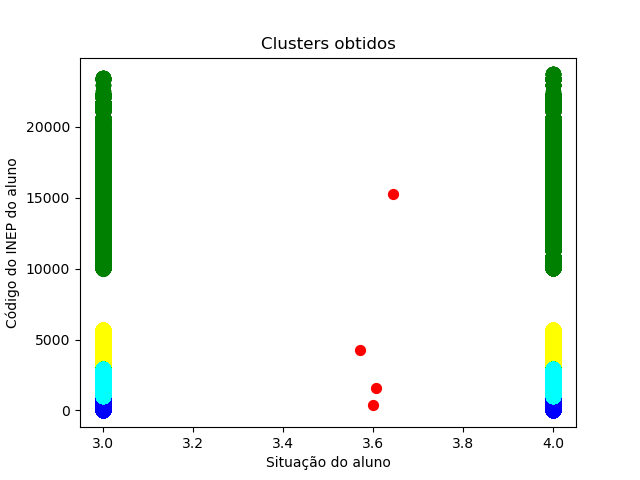
plt.scatter(kmeans.cluster\_centers\_[:, 0], kmeans.cluster\_centers\_[:, 1], *s*=50, *c*='red',*label*= 'Centroids')

plt.title('Clusters obtidos')

plt.xlabel('Situação do aluno')

plt.ylabel('Código do INEP do aluno')

plt.show()



import numpy as np

import pandas as pd

import seaborn as sb

from matplotlib import pyplot as plt

from sklearn.cluster import KMeans

Data = pd.read\_csv('C:/Users/LUCAS MATHEUS/Desktop/TCC/DADOS/DatasetAlunosEvasao.csv')

X = np.array(Data)

kmeans = KMeans(*n\_clusters*=4, *init*='k-means++', *random\_state*=0)

labels = kmeans.fit\_predict(X)

Data['K\_CLASSES'] = kmeans.labels\_

plt.scatter(X[labels==0, 0], X[labels==0, 1], *s*=100, *c*='yellow', *label* ='Cluster 1')

plt.scatter(X[labels==1, 0], X[labels==1, 1], *s*=100, *c*='blue', *label* ='Cluster 2')

plt.scatter(X[labels==2, 0], X[labels==2, 1], *s*=100, *c*='green', *label* ='Cluster 3')

plt.scatter(X[labels==3, 0], X[labels==3, 1], *s*=100, *c*='cyan', *label* ='Cluster 4')

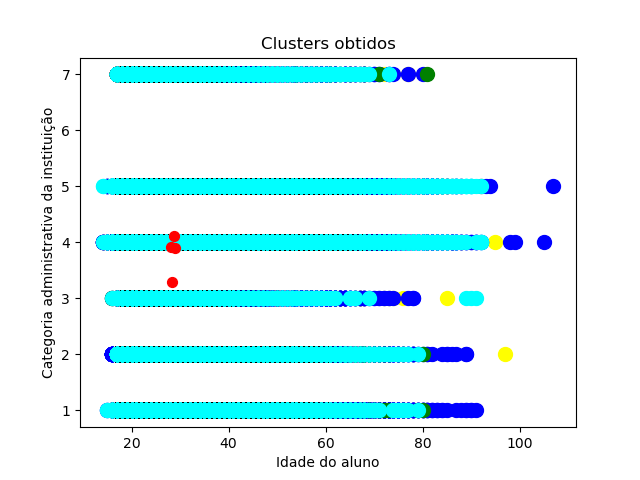
plt.scatter(kmeans.cluster\_centers\_[:, 0], kmeans.cluster\_centers\_[:, 1], *s*=50, *c*='red',*label*= 'Centroids')

plt.xlabel('Idade do aluno')

plt.ylabel('Categoria administrativa da instituição')

plt.title('Clusters obtidos')

plt.show()



**Resultado da análise de perfil**

O cluster que mais se destacou com cerca de 61% (cluster com 9682834 amostras das 15860343 mostras do dataset de alunos) do total do dataset de alunos, tem as características que descrevem o perfil da maioria dos de alunos que evadiram de instituições de ensino superior.

Segue a análise detalhada:

|  |  |  |
| --- | --- | --- |
| Descrição do dado | Descrição e quantidade de amostras | Porcentagem do valor de maior concentração de amostras (valor percentual aproximado) |
| Tipo de situação de vínculo do aluno no curso | Matrícula trancada - 3885412  Desvinculados do curso - 5797422 | Desvinculados do curso - 60% |
| Tipo da Categoria Administrativa da IES | Pública Federal - 1275932  Pública Estadual - 435486  Pública Municipal - 55625  Privada com fins lucrativos - 4177849  Privada - 3688882  Especial - 49060 | Privada com fins lucrativos - 43% |
| Tipo da Organização Acadêmica da IES | Universidade - 7919722  Centro universitario - 1288037  Faculdade - 439539  Instituto Federal de educação, ciência e tecnologia - 24830  Centro federal de educação e Tecnologia - 10706 | Universidade - 81% |
| Tipo do turno do curso ao qual o aluno está vinculado | Informação indisponível - 3186248  Matutino - 1427029  Vespertino - 219870  Noturno - 4115474  Integral - 734213 | Noturno - 42% |
| Tipo do grau acadêmico conferido ao diplomado pelo curso | Informação indisponível - 61248  Bacharelado - 5863055  Licenciatura - 1983134  Tecnologia - 1775397 | Bacharelado - 60% |
| Tipo da cor/raça do aluno | Aluno não quis identificar - 3510867  Branca - 3178535  Preta - 504434  Parda -1900560  Amarela - 119385  Indígena - 42108  Informação Indisponível - 33035 | Branca - 32% |
| Informa o sexo do aluno | Informação Indisponível - 2674556  Feminino - 5099365  Masculino - 1908913 | Feminino - 52% |
| Idade que o aluno completa no ano de referência do Censo | Entre 18 e 42 anos acima de 10 mil amostras por ano  Entre 19 e 28 anos acima de 40 mil amostras por ano | 21 anos - 652673 amostras - 6% |
| Informa se o aluno ingressou no curso por vestibular | Não - 2652414  Sim - 7030420 | Sim - 72% |
| Informa se o aluno ingressou no curso pelo Enem | Não - 8173051  Sim - 1509783 | Não - 82% |
| Informa se o aluno utiliza financiamento estudantil | Não - 8173051  Sim - 1509783 | Não - 84% |
| Tipo de escola que o aluno concluiu ensino médio | Informação indisponível - 1650165  Pública - 6343928  Privada - 1688741 | Pública - 65% |

A partir da tabela acima é possível definir o perfil da maioria dos alunos que evadem de cursos de ensino superior, sendo eles de maioria desvinculados do curso, de sexo feminino, idade entre 19 e 28 anos, identificadas como brancas, de universidades privadas com fins lucrativos, ingressando em cursos de bacharelado de turno noturno, em sua maioria por meio do enem, sem financiamento estudantil e concluintes do ensino médio através de instituições públicas.