Universidade do Minho

Dados e Aprendizagem Automática

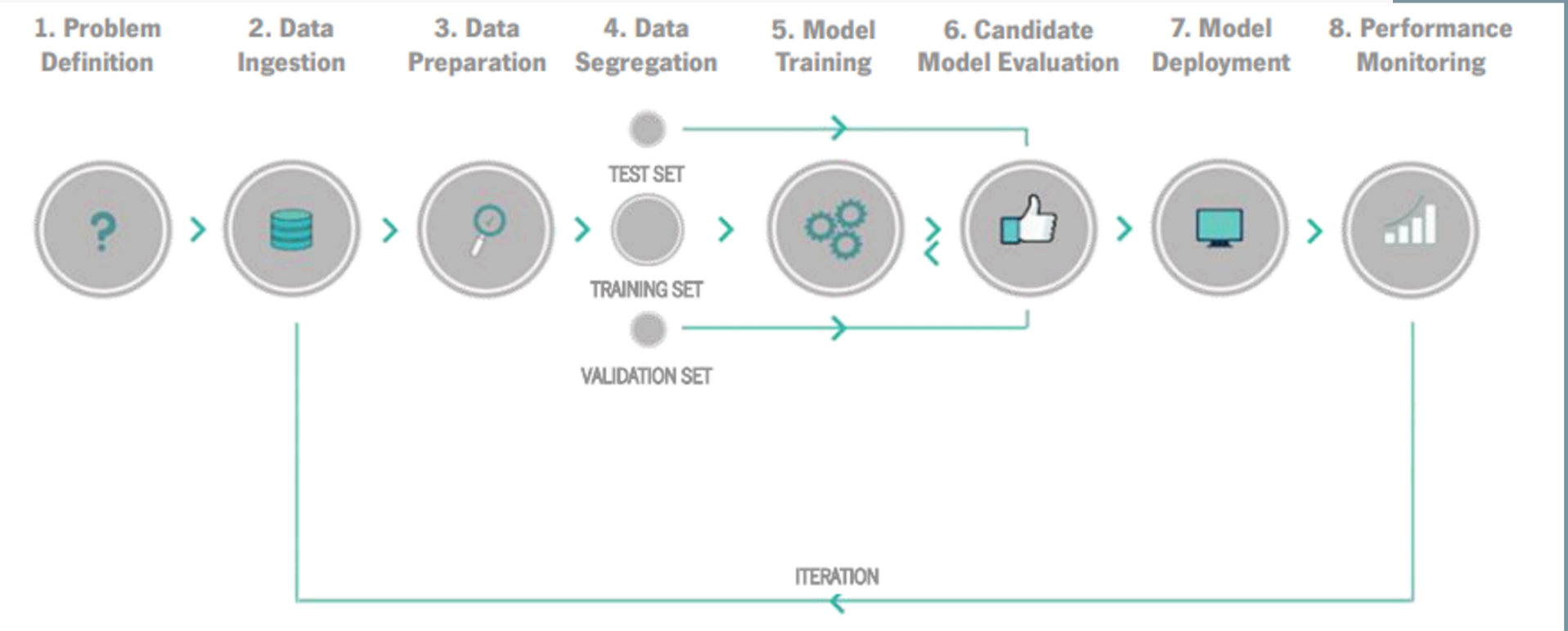
Eduardo Cunha

Jorge Rodrigues

João Magalhães

Rodrigo Gomes

Metodologia



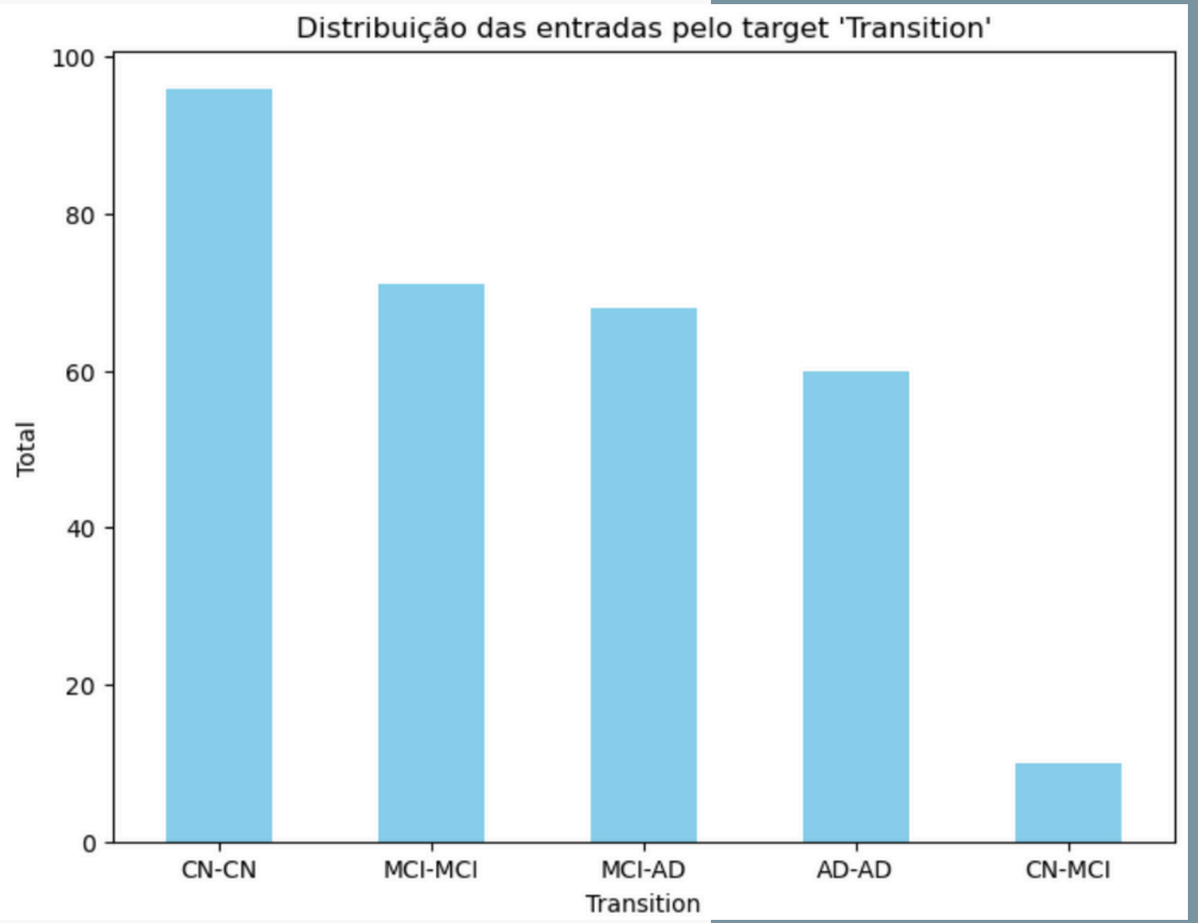
Data Exploration

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 305 entries, 0 to 304
Columns: 2181 entries, ID to Transition
dtypes: float64(2014), int64(147), object(20)
memory usage: 5.1+ MB
```

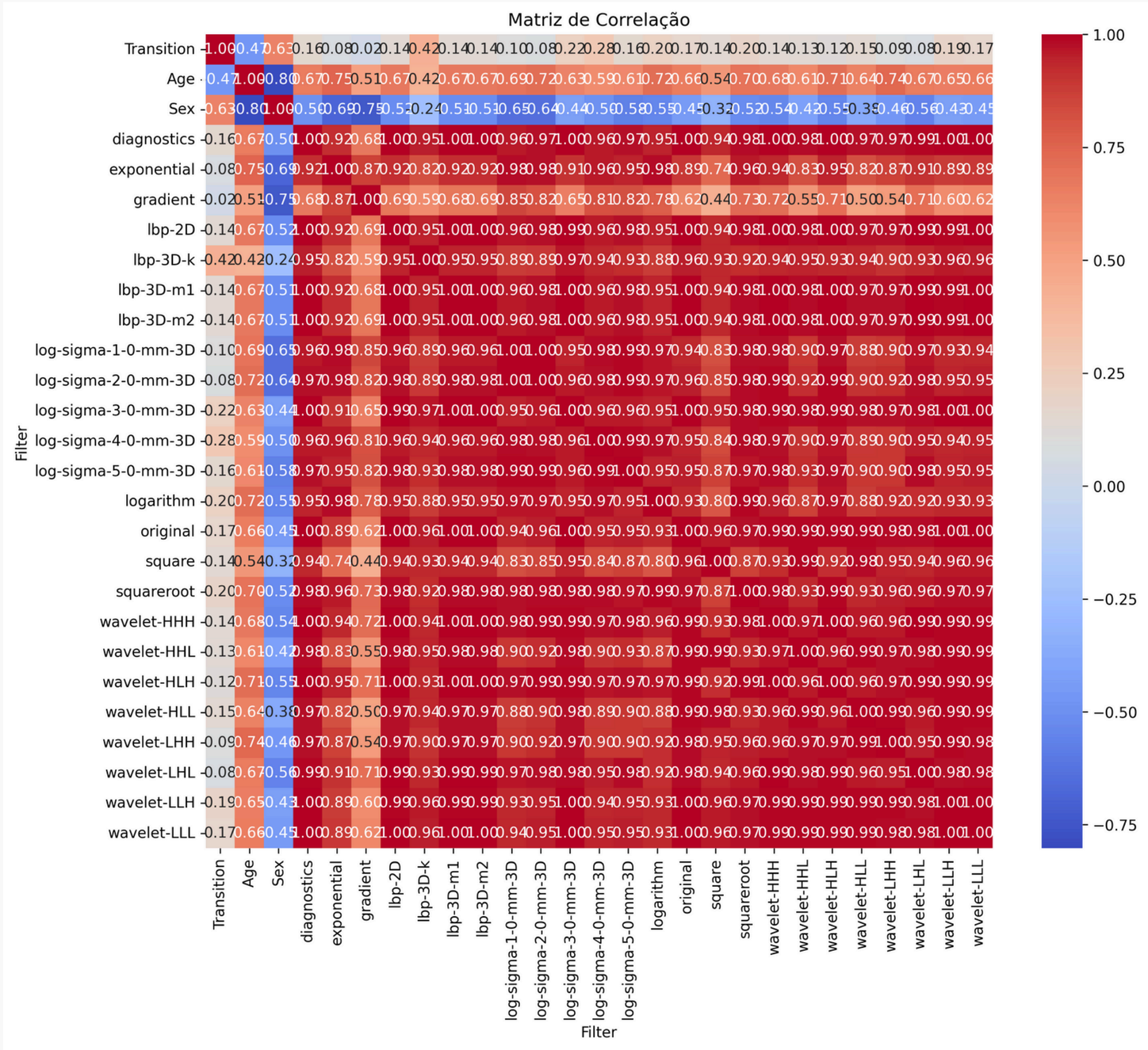
Não há entradas duplicadas

Não há valores em falta

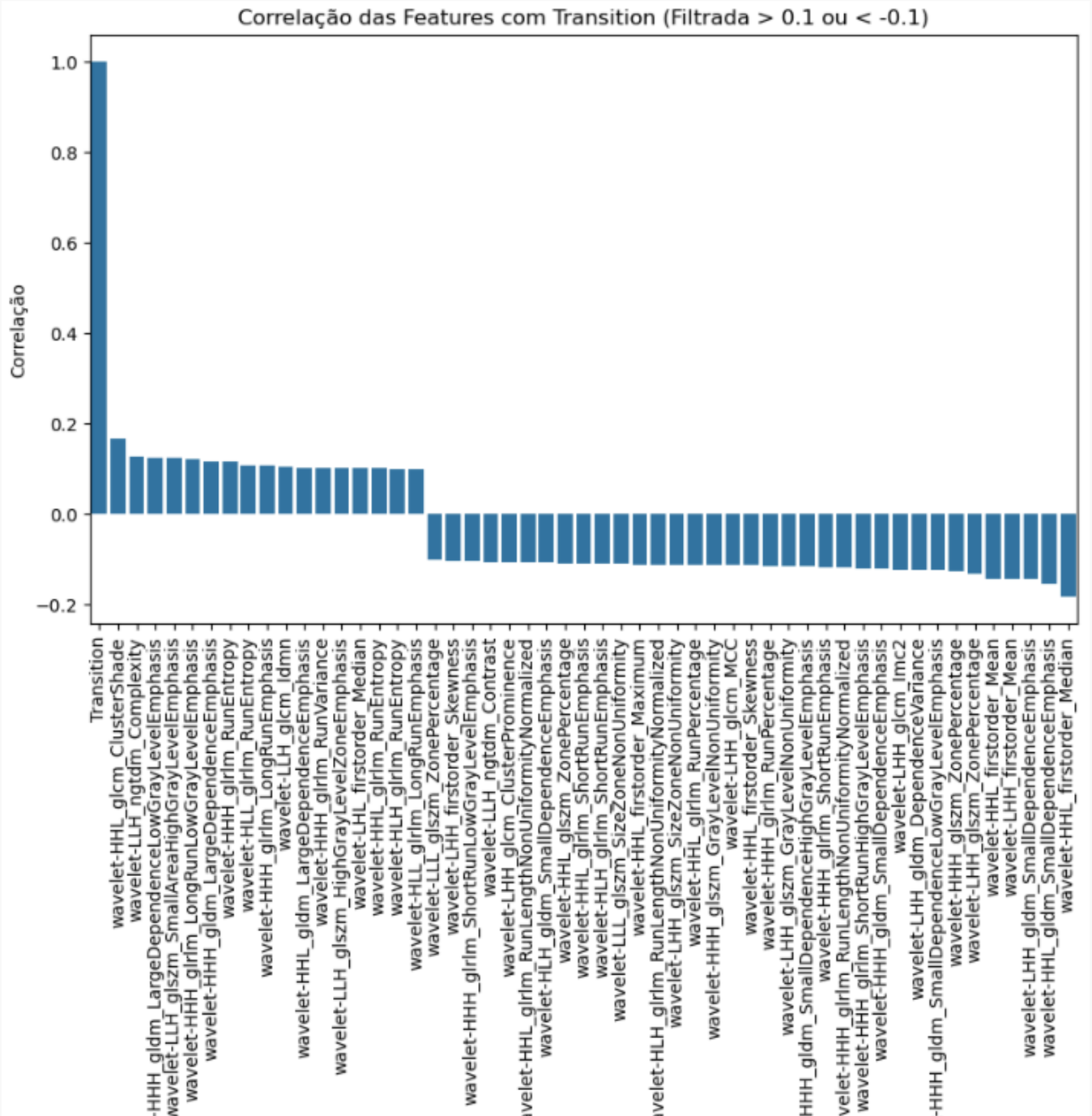
A grande maioria das colunas são numéricas



Visualização dos Dados



Visualização dos Dados



Preparações dos dados

Prep 1

- Drop de colunas identificadoras
- Drop a colunas não numéricas
- Normalização MinMax

Prep 2

- Drop de colunas identificadoras
- Drop a colunas não numéricas
- Drop a colunas constantes
- Outlier Removal Z-score
- Normalização Standard

Prep 3

- Drop de colunas identificadoras
- Drop a colunas não numéricas
- Drop a colunas constantes
- Normalização MinMax

Prep 4

- Drop de colunas identificadoras
- Drop a colunas não numéricas
- Normalização MinMax
- Feature Selection com PCA

Preparações dos dados

Prep 5

- Remoção de colunas não numéricas
- Remoção de colunas constantes
- Normalização MinMax
- Eliminação de Features com Cross-Validation

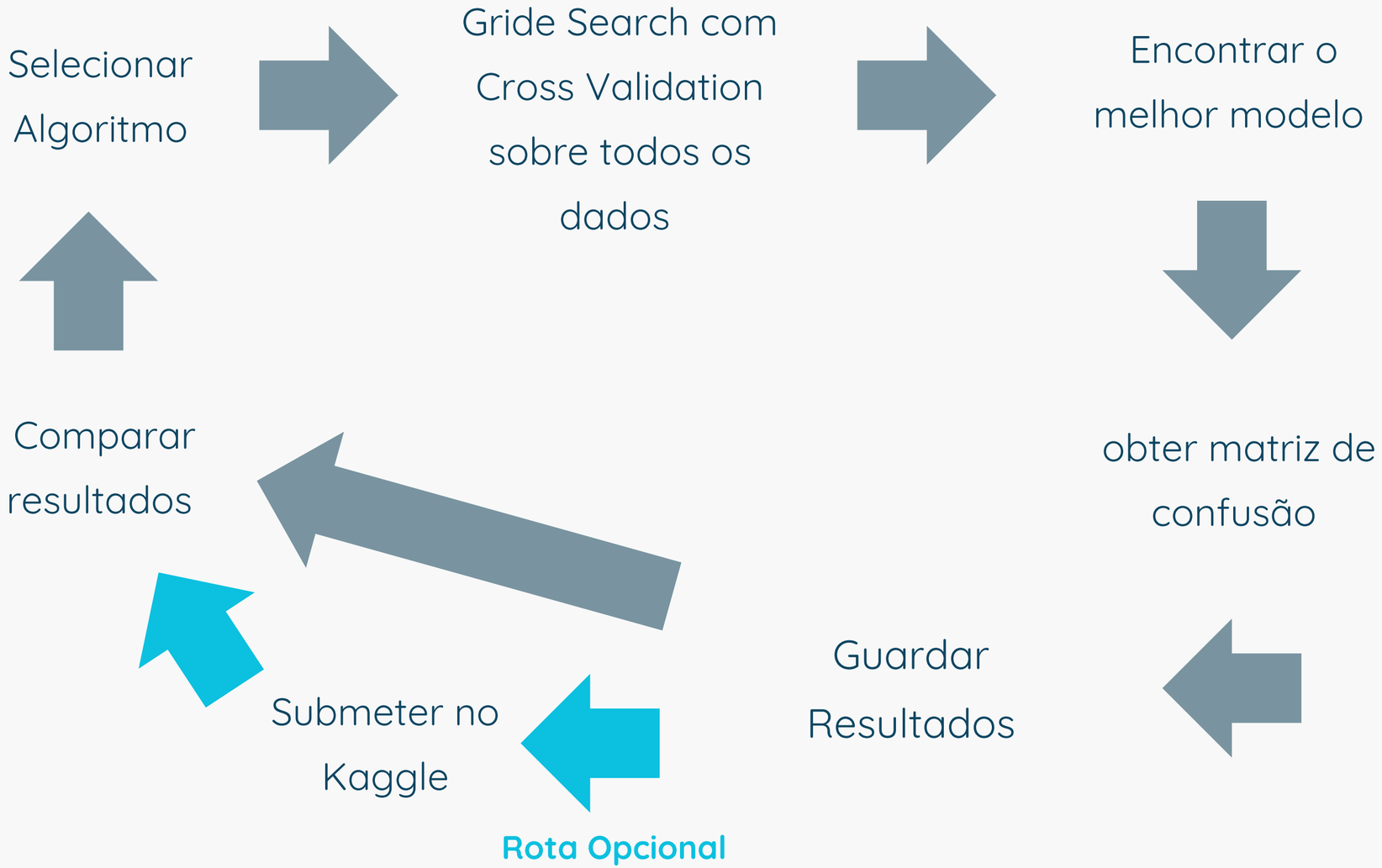
Prep 6

- Remoção de colunas não numéricas
- Remoção de colunas constantes
- Normalização MinMax
- Eliminação de Features com Cross-Validation
- SMOTE

Prep 7

- Remoção de colunas não numéricas
- Remoção de colunas constantes
- Feature Selection com ANOVA
- Normalização MinMax

Ciclo de Treino dos Modelos



Modelos

Essemblers

Bagging

Preparação	n_estimators	F1-macro	Kaggle	Diff
1	100	0.335723023	/	/
2	100	0.327265027	/	/
3	100	0.325427311	/	/
4	100	0.260252105	/	/
5	50	0.373371477	0.3251	0.048271
6	300	0.708276411	/	/
7	300	0.335320281	/	/

Random Forest

Prep	n_estimators	max_depth	criterion	max_features	F1-macro	Kaggle	Diff
1	100	5	entropy	log2	0.352970724	0.39298	0.040009276
2	300	5	entropy	None	0.3365891168	/	/
3	100	20	entropy	log2	0.336931552	0.31972	0.017211552
4	500	10	entropy	sqrt	0.277658465	/	/
5	50	20	gini	None	0.358217787	0.3504	0.007817787
6	500	20	entropy	sqrt	0.726276622	0.29079	0.435486622
7	300	20	entropy	sqrt	0.364177667	0.38099	0.024372333

Essemblers

Gradient Boosting

Prep	n_estimators	max_depth	learning_r	max_features	F1-macro	Kaggle	Diff
1	100	5	0.1	sqrt	0,36091994	0.2730	0,087909946
2	50	5	0.1	sqrt	0,337514053	/	/
3	50	5	0.3	none	0,34531770	0.40015	0,05483229
4	100	5	0.1	none	0,30322347	/	/
5	100	10	0.3	sqrt	0,3417128	0,09364276	/
6	100	10	0.3	sqrt	0,7353971	/	/
7	100	20	0.3	sqrt	0,3343648	/	/

XGBoosting

Preparação	n_estimators	max_depth	learning_rate	F1-macro	Kaggle	Diff
1	50	5	0.1	0.34849889	0.30851	0.03998800
2	300	5	0.1	0.32311066	/	/
3	50	5	0.1	0.34849809	0.31212	0.03637800
4	50	5	0.1	0.28788948	/	/
5	50	5	0.3	0.32751658	/	/
6	300	5	0.3	0.70815849	/	/
7	100	0	0.3	0.33275179	/	/

Essemblers

Stacking

Preparação	Meta model	Modelos	F1-macro	Kaggle	Diff
1	RandomForest	Rf, GB, SVM	0.324699587	0.34652	0.021820413
3	RandomForest	Rf, GB, SVM	0.305552525	/	/
4	RandomForest	Rf, GB, SVM	0.333704187	0.25241	0.081294187
5	RandomForest	Rf, GB, SVM	0.293556556	0.31103	0.017473444
6	RandomForest	Rf, GB, SVM	0.756487516	/	/
7	RandomForest	Rf, GB, SVM	0,332924057	0.3323	0.000624057

Maxvoting

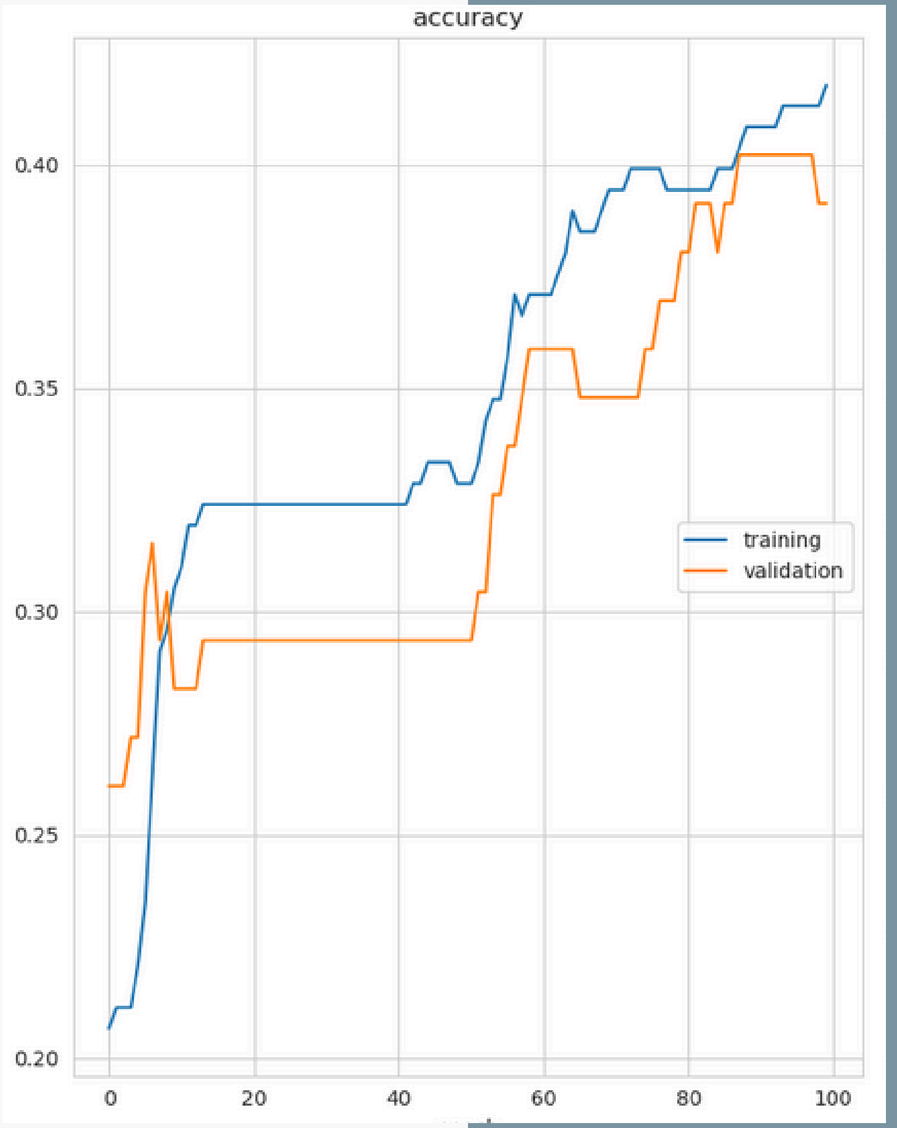
Preparação	Modelos	Pesos	F1-macro	Kaggle	Diff
1	GB, SVM, RF	[2,1,3]	0.338322464	/	/
3	GB, SVM, RF	[3,1,2]	0.345316348	0.27619	0.069126
4	GB, SVM, RF	[5,3,1]	0.303223475	/	/
5	GB, SVM, RF	[5,3,1]	0.357627429	0.25334	0.104287
6	GB, SVM, RF	[3,1,2]	0.7490927	0.3492	0.399893
7	GB, SVM, RF	[3,3,1]	0.342865921	/	/

SVM

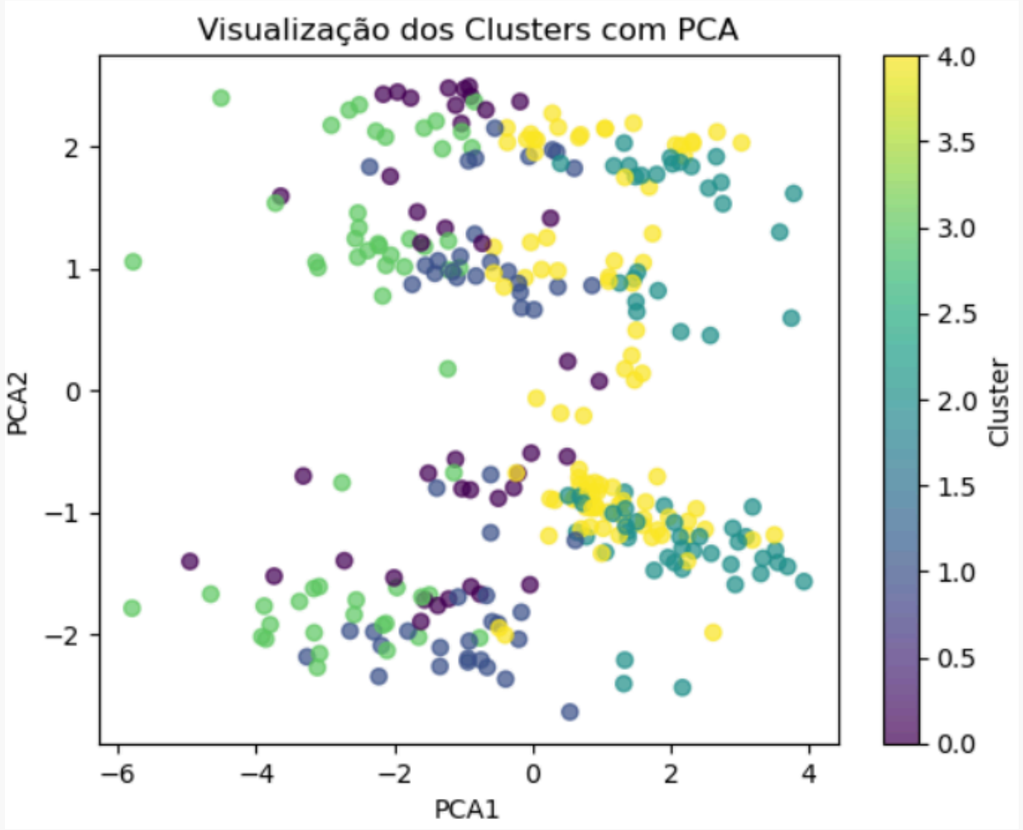
Preparação	C	kernel	gamma	F1-macro	Kaggle	Diff
1	1	linear	scale	0.314632233	/	/
3	1	rbf	scale	0.297462045	/	/
4	1	rbf	scale	0.320738023	/	/
5	100	sigmoid	scale	0.340245549	0.27993	0.060316
6	100	rbf	scale	0.744010658	/	/
7	1	rbf	scale	0.313844853	/	/

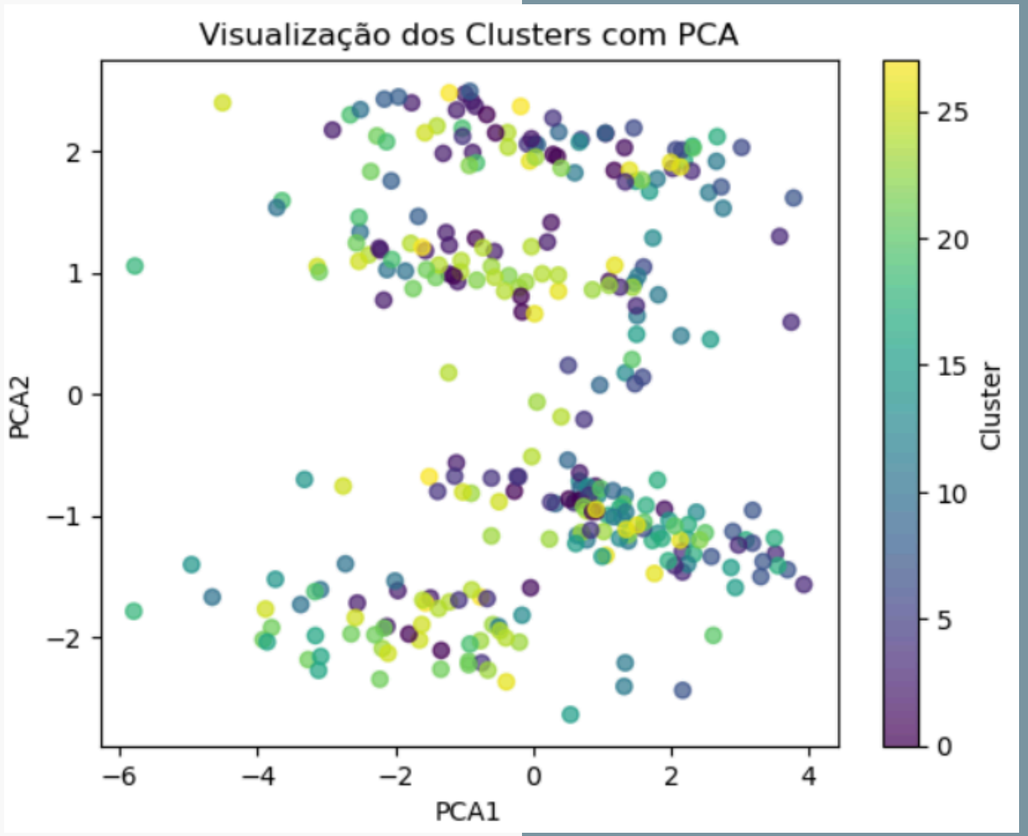
Neural Networks

```
==========================================================================
Layer (type:depth-idx)                   Output Shape              Param #
==========================================================================
MLP_1                                    [213, 5]                  --
├─Linear: 1-1                            [213, 400]                805,600
├─ReLU: 1-2                              [213, 400]                --
├─Linear: 1-3                            [213, 200]                80,200
├─ReLU: 1-4                              [213, 200]                --
├─Linear: 1-5                            [213, 5]                  1,005
├─Softmax: 1-6                           [213, 5]                  --
==========================================================================
```



Clustering


Visualização dos Clusters com PCA
PCA1
PCA2
Cluster
−6
−4
−2
0
2
4
2
1
0
−1
−2
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0


Visualização dos Clusters com PCA
PCA1
PCA2
Cluster
−6
−4
−2
0
2
4
2
1
0
−1
−2
25
20
15
10
5
0

Exploitation

- **Exprimentar com novos hiperparâmetros:**Testámos ajustes mais variados
- **Novas Técnicas:** Explorámos AdaBoost e ExtraTreesClassifier
- **Técnicas para dados desbalanceados:** Métodos como EasyEnsemble e BalancedRandomForest
- **Stacking e Voting:** Resultados inconsistentes
- **Over sampling e Under sampling:** SMOTE e outros métodos geraram ruído ou perda de informação, afetando a performance.

Comparacao de modelos e Preparações de dados

Os modelos de ensemble destacaram-se claramente, com o Random Forest (preparação 5) a obter o melhor F1-score macro, e o Gradient Boost (preparação 3) a ter maior robustez nas submissões públicas do Kaggle.

As preparações 3 e 5 foram as mais eficazes globalmente.

Dificuldades Identificadas:

- Classes desbalanceadas afetaram negativamente as classes minoritárias.
- Overfitting prejudicou a generalização de alguns modelos.
- Dados limitados com muitos atributos reduziram o desempenho geral.

Desempenho no Kaggle:

- **Testes Públicos:** 29ª posição.
- **Testes Privados:** 26ª posição, com o Random Forest a demonstrar boa generalização.

Prep	n_estimators	max_depth	criterion	max_features	F1-macro	Kaggle	Diff
5	50	20	gini	None	0.358217787	0.3504	0.007817787

- A melhor submissão alcançou um F1 de 0,41 (6ª posição), mas foi

Prep	n_estimators	max_depth	learning_r	max_features	F1-macro	Kaggle	Diff
3	50	5	0.3	none	0,34531770	0.40015	0,05483229

Conclusão

Os objetivos foram atingidos, mas a principal limitação foi o reduzido dataset (100 entradas), que dificultou a criação de modelos robustos e generalizáveis, especialmente em redes neuronais.

Diversas técnicas de preparação e modelação foram aplicadas, como normalização e seleção de variáveis, resultando em progressos promissores. Contudo, a ampliação do dataset é crucial para melhorar a eficácia dos modelos, especialmente os mais complexos.

Este trabalho oferece uma base sólida para futuras investigações com dados mais abrangentes.

Universidade do Minho

Dados e Aprendizagem Automática

Eduardo Cunha

Jorge Rodrigues

João Magalhães

Rodrigo Gomes