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

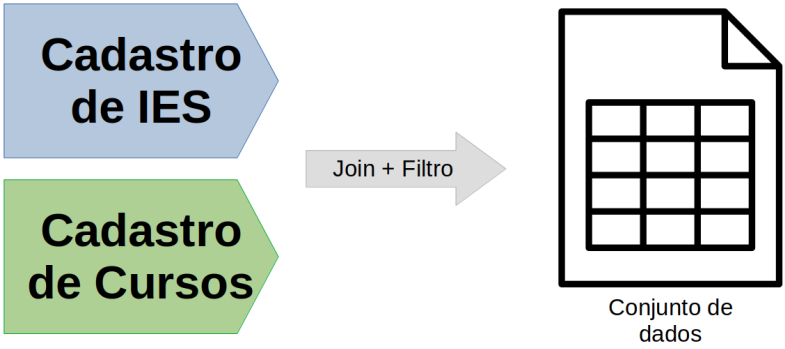
Agenda

1. Descrição do dataset e análise exploratória
2. Pré-processamento e PCA
3. Seleção de características
4. Naive Bayes Classifier
5. Estimação de desempenho
6. Estimação de desempenho e comparação entre classificadores
7. Stacking Classifier
8. Considerações finais

Descrição do dataset e análise exploratória

Pré-processamento de dados

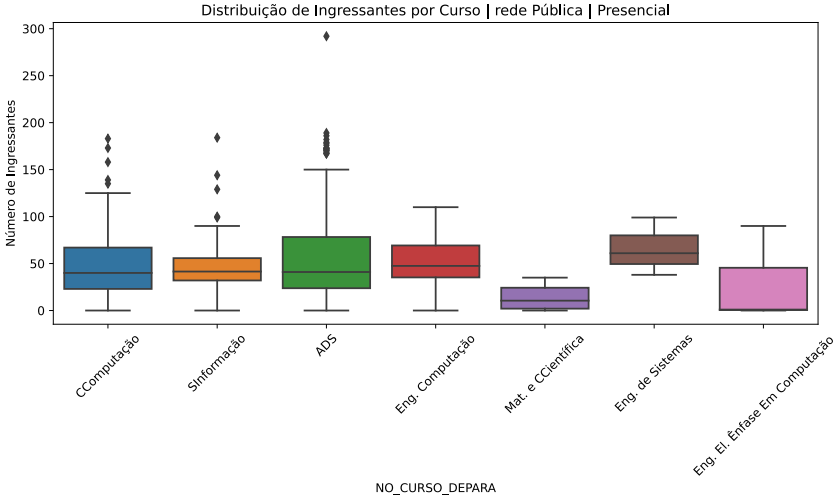
Cursos de tecnologia



Variáveis numéricas: 05 | Variáveis categóricas: 22 + 1 (ID:ANO)

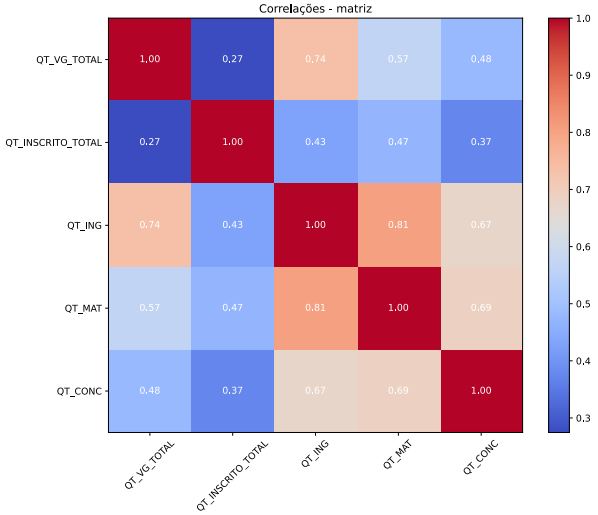
34 cursos | Instâncias: 19158 | 558 cursos distintos

Análise exploratória



Análise exploratória

Correlações - matriz | rede Pública | Presencial



Pré-processamento e PCA

One-hot encoding

Utilização do pacote Pandas: get_dummies


```
categories = ['NO_REGIAO', 'NO_UF', 'IN_CAPITAL_DEPARA', 'NO_CURSO_DEPARA',
              'TP_GRAU_ACADEMICO_DEPARA','IN_GRATUITO_DEPARA','TP_MODALIDADE_ENSINO_DEPARA',
              'TP_NIVEL_ACADEMICO_DEPARA','TP_DIMENSAO_DEPARA','TP_ORGANIZACAO_ACADEMICA_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))

encoded_data = encoded_data.merge(df[['QT_VG_TOTAL','QT_INSCRITO_TOTAL','QT_ING',
                                      'QT_MAT','QT_CONC','TP_REDE_DEPARA']],
                                  left_index=True, right_index=True, how='left')
```

StandardScaler

Utilização do pacote StandardScaler do Sklearn

EACH | USP campus capital
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

```
################################################################################
print(f' Normalização de variáveis '.center( __width: 80, __fillchar: '#'))

scaler = StandardScaler()

columns=['QT_VG_TOTAL','QT_INSCRITO_TOTAL','QT_ING','QT_MAT','QT_CONC']

normalized_data = scaler.fit_transform(encoded_data[columns])
normalized_data = pd.DataFrame(normalized_data, columns=columns)
print(normalized_data)
```

Dimensões do conjunto de dados

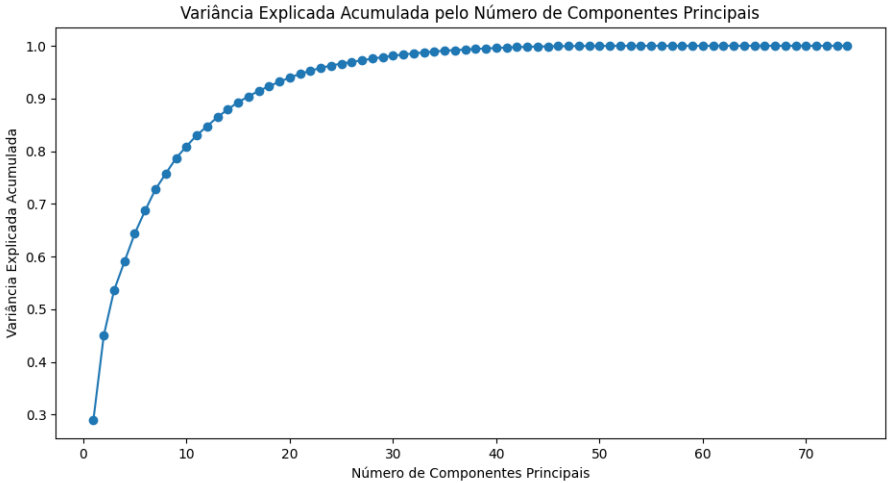
Após a padronização

EACH | campus capital USP leste
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

- 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



PCA

Principal Component Analysis


```
Número de Componentes Principais para 95% de Variância: 22
#################### PCA com número de componentes esperado ####################
            PC1       PC2       PC3  ...      PC20      PC21      PC22
0      0.190351  3.345125 -0.140522  ... -1.051111 -1.318289 -0.026662
1     -0.271789  0.161091 -0.924383  ... -1.180369 -1.477429 -0.046666
2     -0.309983  0.141671  0.358834  ... -1.032111 -1.547045 -0.003615
3     -0.309983  0.141671  0.358834  ... -1.032111 -1.547045 -0.003615
4      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 -0.069639  0.916337  ...  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]
```

Seleção de características

RELIEF (Kira and Rendell, 1992)

Tipo filtro: seleção de características independente do classificador

EACH | USP campus capital
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

```python
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")
```

```python
        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
```

Naive Bayes Classifier

Naive Bayes Classifier

Conjunto de dados Total

```
################################## Carga de dados ##################################
   NO_REGIAO_Centro-Oeste  NO_REGIAO_Nordeste  ...  QT_CONC  TP_REDE_DEPARA
0                       0                   0  ... -0.20961         Pública
1                       0                   0  ... -0.20961         Pública
2                       0                   0  ... -0.20961         Pública
3                       0                   0  ... -0.20961         Pública
4                       0                   0  ... -0.20961         Pública

[5 rows x 75 columns]
(19158, 75)
############################## Naive Bayes Classifier ##############################
Acurácia do modelo Naive Bayes: 1.0
```

Naive Bayes Classifier

Conjunto de dados PCA

EACH | campus capital USP leste
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

```
################################## Carga de dados ##################################
        PC1       PC2       PC3  ...      PC21      PC22  TP_REDE_DEPARA
0 -0.271789  0.161091 -0.924383  ... -1.477637 -0.047432         Pública
1  0.190351  3.345125 -0.140522  ... -1.318910 -0.022477         Pública
2 -0.245303  0.359979 -0.920680  ... -1.464840 -0.046929         Pública
3  0.237584  3.408979 -0.922878  ... -1.264806 -0.039676         Pública
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

Conjunto de dados Selecionado - 22 atributos do Relief

```
################################## Carga de dados ##################################
   NO_UF_Rondônia  ...  TP_REDE_DEPARA
0               0  ...         Pública
1               0  ...         Pública
2               0  ...         Pública
3               0  ...         Pública
4               0  ...         Pública

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

Estimação de desempenho

Planejamento do experimento

Validação cruzada estratificada

- Estratégia: grid-search
- Maximização: F1-score
- Valor de k: 5
- Modelo: Naive Bayes
- Valores de hiperparâmetros testados:
  $'var\_smoothing' : np.logspace(0, -15, num = 100)$

* var_smoothingfloat, default=1e-9
Suavização à estimativa da variância.

Planejamento do experimento

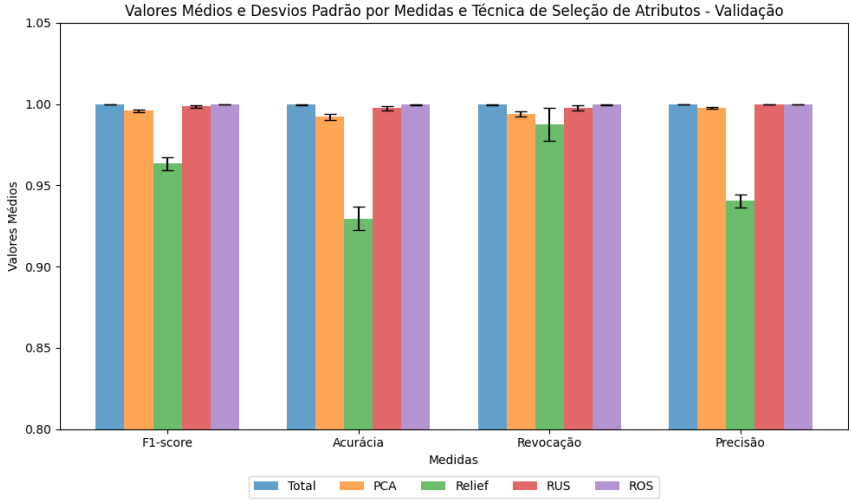
Validação cruzada estratificada

EACH | campus capital USP leste
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Médias	F1-score	Acurácia	Revocação	Precisão	var_smoothing(avg)
Total	0.9998	0.9997	0.9997	0.9999	5.3e-04
PCA	0.9958	0.9921	0.9939	0.9977	5.1e-01
Relief	0.9634	0.9296	0.9876	0.9404	8.0e-01
RUS	0.9986	0.9974	0.9975	0.9998	1.0e-01
ROS	0.9998	0.9997	0.9997	0.9999	7.4e-05

Planejamento do experimento

Resultados



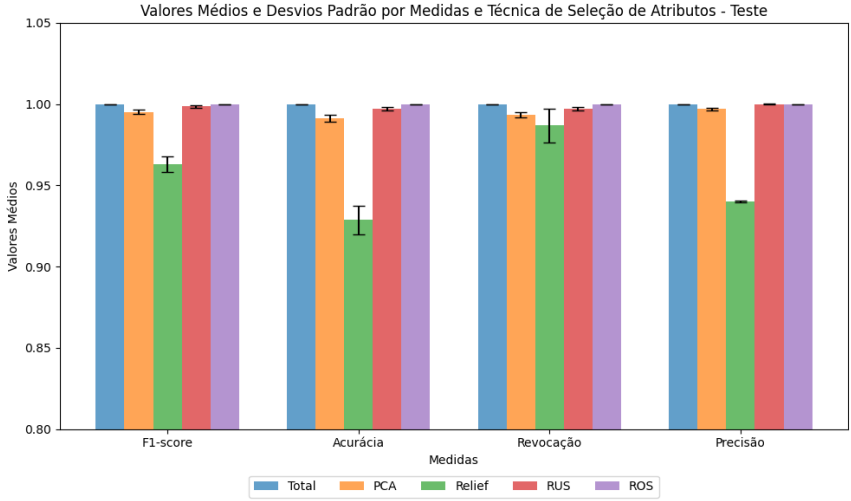
Planejamento do experimento

Valores obtidos com os conjuntos de teste

Médias	F1-score	Acurácia	Revocação	Precisão
Total	0.9999	0.9998	0.9999	0.9999
PCA	0.9952	0.9911	0.9934	0.9971
Relief	0.9629	0.9287	0.9868	0.9402
RUS	0.9985	0.9972	0.9972	0.9999
ROS	0.9999	0.9998	0.9999	0.9999

Planejamento do experimento

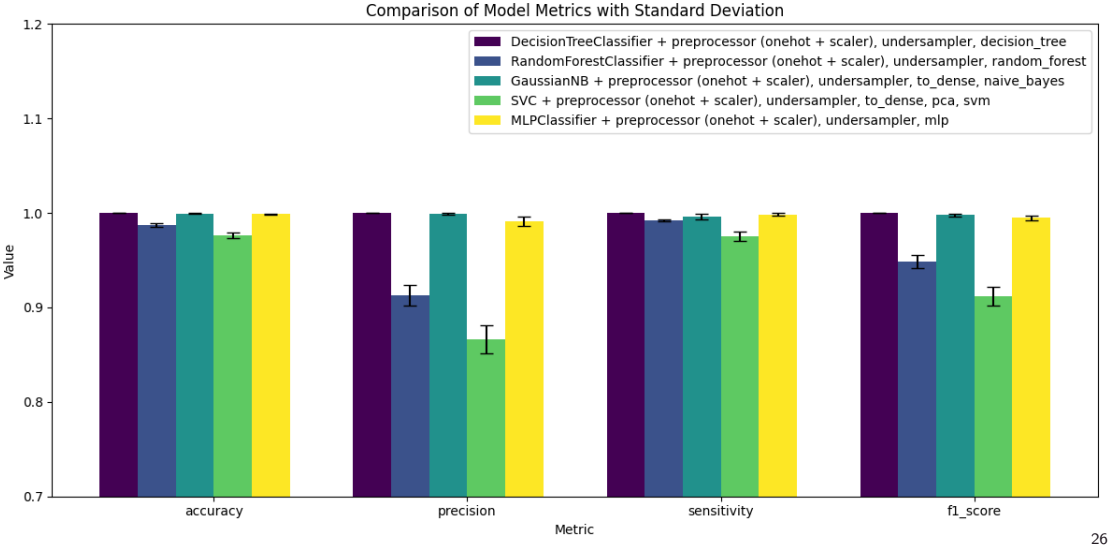
Resultados utilizando o conjunto de teste



Estimação de desempenho e comparação entre classificadores

Comparativo do experimento

Resultados com o conjunto de teste



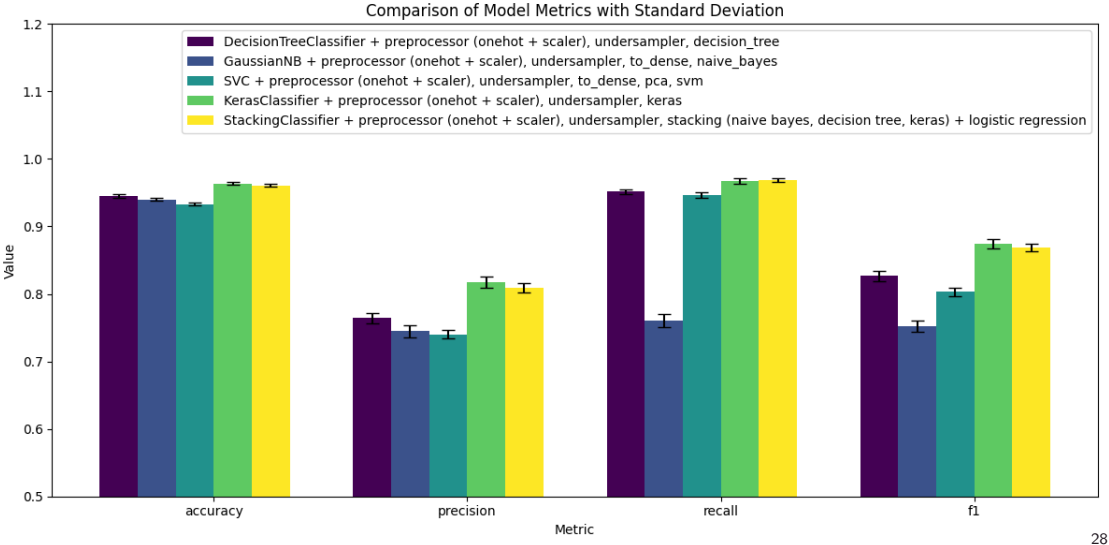
Utilizando Stacking Classifier

- Classificadores:
  - Naive Bayes
  - Decision Tree
  - Neural Network (Keras)

- Classificador final:
  - Logistic Regression

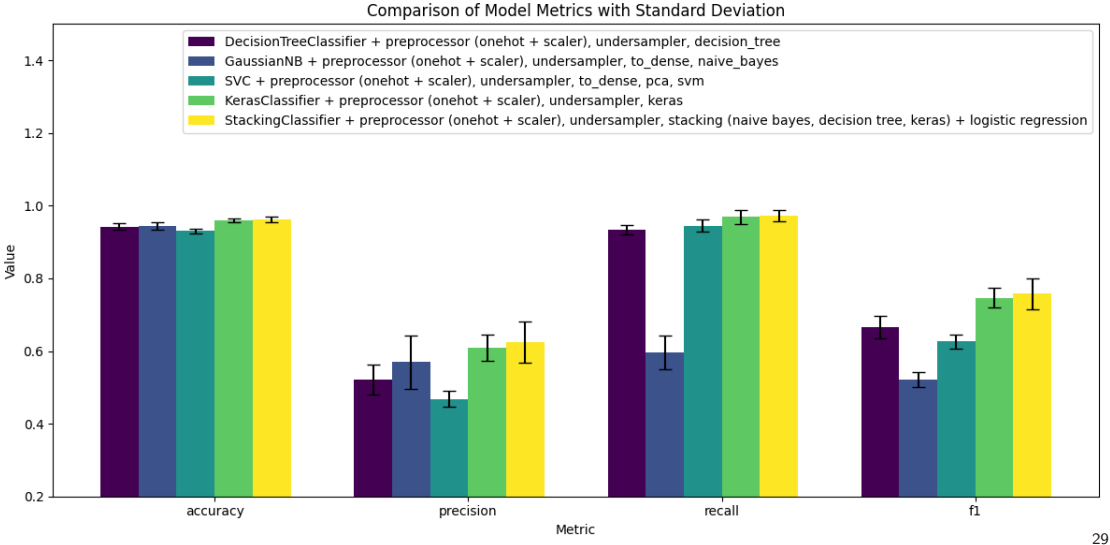
Comparativo do experimento

Resultados – Amostra de teste + resample



Comparativo do experimento

Resultados – Amostra de calibração do cross validation



Considerações finais

- O pré-processamento e a seleção de atributos foram de fundamental importância para o entendimento sobre o conjunto de dados;
- As técnicas de redução de dimensionalidade auxiliaram na diminuição do tempo de processamento e no entendimento sobre a variabilidade dos dados;
- A retirada de atributos (in_gratuito e categoria_administrativa) melhorou a performance de indicadores quando aplicado o modelo de árvore de decisão;
- A utilização do Pipeline ajudou na execução de vários modelos com diferentes configurações, de modo simples e eficiente;
- A abordagem com o StackingClassifier apresentou melhores resultados para os indicadores investigados. A maior variabilidade detectada ocorreu na medida de precisão.

Considerações finais

- O pré-processamento e a seleção de atributos foram de fundamental importância para o entendimento sobre o conjunto de dados;
- As técnicas de redução de dimensionalidade auxiliaram na diminuição do tempo de processamento e no entendimento sobre a variabilidade dos dados;
- A retirada de atributos (in_gratuito e categoria_administrativa) melhorou a performance de indicadores quando aplicado o modelo de árvore de decisão;
- A utilização do Pipeline ajudou na execução de vários modelos com diferentes configurações, de modo simples e eficiente;
- A abordagem com o StackingClassifier apresentou melhores resultados para os indicadores investigados. A maior variabilidade detectada ocorreu na medida de precisão.

Considerações finais

EACH | campus capital USP LESTE
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

- O pré-processamento e a seleção de atributos foram de fundamental importância para o entendimento sobre o conjunto de dados;
- As técnicas de redução de dimensionalidade auxiliaram na diminuição do tempo de processamento e no entendimento sobre a variabilidade dos dados;
- A retirada de atributos (in_gratuito e categoria_administrativa) melhorou a performance de indicadores quando aplicado o modelo de árvore de decisão;
- A utilização do Pipeline ajudou na execução de vários modelos com diferentes configurações, de modo simples e eficiente;
- A abordagem com o StackingClassifier apresentou melhores resultados para os indicadores investigados. A maior variabilidade detectada ocorreu na medida de precisão.

Considerações finais

EACH | campus capital USP LESTE
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

- O pré-processamento e a seleção de atributos foram de fundamental importância para o entendimento sobre o conjunto de dados;
- As técnicas de redução de dimensionalidade auxiliaram na diminuição do tempo de processamento e no entendimento sobre a variabilidade dos dados;
- A retirada de atributos (in_gratuito e categoria_administrativa) melhorou a performance de indicadores quando aplicado o modelo de árvore de decisão;
- A utilização do Pipeline ajudou na execução de vários modelos com diferentes configurações, de modo simples e eficiente;
- A abordagem com o StackingClassifier apresentou melhores resultados para os indicadores investigados. A maior variabilidade detectada ocorreu na medida de precisão.

Considerações finais

EACH | campus capital USP LESTE
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

- O pré-processamento e a seleção de atributos foram de fundamental importância para o entendimento sobre o conjunto de dados;
- As técnicas de redução de dimensionalidade auxiliaram na diminuição do tempo de processamento e no entendimento sobre a variabilidade dos dados;
- A retirada de atributos (in_gratuito e categoria_administrativa) melhorou a performance de indicadores quando aplicado o modelo de árvore de decisão;
- A utilização do Pipeline ajudou na execução de vários modelos com diferentes configurações, de modo simples e eficiente;
- A abordagem com o StackingClassifier apresentou melhores resultados para os indicadores investigados. A maior variabilidade detectada ocorreu na medida de precisão.

Obrigado!

Thanks! / ¡Gracias!

EACH | campus capital USP leste
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Leonardo Cunha dos Santos
lattes.cnpq.br/5620610314140397
leonardo.cunha.santos@usp.br

Gabriel Francisco dos Santos Silva
gabfssilva@gmail.com