Protótipo

Prognóstico do Câncer do Colo Uterino

Implementação e Avaliação de Modelos Preditivos

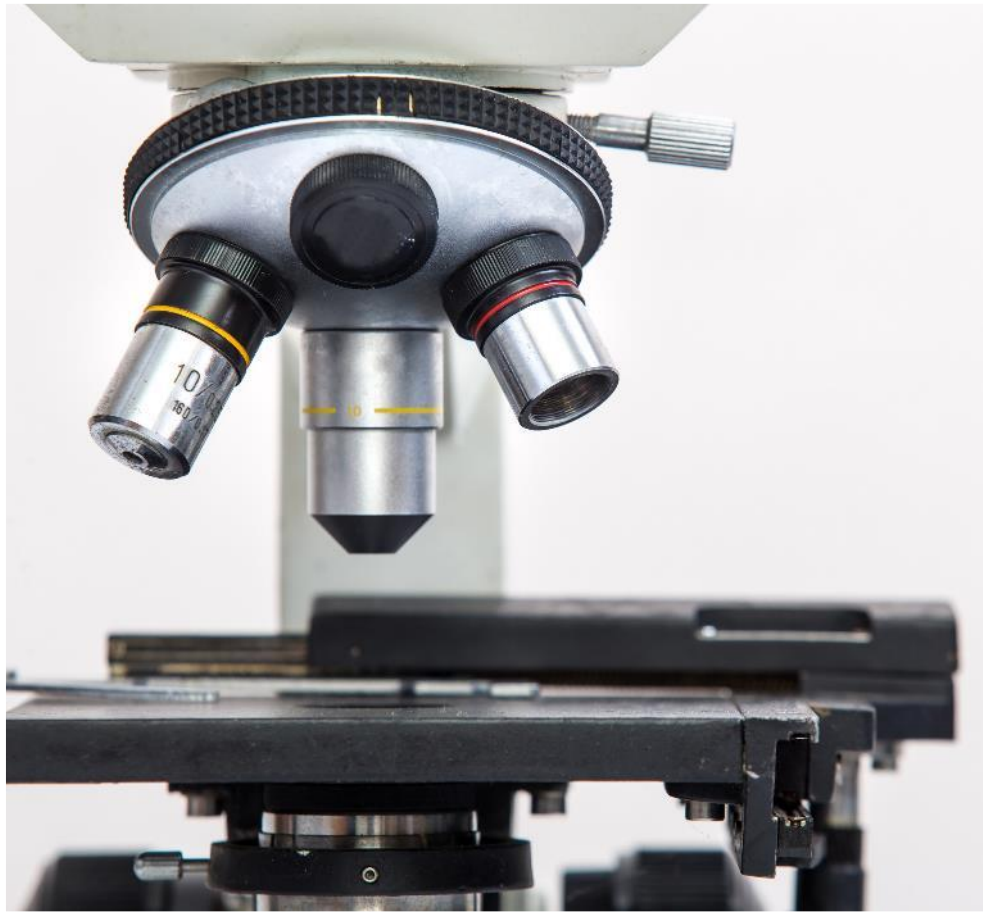
Motivação

Por que o assunto nos interessou?

A doença

Câncer do Colo do Útero (ou Cervical)

- Papilomavírus Humano (HPV)
- Fatores de risco:
  - início precoce da atividade sexual;
  - múltiplos parceiros sexuais;
  - uso prolongado de pílulas anticoncepcionais;
  - tabagismo.
- Diagnóstico e tratamento precoces



SINTOMAS

Secreção vaginal anormal

Sangramento vaginal intermitente ou após a relação sexual

Dor abdominal associada a queixas urinárias ou intestinais nos casos mais avançados

VOCÊ SABIA?

- O pico de sua incidência se dá na faixa etária de 45 a 50 anos
- Aproximadamente 100% dos casos do câncer resultam do HPV
- Existem 13 tipos de HPV considerados oncogênicos
- A mortalidade reduziu mais de 50% desde a introdução do Papanicolau

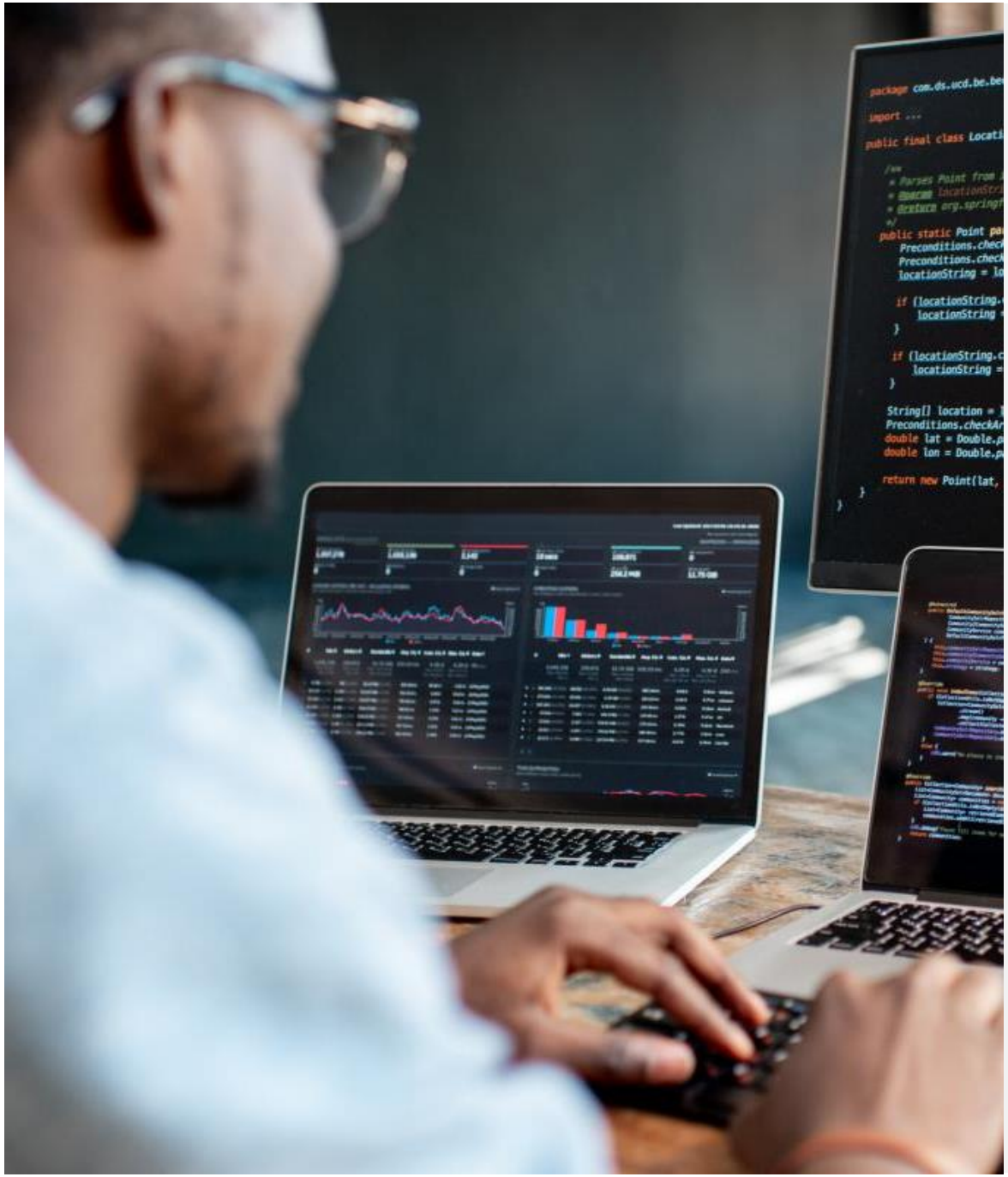
Protótipo

Conjunto de Dados

Cervical cancer (Risk Factors) Data Set

- UCI - University of California, Irvine
  - Center for Machine Learning and Intelligent Systems

Data Set Characteristics:	Multivariate
Attribute Characteristics:	Integer, Real
Associated Tasks:	Classification
Number of Instances:	858
Number of Attributes:	36
Missing Values?	Yes
Area:	Life
Date Donated:	03/03/2017
Number of Web Hits:	129.087



Pré-processamento

O que desenvolvemos até agora?

Data Extraction

risk_factors_cervical_cancer.csv

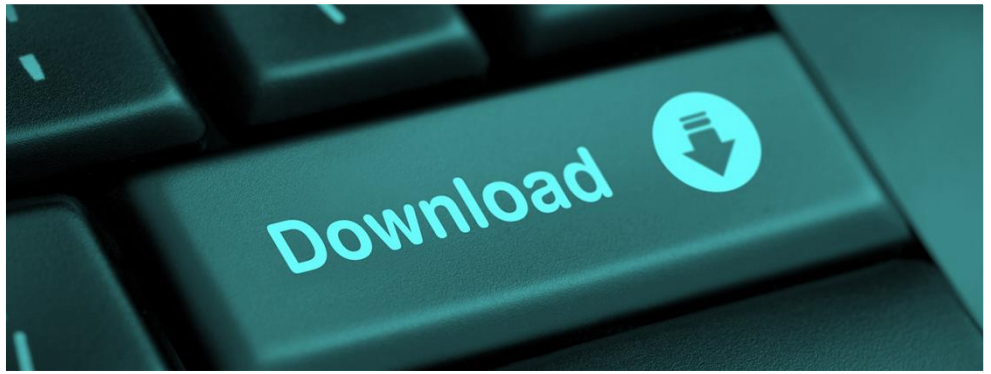
```
import requests


url = 'https://archive.ics.uci.edu/ml/machine-
learning-
databases/00383/risk_factors_cervical_cancer.csv'


arquivo = url.rsplit('/', 1)[1]



r = requests.get(url, allow_redirects=True)
open(arquivo, 'wb').write(r.content)
```



Exploratory Analysis

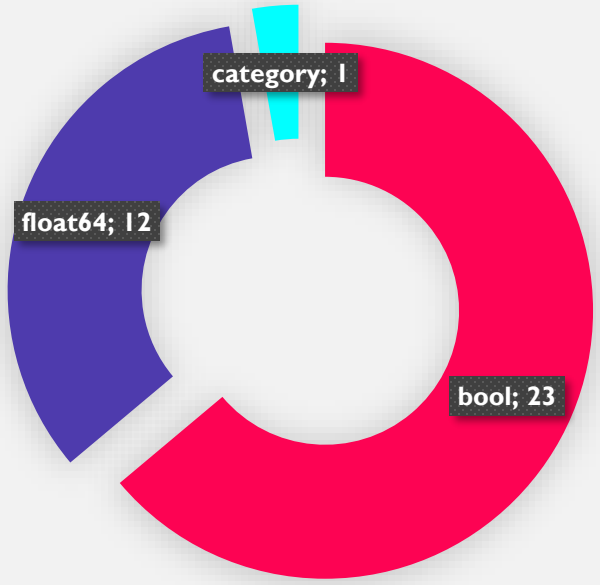
"Profile - Colunas Originais"

- 858 observações
- Entendimento de Domínio (35)
- 4 Targets candidatas:
  - Hinselmann
  - Schiller
  - Citology
  - Biopsy
- Variável Resposta: Biópisa (1)

Feature Engineering

"Profile - Colunas Padronizadas"

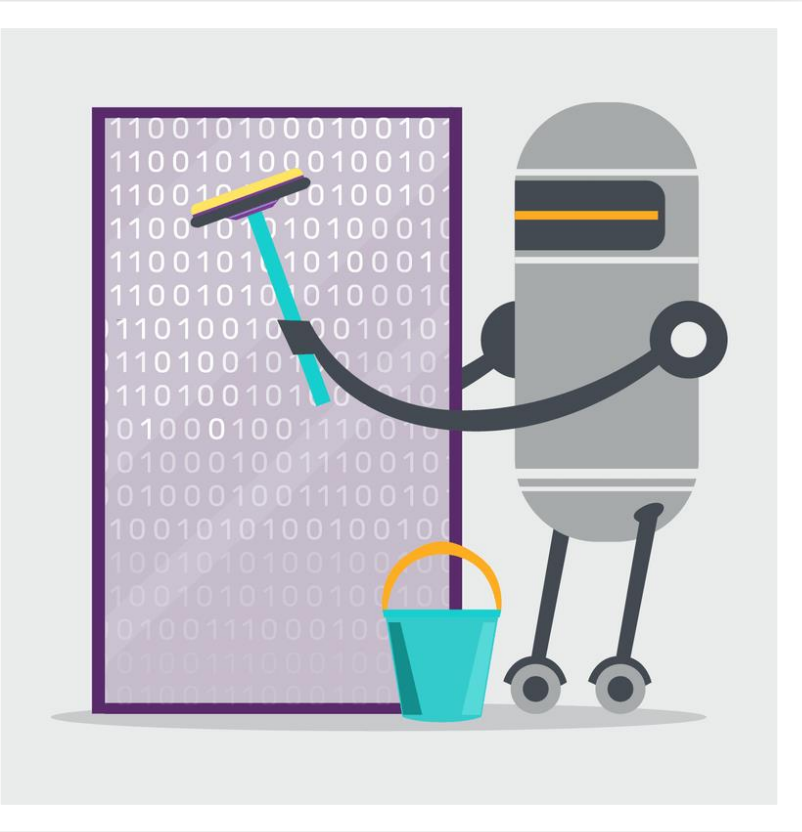
- Ações sobre as Colunas:
  - Renomeação
  - Ordenação
  - Tipagem



Data Cleaning

"Profile - Valores Padronizados"

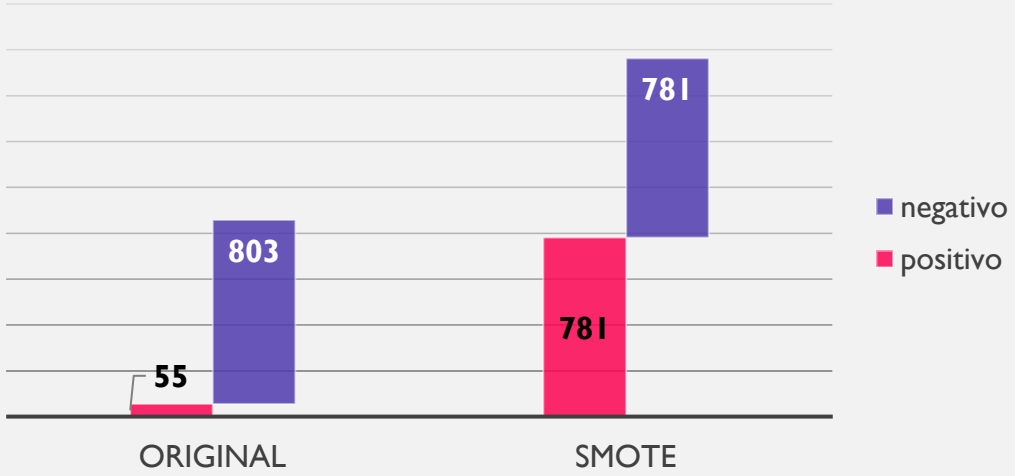
- Registros Duplicados (23): Remoção
- Valores Nulos (1.957): Mediana
- Padronização: MinMaxScaler (0,1) float64
- One-hot Encoder: pd.get_dummies
  - de bool para float64
  - [coluna_valor]
  - 57 features
  - 100% float64
- Persistência da lista da Features (colunas_final.csv)

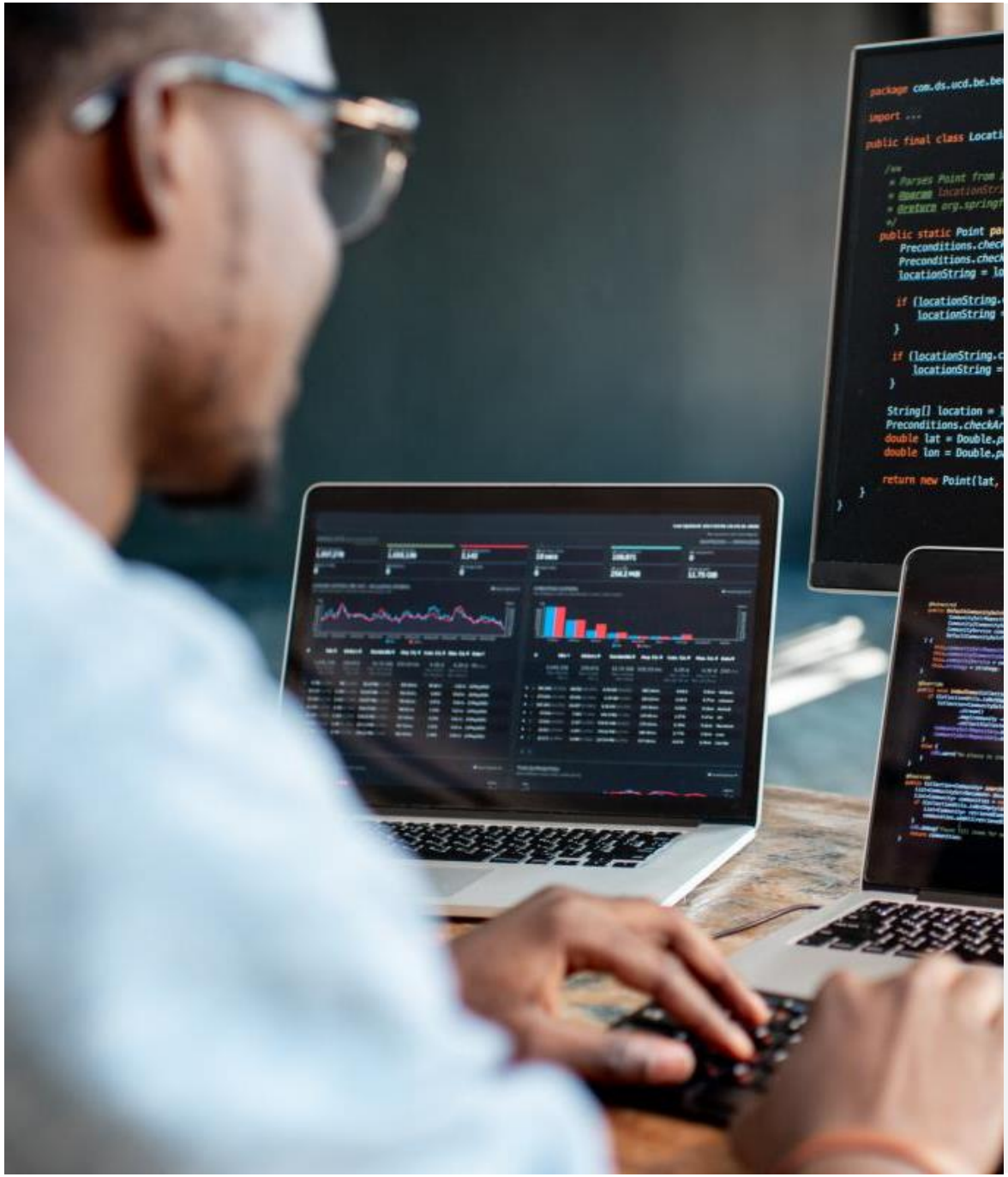


Data Generation

Synthetic Minority Oversampling TEchnique

- SMOTE:
  - Target: de 803 | 55 para 781 | 781
- Persistência das Dados Pré-processados (dados_final.csv)





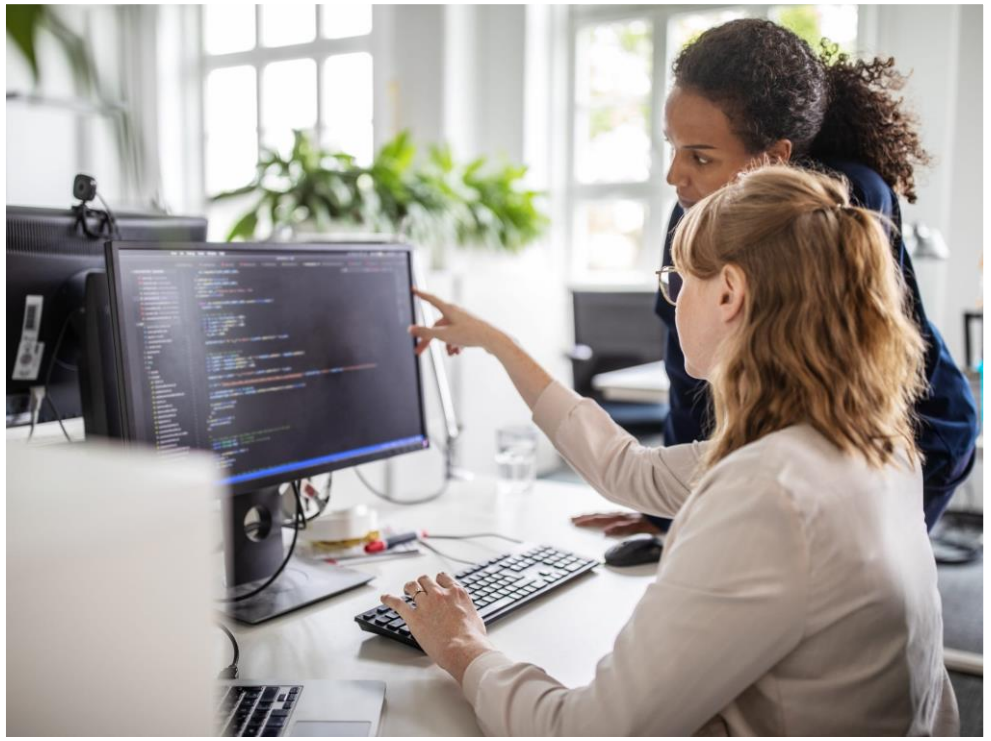
Treinamento

O que já foi possível implementar?

Data Clustering

Modelagem Experimental

- train_test_split – 70/30%
- Acurácia dos Modelos:
  - Support Vector Machines (SVM): 94,37%
  - K-Nearest Neighbors (KNN): 97,44%
  - Decision Tree Classifier (DTC): 96,42%
- Avaliação:
  - Coeficiente de determinação
  - R2 Score KNN: 0,8976
  - A melhor pontuação possível é 1, obtida quando os valores previstos são iguais aos valores reais.



Próximos passos

Quais pontos ainda não foram tratados?

Overview

Overview | Warnings 24 | Reproduction

Warnings

dsts_condilomatose_colo_utero is highly correlated with dsts_aids and 8 other fields	High correlation
dsts_aids is highly correlated with dsts_condilomatose_colo_utero and 8 other fields	High correlation
dsts_condilomatose_vaginal is highly correlated with dsts_aids and 6 other fields	High correlation
dsts_condilomatose_vulvo_perineal is highly correlated with dsts_condilomatose	High correlation
dsts_condilomatose is highly correlated with dsts_condilomatose_vulvo_perineal	High correlation
dsts_doenca_inflamatoria_pelvica is highly correlated with dsts_aids and 8 other fields	High correlation
dsts_hepatite_b is highly correlated with dsts_aids and 8 other fields	High correlation
dsts_herpes_genital is highly correlated with dsts_aids and 8 other fields	High correlation
dsts_hiv is highly correlated with dsts_aids and 5 other fields	High correlation
dsts_hpv is highly correlated with dsts_aids and 6 other fields	High correlation
dsts_molusco_contagioso is highly correlated with dsts_aids and 8 other fields	High correlation
dsts_sifilis is highly correlated with dsts_aids and 5 other fields	High correlation
dsts_ultimo_diagnostico_anos is highly correlated with dsts_primeiro_diagnostico_anos	High correlation
dsts_primeiro_diagnostico_anos is highly correlated with dsts_ultimo_diagnostico_anos	High correlation
df_index has unique values	Unique
contraceptivos_hormonais_anos has 255 (30.5%) zeros	Zeros
diu_anos has 752 (90.1%) zeros	Zeros
dsts_numero has 756 (90.5%) zeros	Zeros
dsts_primeiro_diagnostico_anos has 15 (1.8%) zeros	Zeros
dsts_ultimo_diagnostico_anos has 17 (2.0%) zeros	Zeros
fuma_anos has 712 (85.3%) zeros	Zeros
fuma_macos_ano has 712 (85.3%) zeros	Zeros
gestacoes_numero has 16 (1.9%) zeros	Zeros
parceiros_sexuais_numero has 193 (23.1%) zeros	Zeros

+ Analisar correlações significativas

Protótipo

+ Melhorias gerais

Revisão dos passos anteriores

- Reparametrizar:
  - Support Vector Machines (SVM)
  - K-Nearest Neighbors (KNN)
  - Decision Tree Classifier (DTC)
- Avaliar métricas relevantes
- + Rede Neural (?)
- Criar Pipeline(s)
- Aplicar Versionamento (Github)
- Estruturar o Projeto (Cookiecutter)
- Rever Engenharia de Features (Featuretools)

Obrigad@!

Projeto Integrado em Aprendizado de Máquina

✉ {marcelohonoliveira, mari.ikoma, nayara.mohana} @gmail.com