

Redução da dimensionalidade em bigdata

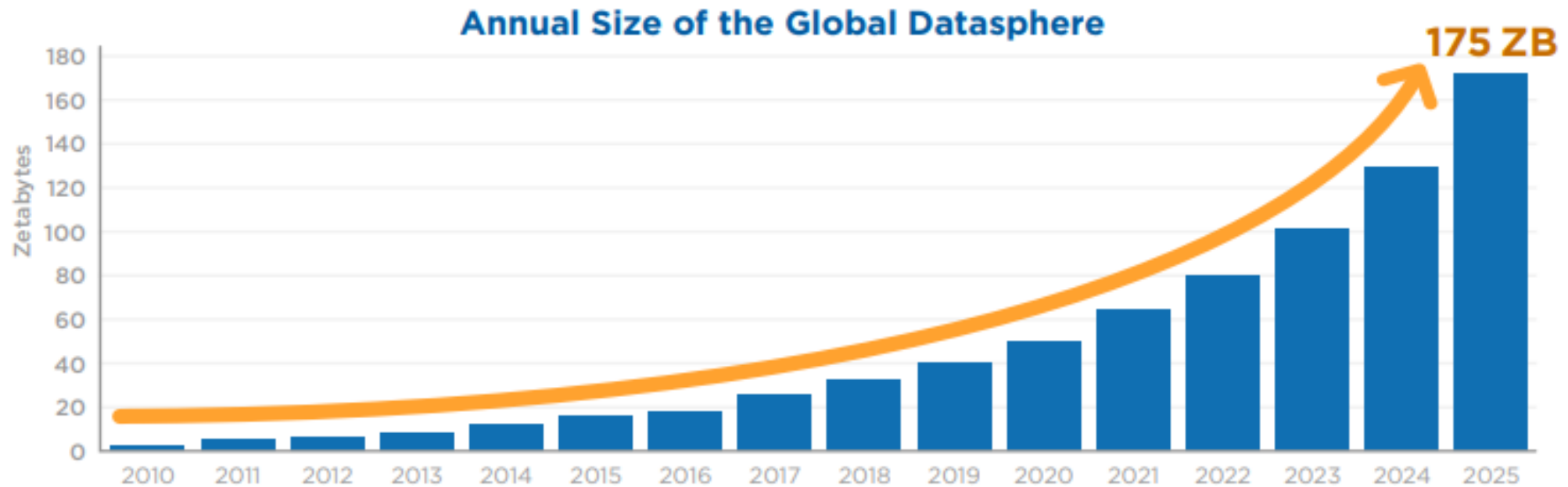
Lucca de Castro Machado – 32292783

Prof. Dr. Anderson Borba

Introdução

Problemas

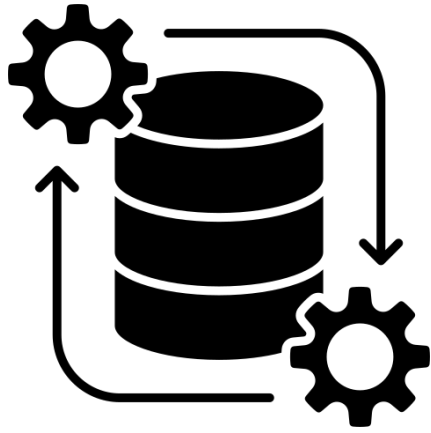
Crescimento exponencial dos dados



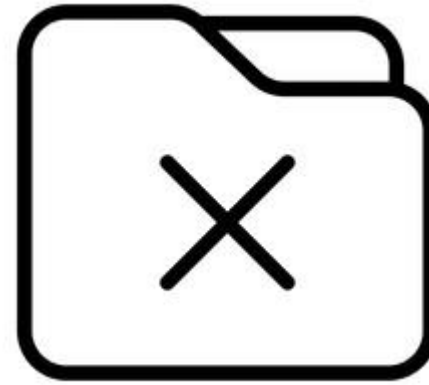
Fonte: <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>

Introdução

Problemas



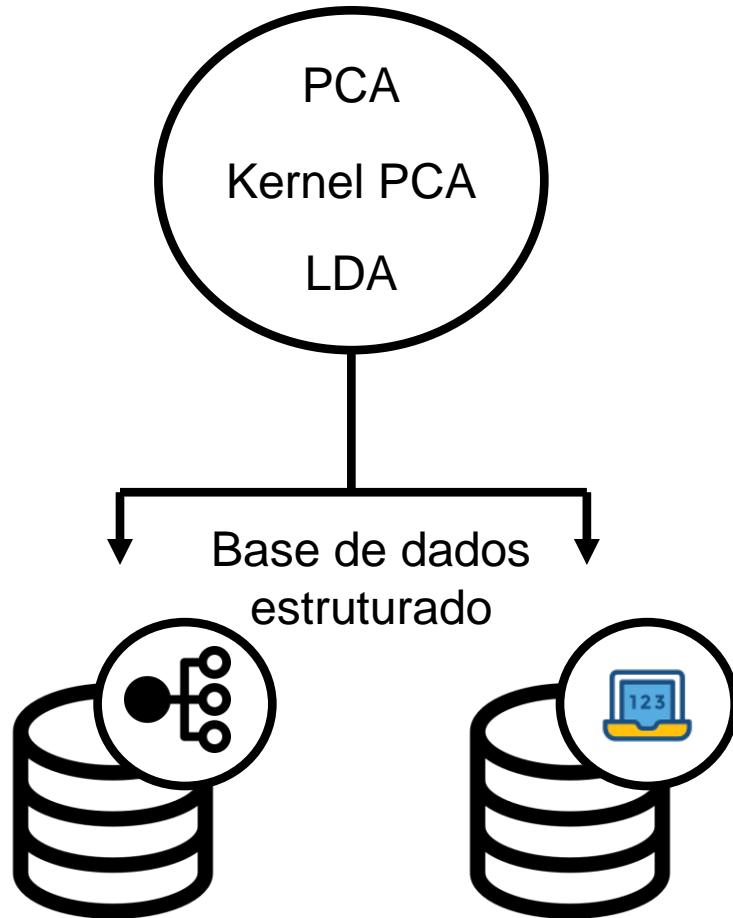
Processamento de
dados



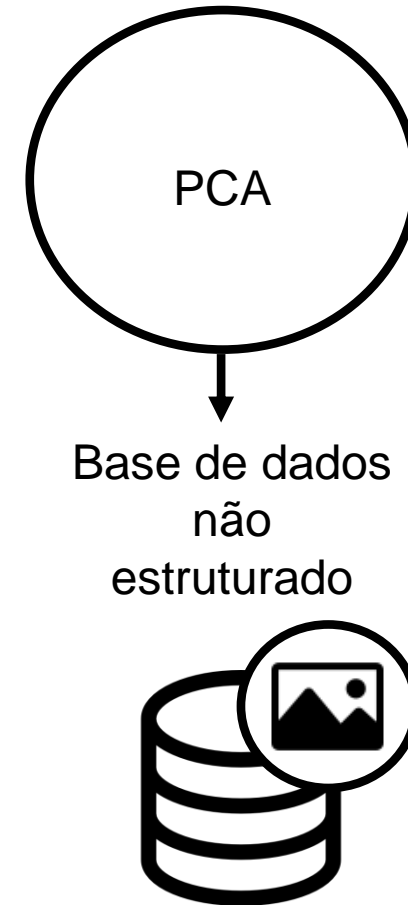
Espaço de
armazenamento

Introdução

Objetivo principal

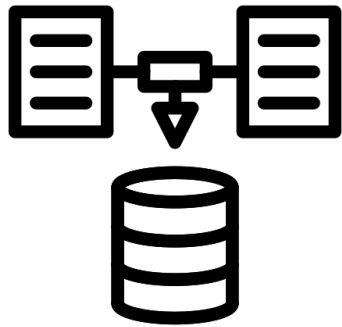


Técnicas de
redução da
dimensionalidade

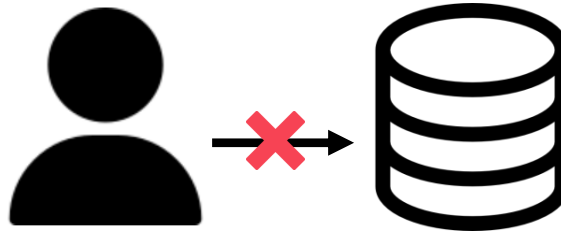


Background Teórico

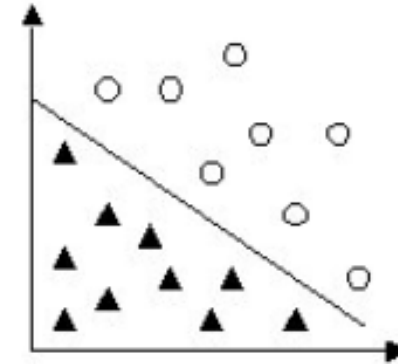
PCA (Principal Component Analysis)



Extração de
características



Aprendizagem não
supervisionada



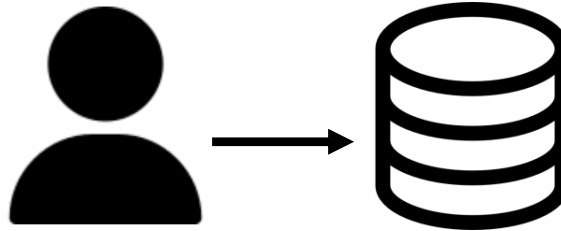
Processamento de
dados lineares

Background Teórico

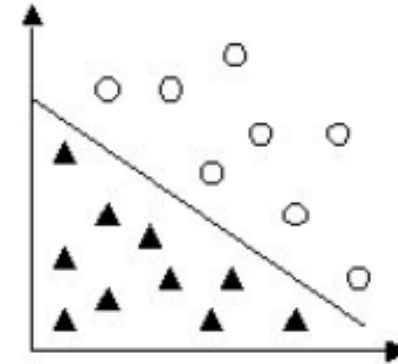
LDA (Linear Discriminant Analysis)



Seleção de
características



Aprendizagem
supervisionada

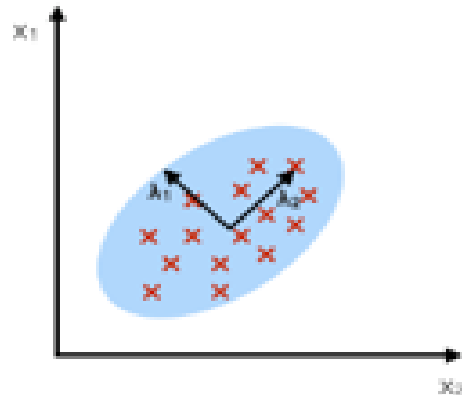


Processamento de
dados lineares

Background Teórico

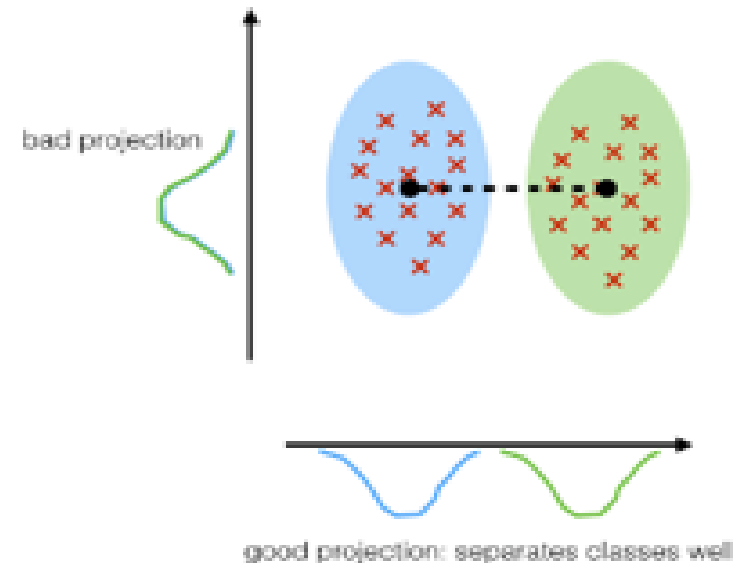
PCA:

component axes that
maximize the variance



LDA:

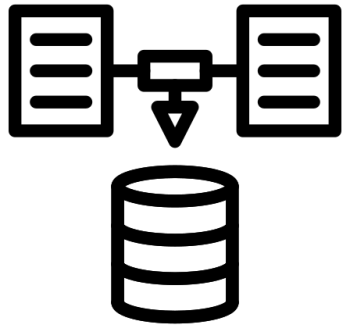
maximizing the component
axes for class-separation



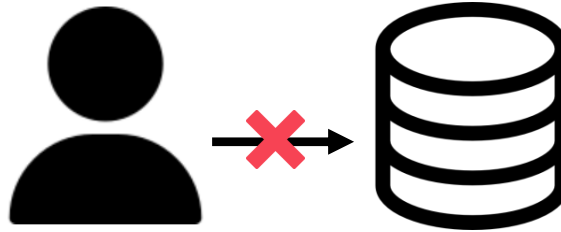
Fonte: https://sebastianraschka.com/Articles/2014_python_lda.html

Background Teórico

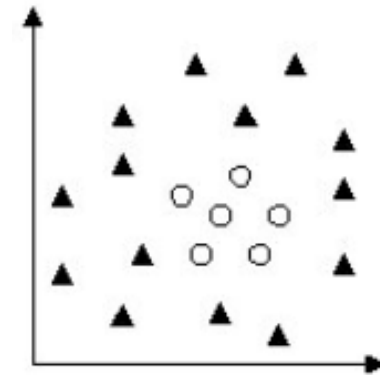
Kernel PCA



Extração de
características

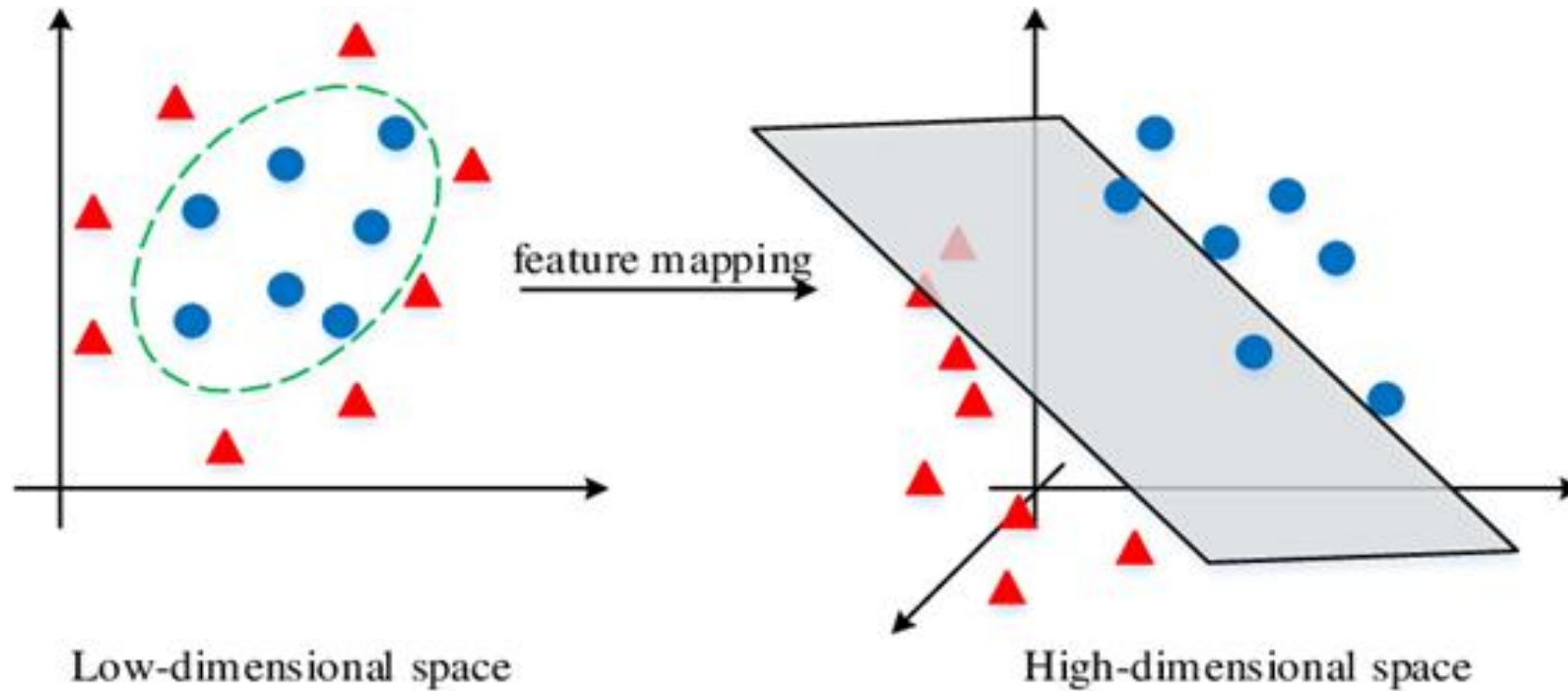


Aprendizagem não
supervisionada



Processamento de
dados não lineares

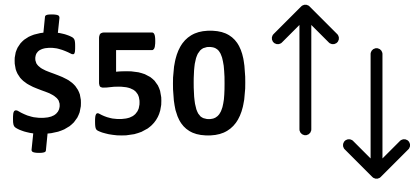
Background Teórico



Fonte: <https://www.semanticscholar.org/paper/Software-defect-prediction-based-on-kernel-PCA-and-Xu-Liu/fe246ded4da28ab5668f799fb08cb32a797c009e>

Procedimentos metodológicos

Bases de dados utilizadas



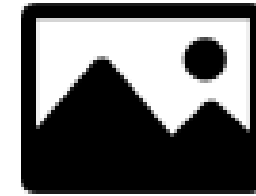
Adult data set

50 mil registros
15 atributos



House prices

22 mil registros
21 atributos



Imagens

200 imagens
31,6 MB

Procedimentos metodológicos

Hardware

- AMD Ryzen 7 5700G 3.80 GHz
- Memória RAM 32GB (2x16) DDR4 3600MHz
- SSD 512GB
 - Leitura 3100MBs
 - Gravação 1500MBs

Sotfwares

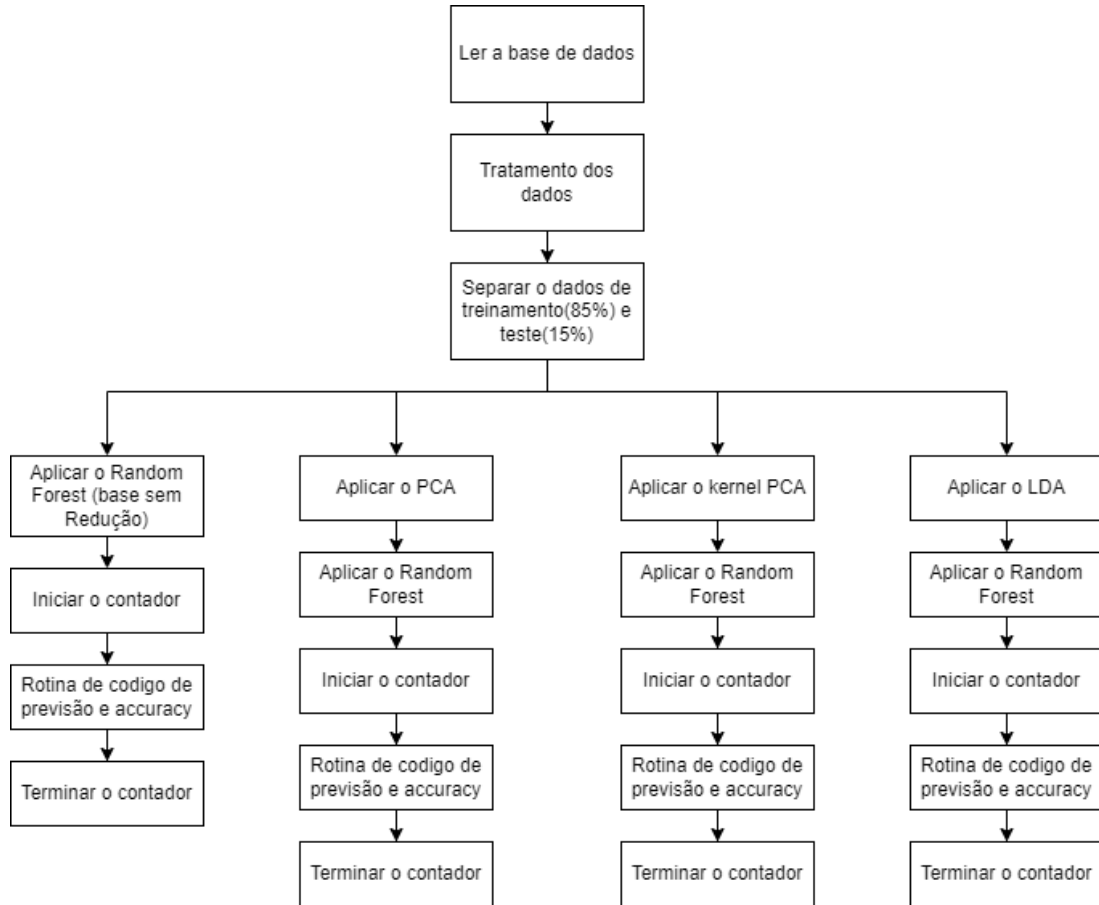
- Jupyter Notebook 6.5.2
- Python 3.10.2

Bibliotecas:

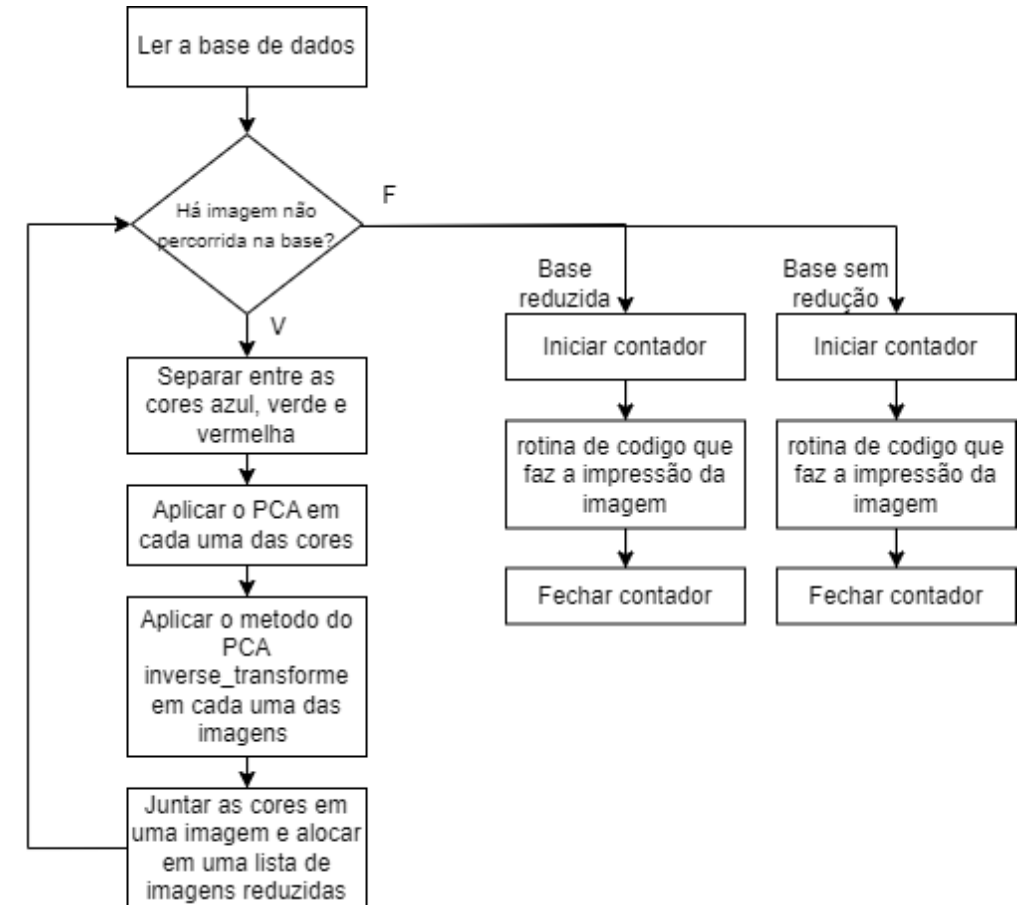
- Time
- Panda
- Seaborns
- Matplotlib
- Sklearn

Procedimentos metodológicos

Base de dados estruturado



Base de dados Não estruturado



Resultados

Bases de dados estruturado - Classificação

Base sem redução da dimensionalidade

Taxa de acerto	Tempo de execução
0.9404	2.563

Base reduzida pelo LDA

n_components	Taxa de acerto	Tempo de execução
1	0.8908	2.981

Base reduzida pelo PCA

n_components	Taxa de acerto	Tempo de execução
14	0.9404	2.341
12	0.9404	2.303
10	0.9404	2.244
8	0.9404	2.349
6	0.9404	2.401
4	0.9404	2.474
2	0.9404	2.821

Base reduzida pelo kernel PCA

n_components	Taxa de acerto	Tempo de execução
14	0.9396	2.201
12	0.9384	2.274
10	0.9355	2.311
8	0.9312	2.383
6	0.9304	2.424
4	0.9186	2.481
2	0.8972	2.868

Resultados

Bases de dados estruturado - Regressão

Base sem redução da dimensionalidade

Taxa de acerto	Erro	Tempo de execução
0.875	76620	274.181

Base reduzida pelo LDA

N components	Taxa de acerto	Erro	Tempo de execução
16	0.881	76.323	275.234
14	0.871	77.894	288.409
12	0.852	84.612	274.185
10	0.850	84.187	271.937
8	0.823	91.019	265.443
6	0.800	97.137	261.267
4	0.781	102.447	255.100
2	0.762	108.197	253.292

Base reduzida pelo PCA

N components	Taxa de acerto	Erro	Tempo de execução
16	0.881	74.811	274.498
14	0.875	77.400	278.630
12	0.865	78.189	268.266
10	0.829	89.368	264.268
8	0.813	89.841	257.877
6	0.763	105.411	253.703
4	0.666	135.690	252.057
2	0.632	143.166	245.153

Base reduzida pelo kernel PCA

N components	Taxa de acerto	Erro	Tempo de execução
16	0.716	90.505	305.011
14	0.705	95.932	297.465
12	0.688	101.872	269.198
10	0.666	102.646	273.449
8	0.659	107.767	277.718
6	0.597	119.540	260.827
4	0.530	141.316	248.274
2	0.463	159.849	237.248

Resultados

Bases de dados não estruturado - 200 Imagens

Base sem redução da dimensionalidade

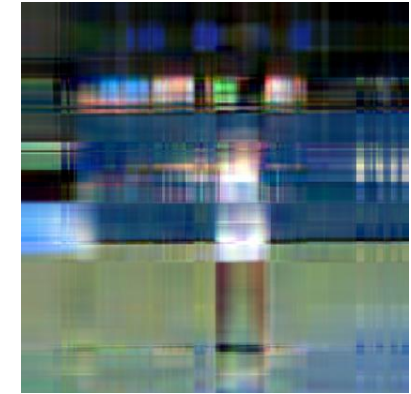
Tamanho em MB	Tempo de impressão
31,6	26.153

Base reduzida pelo PCA

N_components	Tamanho em MB	Tempo de impressão
100	4,45	20.521
50	4,21	19.842
20	3,68	17.987
5	2.80	17.130

Resultados

Analizando uma imagem




N_components	-	100	50	20	5
Armazenamento	139KB	36KB	35,5KB	32,3KB	25KB
Variância	-	0.987	0.965	0.913	0.776

Considerações finais

- Na base de classificação, o algoritmo **PCA** teve resultados melhores, não perdendo informações que faria o a precisão do Random Forest decrescer e aumentando a performance em 12.5% nos melhores dos seus parâmetros.
- Na base de regressão, o algoritmo mais eficaz foi o **LDA**, perdendo em média de 10.29% da precisão calculada pelo Random Forest, resultando melhor precisão em comparado com o PCA e Kernel PCA que teve uma perda de informação de 22.80% e 33.94% respectivamente
- O **kernel PCA** teve resultados satisfatórios na base de classificação, porém para haver total do seu desempenho, a base teria que ser não-linear, diferentes das bases testadas.
- Na base de imagens, o melhor resultado que tivemos foi com o parâmetro **PCA n_components = 100**, que teve uma diminuição de armazenamento de sua base de 85.9%, um aumento na velocidade da sua impressão de 21.6%, tudo isso perdendo somente 1.30% de suas informações.

Dúvidas?

 <https://www.linkedin.com/in/lucca-de-castro-machado/>

 lucca465@gmail.com

 Lucca de Castro Machado - 32292783

Referências

- ANOWAR, Farzana; SADAOU, Samira; SELIM, Bassant. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). Computer Science Review. Montreal, Canada, p. 1-13. jan. 2021.
- H., Telgaonkar Archana; SACHIN, Deshmukh. Dimensionality Reduction and Classification through PCA and LDA. International Journal Of Computer Applications. Aurangabad, p. 0975-8887. jul. 2015.
- Belarbi, Mohammed Amin & Saïd, Mahmoudi & Belalem, Ghalem. (2017). PCA as Dimensionality Reduction for Large-Scale Image Retrieval Systems. International Journal of Ambient Computing and Intelligence. 8. 14. 10.4018/IJACI.2017100104.