

Programa de Pós-graduação em Sistemas de Informação

SIN5007 - Reconhecimento de Padrões (2023)

Censo da Educação Superior - Cursos 2021

MSc. Leonardo Cunha dos Santos Gabriel Francisco dos Santos Silva

São Paulo / 2023

Agenda



- 1 Descrição do dataset e análise exploratória
- Pré-processamento e PCA
- 3 Seleção de características
- Maive Bayes Classifier
- **5** Estimação de desempenho



Descrição do dataset e análise exploratória

Introdução







Introdução

Microdados



- Os microdados do Inep reúnem informações detalhadas sobre pesquisas do INEP;
- As estatísticas produzidas pelo Inep visam fornecer os subsídios para a formulação e implementação de políticas voltadas para a melhoria contínua da educação no país;
- Os formatos de apresentação foram reestruturados de acordo com a Lei Geral de Proteção de Dados Pessoais (LGPD).

https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior

Conjunto de dados

Microdados do Censo da Educação Superior



Cadastro de IES

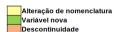
- Localização
- · Quadro de funcionários
- Infraestrutura
- Estatísticas sobre professores

Cadastro de Cursos

- Grau acadêmico
- Modalidade
- Estatísticas sobre estudantes

Dicionário de dados





Cadastro IES

N	Nome da Variável	Descrição da Variável	Tipo	Tam.	Categoria				
1	NU_ANO_CENSO	Ano de referência do Censo da Educação Superior	Num	4					
	DADOS DA INSTITUIÇÃO DE ENSINO SUPERIOR (IES) - SEDE ADMINISTRATIVA/REITORIA								
2	NO_REGIAO_IES	Nome da região geográfica da sede administrativa ou reitoria da IES	Char	20					
3	CO_REGIAO_IES	Código da região geográfica da sede administrativa ou reitoria da IES	Num	2					
4	NO_UF_IES	Nome da Unidade da Federação da sede administrativa ou reitoria da IES	Char	50					
5	SG_UF_IES	Sigla da Unidade da Federação da sede administrativa ou reitoria da IES	Char	2					
6	CO_UF_IES	Código da Unidade da Federação da sede administrativa ou reitoria da IES	Num	2					
7	NO_MUNICIPIO_IES	Nome do Município da sede administrativa ou reitoria da IES	Char	150					
8	CO_MUNICIPIO_IES	Código do Município da sede administrativa ou reitoria da IES	Num	7					
9	IN_CAPITAL_IES	Informa se a sede administrativa ou reitoria da IES está localizada na capital da Unidade da Federação	Num	2	0. Não 1. Sim				
10	NO_MESORREGIAO_IES	Nome da Mesorregião da sede administrativa ou reitoria da IES	Char	100					
11	CO_MESORREGIAO_IES	Código da Mesorregião da sede administrativa ou reitoria da IES	Num	4					
12	NO_MICRORREGIAO_IES	Nome da Microrregião da sede administrativa ou reitoria da IES	Char	100					
13	CO_MICRORREGIAO_IES	Código da Microrregião da sede administrativa ou reitoria da IES	Num	5					
14	TP_ORGANIZACAO_ACADEMICA	Tipo de Organização Acadêmica da IES	Num	1	Universidade Centro Universitário Faculdade Instituto Federal de Educação, Ciência e Tecnologia Centro Federal de Educação Tecnológica				

Microdados 2021 - Variáveis



Cadastro de IES

• Categóricas: 32 + 1 (ID:ANO)

• Numéricas: 48

• Total: 81

Cadastro de CURSOS

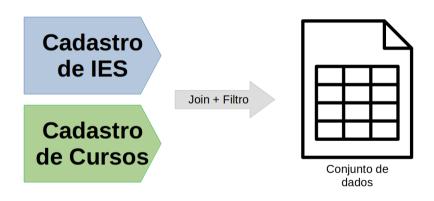
• Categóricas: 27 + 1 (ID:ANO)

• Numéricas: 172

• Total: 200

Cursos de tecnologia





Variáveis numéricas: $05 \mid Variáveis categóricas: 22 + 1 (ID:ANO)$

34 cursos | Instâncias: 19158 | 558 cursos distintos

Depara de cursos



NO CURSO DEPARA	NO CURSO
Agrocomputação	Agrocomputação
Análise e Desenvolvimento de Sistemas	Administração Em Sistemas E Serviços De Saúde
	Análise De Infraestrutura De Redes E Sistemas Computacionais
	Análise De Sistemas
	Análise E Desenvolvimento De Sistemas
	Desenvolvimento De Sistemas
	Sistemas Para Internet
Ciências da Computação	Abi - Ciência Da Computação
olenolas da Compatação	Ciência Da Computação
	Ciências Da Computação
	Ciências De Computação
	Computação
	Computação E Informática
	Computação E Robótica Educativa
	Computação Em Nuvem
	Computação Gráfica
	Internet Das Coisas E Computação Em Nuvem
Engenharia da Computação	Engenharia Da Computação
Engennana da compatação	Engenharia De Computação
	Engenharia De Computação - Énfase Sistemas Corporativos
	Engenharia De Computação E Informação
Engenharia de Sistemas	Engenharia De Automação E Sistemas
Engelinara de disternas	Engenharia De Produção E Sistemas
	Engenharia De Sistemas
	Engenharia De Sistemas Ciber Físicos
Engenharia Elétrica - Énfase Em Computação	Engenharia Elétrica - Ênfase Em Computação
Engermana Eletrica - Emase Em Compatação	Engenharia Elétrica - Énfase Em Eletrônica E Sistemas Computacionais
	Engenharia Eletrônica E De Computação
Matemática Aplicada e Computação Científica	Interdisciplinar Em Matemática E Computação E Suas Tecnologias
	Matemática Aplicada Com Habilitação Em Sistemas E Controle
	Matemática Aplicada E Computação Científica
	Matemática Aplicada E Computacional Com Habilitação Em Sistemas E Controle
Sistemas de Computação	Sistemas De Computação
Sistemas de Informação	Sistemas De Informação
orotomas as mismayas	proteinas de mormação

Método info da biblioteca Pandas para variáveis numéricas



RangeIndex: 19158 entries, 0 to 19157								
Data columns (total 28 columns):								
#		Non-Null Count	Dtype					
0	NU_ANO_CENSO	19158 non-null	object					
1	NO_REGIAO	19158 non-null	object					
2	NO_UF	19158 non-null	object					
3	SG_UF	19158 non-null	object					
4	NO_MUNICIPIO	19158 non-null	object					
5	CO_MUNICIPIO	19158 non-null	object					
6	IN_CAPITAL_DEPARA	19158 non-null	object					
7	CO_IES	19158 non-null	object					
8	SG_IES	19158 non-null	object					
9	NO_IES	19158 non-null	object					
10	CO_MANTENEDORA	19158 non-null	object					
11	NO_MANTENEDORA	19158 non-null	object					
12	NO_CURSO	19158 non-null	object					

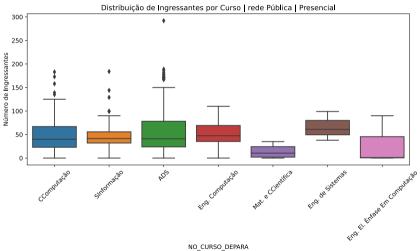
13	NO_CURSO_DEPARA	19158	non-null	object
	CO_CURSO	19158	non-null	object
	TP_GRAU_ACADEMICO_DEPARA	19155	non-null	object
	IN_GRATUITO_DEPARA	19158	non-null	object
17	TP_MODALIDADE_ENSINO_DEPARA	19158	non-null	object
	TP_NIVEL_ACADEMICO_DEPARA	19158	non-null	object
	TP_DIMENSAO_DEPARA	19158	non-null	object
	TP_ORGANIZACAO_ACADEMICA_DEPARA	19158	non-null	object
21	TP_CATEGORIA_ADMINISTRATIVA_DEPARA	19158	non-null	object
	TP_REDE_DEPARA	19158	non-null	object
	QT_VG_TOTAL	19158	non-null	int64
	QT_INSCRITO_TOTAL	19158	non-null	int64
	QT_ING	19158	non-null	int64
	QT_MAT	19158	non-null	int64
27	QT_CONC	19158	non-null	int64
dtyp	es: int64(5), object(23)			
memoi	ry usage: 4.1+ MB			

Método describe para variáveis numéricas (base total)

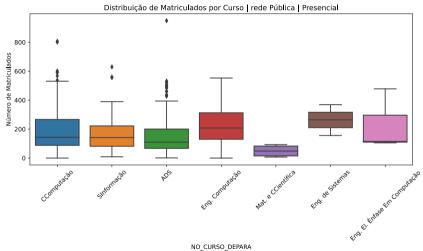


	QT_VG_TOTAL	QT_INSCRITO_TOTAL	QT_ING	QT_MAT	QT_CONC
count	19158	19158	19158	19158	19158
mean	42,84	36,20	10,22	19,61	2,22
std	754,03	468,54	37,86	68,31	10,61
min	0	0	0	0	0
25%	0	0	1	1	0
50%	0	0	2	2	0
75%	0	0	6	8	1
max	73280	32024	1794	3028	608

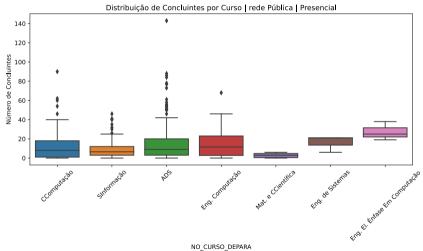






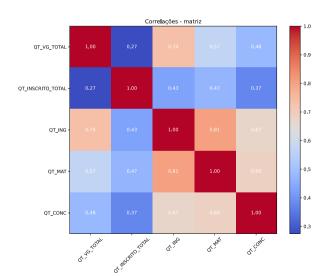




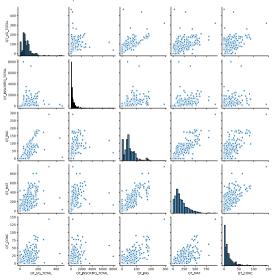


Correlações - matriz | rede Pública | Presencial





Correlações - Pairplot | rede Pública | Presencial





Considerações finais



Com este conjunto de dados, podemos:

- regionalizar (geograficamente) as análises;
- analisar cursos, modalidade, grau acadêmico e rede de ensino separadamente;
- ampliar o conjunto de dados com informações não utilizadas na primeira versão ou complementar com dados de anos anteriores.



Pré-processamento e PCA

Conjunto de dados

Versão depara



RangeIndex: 19158 entries, 0 to 19157								
Data columns (total 17 columns):								
		Non-Null Count	Dtype					
	NO_REGIAO	19158 non-null	object					
	NO_UF	19158 non-null	object					
	IN_CAPITAL_DEPARA	19158 non-null	object					
	NO_CURSO_DEPARA	19158 non-null	object					
	TP_GRAU_ACADEMICO_DEPARA	19155 non-null	object					
	IN_GRATUITO_DEPARA	19158 non-null	object					
	TP_MODALIDADE_ENSINO_DEPARA	19158 non-null	object					
	TP_NIVEL_ACADEMICO_DEPARA	19158 non-null	object					
	·							

	TP_DIMENSAO_DEPARA	19158	non-null	object				
	TP_ORGANIZACAO_ACADEMICA_DEPARA	19158	non-null	object				
	TP_CATEGORIA_ADMINISTRATIVA_DEPARA	19158	non-null	object				
11	TP_REDE_DEPARA	19158	non-null	object				
12	QT_VG_TOTAL	19158	non-null	int64				
13	QT_INSCRITO_TOTAL	19158	non-null	int64				
	QT_ING	19158	non-null	int64				
	TAM_TQ	19158	non-null	int64				
	QT_CONC	19158	non-null	int64				
dtypes: int64(5), object(12)								
memo	memory usage: 2.5+ MB							

One-hot encoding

Utilização do pacote Pandas: get_dummies



```
categories = ['NO_REGIAO', 'NO_UF', 'IN_CAPITAL_DEPARA', 'NO_CURSO_DEPARA',
              'TP_CATEGORIA_ADMINISTRATIVA_DEPARA']
encoded_data = pd.get_dummies(df[categories], prefix=categories, prefix_sep='_')
encoded data = encoded data.apply(lambda x: x.astype(bool).astype(int))
<u>|encoded_data = encod</u>ed_data.merge(df[['QT_VG_TOTAL','QT_INSCRITO_TOTAL','QT_ING',
```

One-hot encoding

Result set



	!#####################################				######################################
P	NO_REGIAO_Centro-Oeste	NU_KEGIAU_		QT_CONC	TP_REDE_DEPARA
0					Pública
					Privada
19153					Privada
19154					Privada
19155					Privada
19156					Privada
19157					Privada
[19158 r	rows x 75 columns]				

StandardScaler

Utilização do pacote StandardScaler do Sklearn



```
print(f' Normalização de variáveis '.center( _width: 80 , _fillchar: '#'))
scaler = StandardScaler()
normalized data = scaler.fit transform(encoded data[columns])
normalized_data = pd.DataFrame(normalized_data, columns=columns)
print(normalized_data)
```

StandardScaler

Result set



####	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	####### Normalizaçã	ão de variá	veis #####	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	############
	QT_VG_TOTAL	QT_INSCRITO_TOTAL	QT_ING	QT_MAT	QT_CONC	
0	0.724341	3.745327	-0.269880	-0.287079	-0.20961	
1	-0.056820	-0.077259	-0.269880	-0.287079	-0.20961	
2	-0.056820	-0.077259	-0.269880	-0.287079	-0.20961	
3	-0.056820	-0.077259	-0.269880	-0.287079	-0.20961	
4	12.420536	12.534500	-0.269880	-0.287079	-0.20961	
191	-0.056820	-0.077259	-0.243467	-0.272439	-0.20961	
191	-0.056820	-0.077259	-0.164229	-0.169958	-0.20961	
191	-0.056820	-0.077259	-0.111404	-0.228518	-0.11534	
191	-0.056820	-0.077259	-0.243467	-0.272439	-0.20961	
1919	-0.056820	-0.077259	-0.243467	-0.272439	-0.20961	

Dimensões do conjunto de dados

Após a padronização



- Variáveis qualitativas (binárias): 69
- Variáveis quantitativas (padronizadas): 5
- Alvo: TP_REDE_DEPARA
- Total de variáveis no result set: 75
- Instâncias: 19158

PCA

Principal Component Analysis

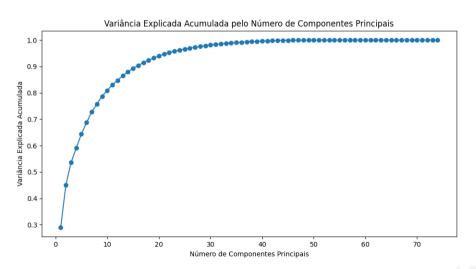


```
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
data_X = normalized_data_vf.drop('TP_REDE_DEPARA', axis=1)
pca = PCA()
pca.fit(data_X)
explained_variance_ratio = pca.explained_variance_ratio_
```

PCA

Principal Component Analysis





PCA

Principal Component Analysis



```
Número de Componentes Principais para 95% de Variância: 22
PC2
                              PC3
                                           PC20
                                                     PC21
                                                              PC22
      0.190351
                3.345125 -0.140522
                                   ... -1.051111 -1.318289 -0.026662
      -0.271789
                0.161091 -0.924383
                                   ... -1.180369 -1.477429 -0.046666
                0.141671
                                   ... -1.032111 -1.547045 -0.003615
     -0.309983
                         0.358834
     -0.309983
                0.141671
                         0.358834
                                   ... -1.032111 -1.547045 -0.003615
      1.720391
                17.651678 -0.612151
                                   ... -0.292294 -0.570287 -0.030889
19153 -0.535000
                -0.050070
                          0.392255
                                       -0.039530 -0.041791
                                                          0.950386
19154 -0.439304
                          0.916337
                -0.069639
                                        0.018221 -0.060235
                                                          0.949860
19155 -0.392884
                -0.077730
                          0.929753
                                   ... -0.017430 -0.033971
                                                          0.951916
19156 -0.447089
               -0.046038
                          0.713141
                                        0.101151
                                                 0.011569
                                                          0.959502
19157 -0.535000
               -0.050070
                         0.392255
                                   ... -0.039530 -0.041791
                                                          0.950386
[19158 rows x 22 columns]
```

Considerações finais



Com esta atividade:

- Efetuamos o processo one-hot encoding, que transformou as variáveis categóricas em numéricas;
- Reduzimos a quantidade de variáveis para 22 componentes principais.



Seleção de características

RELIEF (Kira and Rendell, 1992)



Tipo filtro: seleção de características independente do classificador

```
def relief(X, y):
    X = X.to_numpy()
    y = y.to_numpy()
    print(f'Dimensões da entrada X: {X.shape}')
    print(f'Dimensões da entrada y: {y.shape}')
    num_samples, num_features = X.shape
    weights = np.zeros(num_features)

for i in range(num_samples):
    current_instance = X[i, :]

    nearest_hit = None
    nearest_miss = None
    min_hit_distance = float("inf")
    min_miss_distance = float("inf")
```

```
for j in range(num_samples):
    if i != j:
        distance = np.linalg.norm(current_instance - X[j, :])
    if y[i] == y[j]:
        if distance < min_hit_distance:
            min_hit_distance = distance
            nearest_hit = X[j, :]
    else:
        if distance < min_miss_distance:
            min_miss_distance = distance
            nearest_miss = X[j, :]

weights += np.abs(current_instance - nearest_hit) - \
            np.abs(current_instance - nearest_miss)

return weights / num_samples</pre>
```

RELIEF (Kira and Rendell, 1992)

Tipo filtro: seleção de características independente do classificador



```
Dimensões da entrada X: (19158, 74)
Dimensões da entrada v: (19158.)
Selected Feature Names:
NO_CURSO_DEPARA_Agrocomputação
NO UF Tocantins
NO_UF_Rio Grande do Norte
NO UF Amazonas
NO UF Sergine
NO_UF_Distrito Federal
NO_UF_Acre
NO UF Amapá
NO UF Roraima
TP_DIMENSAO_DEPARA_Cursos a distância com dimensão de dados somente a nível Brasil
NO CURSO DEPARA Eng. El. Énfase Em Computação
TP_ORGANIZACAO_ACADEMICA_DEPARA_Centro Federal de Educação Tecnológica
TP_DIMENSAO_DEPARA_Cursos a distância ofertados por instituições brasileiras no exterior
NO_UF_Sem_uf
NO_REGIAO_Sem_regiao
NO CURSO DEPARA Mat. e CCientífica
NO_CURSO_DEPARA_Eng. de Sistemas
TP_NIVEL_ACADEMICO_DEPARA_Sequencial de Formação Específica
TP_NIVEL_ACADEMICO_DEPARA_Graduação
```



Bibliotecas



```
print(f' Carga de bibliotecas '.center(80 ,'#'))
import pandas as pd
import numpy as np
from sklearn.model selection import train test split
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy score
# Defina a semente
seed value = 45
np.random.seed(seed value)
```



Separação em conjuntos de treinamento e teste

```
print(f' Carga de dados '.center(80 .'#'))
df = pd.read csv('arguivos/dados/tema02 cursos 2021 ti 03 normalized.csv',
              index col=0,sep='|')
print(df.head())
print(df.shape)
X = df.drop('TP REDE DEPARA', axis=1)
v = df['TP REDE DEPARA']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
                              stratify=y, random state=seed value)
```



Treinamento, teste e cálculo da acurária

```
print(f' Naive Bayes Classifier '.center(80 ,'#'))
# Inicialize o classificador Naive Bayes Gaussiano
naive bayes classifier = GaussianNB()
# Treine o classificador com os dados de treinamento
naive bayes classifier.fit(X train, v train)
# Faça previsões nos dados de teste
y pred = naive bayes classifier.predict(X test)
# Avalie a precisão do modelo
```

accuracy = accuracy_score(y_test, y_pred)
print("Acurácia do modelo Naive Bayes:", accuracy)

Naive Bayes Classifier

Conjunto de dados Total



```
NO_REGIAO_Centro-Oeste NO_REGIAO_Nordeste ... OT_CONC TP_REDE_DEPARA
                                      Pública
                           ... -0.20961
                          ... -0.20961
                                      Pública
                          ... -0.20961
                                      Pública
                                      Pública
                         0 ... -0.20961
                         0 ... -0.20961
                                      Pública
[5 rows x 75 columns]
Acurácia do modelo Naive Bayes: 1.0
```

Naive Bayes Classifier

Conjunto de dados PCA



```
PC1
                         PC3
                                      PC21
                                               PC22
                                                     TP_REDE_DEPARA
0 -0.271789
           0.161091 -0.924383
                              ... -1.477637 -0.047432
                                                            Pública
  0.190351
           3.345125 -0.140522
                              ... -1.318910 -0.022477
                                                            Pública
 -0.245303 0.359979 -0.920680
                              ... -1.464840 -0.046929
                                                           Pública
                                                           Pública
           3.408979 -0.922878
                              ... -1.264806 -0.039676
4 -0.284295
           0.157172 -0.929598
                              ... -1.490612 -0.019619
                                                           Pública
[5 rows x 23 columns]
(19158, 23)
############################# Naive Bayes Classifier ############################
Acurácia do modelo Naive Bayes: 0.9835594989561587
```

Naive Bayes Classifier





```
NO_UF_Rondônia ... TP_REDE_DEPARA
                        Pública
                        Pública
                        Pública
                        Pública
                        Pública
[5 rows x 23 columns]
(19158, 23)
############################## Naive Bayes Classifier ###########################
Acurácia do modelo Naive Bayes: 0.9191022964509394
```

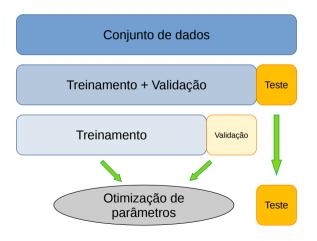


Estimação de desempenho

Separação de conjuntos de dados

Treinamento, validação e teste





EACH Escola de Artes, Clâncias e Humanidades Universidade de São Paulo

- Validação cruzada estratificada
 - Estratégia: grid-serch
 - Maximização: F1-score
 - Valor de k: 10
 - Modelo: Naive Bayes
 - Valores de hiperparâmetros testados:
 - 'var_smoothing': [1e-9, 1e-8, 1e-7]

Portion of the largest variance of all features that is added to variances for calculation stability.

^{*} var_smoothingfloat, default=1e-9

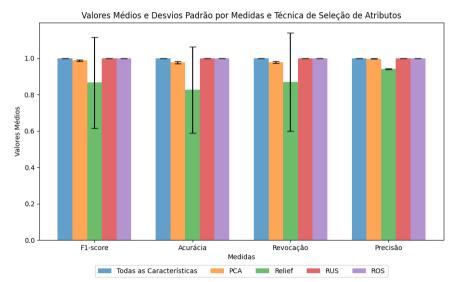
Validação cruzada estratificada



Médias	F1-score	Acurácia	Revocação	Precisão	Hiperparâmetros
Todas as características	0.9999	0.9998	0.9998	1.0000	1e-09
PCA	0.9877	0.9771	0.9780	0.9975	1e-09
Relief	0.8666	0.8266	0.8697	0.9401	1e-09
RUS	1.0000	1.0000	1.0000	1.0000	1e-09
ROS	0.9999	0.9999	0.9999	1.0000	1e-09

Resultados





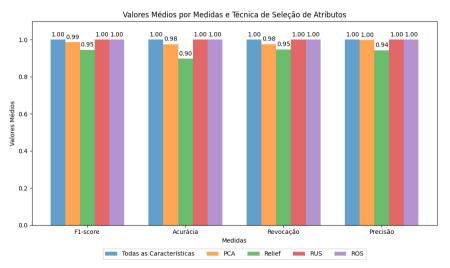
Valores obtidos com os conjuntos de teste



Médias	F1-score	Acurácia	Revocação	Precisão
Todas as características	0.9999	0.9998	0.9998	1.0000
PCA	0.9872	0.9762	0.9762	0.9984
Relief	0.9456	0.8975	0.9480	0.9432
RUS	1.0000	1.0000	1.0000	1.0000
ROS	1.0000	1.0000	1.0000	1.0000

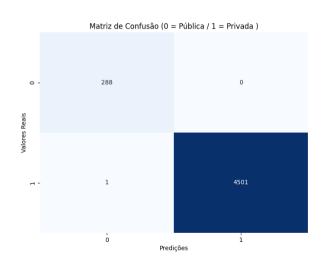
Resultados utilizando o conjunto de teste





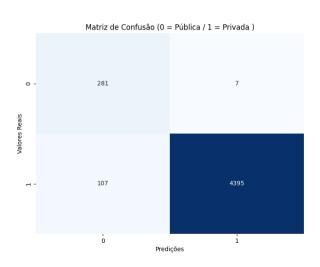
Matriz de confusão: total de atributos





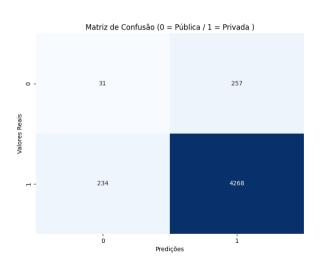
Matriz de confusão: PCA





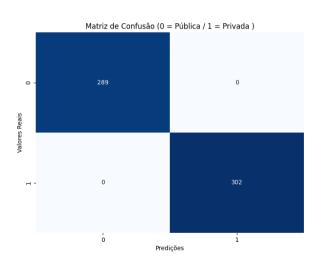
Matriz de confusão: Relief





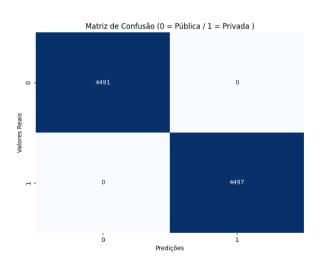
Matriz de confusão: Rus





Matriz de confusão: Ros





Obrigado!

Thanks! / ¡Gracias!



Leonardo Cunha dos Santos lattes.cnpq.br/5620610314140397

leonardo.cunha.santos@usp.br

Gabriel Francisco dos Santos Silva gabfssilva@gmail.com