A3 Inteligência Artificial

Classificação de phishing URLs

Matheus Pereira Silva
Nilso José Miguel da Silva Júnior

Mentores:
Adriana Neves
Vitor Leães

O Que é Phishing?

Phishing é um tipo de ciberataque usado para enganar usuários, fazendo com que, forneçam informações como:

- ➔ Senhas
- ➔ Nome de Usuário
- ➔ Cartão de Crédito
- ➔ Outros Tipos de Data Pessoal

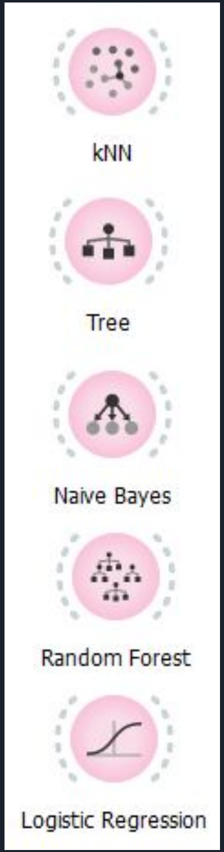
Estrutura do Projeto

- 4 datasets
- Mais de 80 features
- Orange
- 5 Algoritmos
- Identificar Phishing URLs
- Principais Fatores

1	length_url	N	numeric	feature
2	length_hostname	N	numeric	feature
3	ip	C	categorical	feature
4	nb_dots	N	numeric	feature
5	nb_hyphens	N	numeric	feature

Algoritmos

- kNN
- Árvore de Decisão
- Naive Bayes
- Random Forest
- Regressão Logística



Pré-Processamento

- ➔ Datasets não contém missing data
- ➔ Pré-Processamento apenas no kNN
- ➔ Métrica de normalização

Normalize Features

Standardize to $\mu=0$, $\sigma^2=1$

Exemplo No Dataset 3

Sem Pré-Processamento

Model	AUC	CA	F1	Prec	Recall	MCC
Unprocessed kNN	0.916	0.851	0.851	0.851	0.851	0.701

Predicted

	0	1	Σ
0	85.5 %	15.3 %	56000
1	14.5 %	84.7 %	61290
Σ	54281	63009	117290

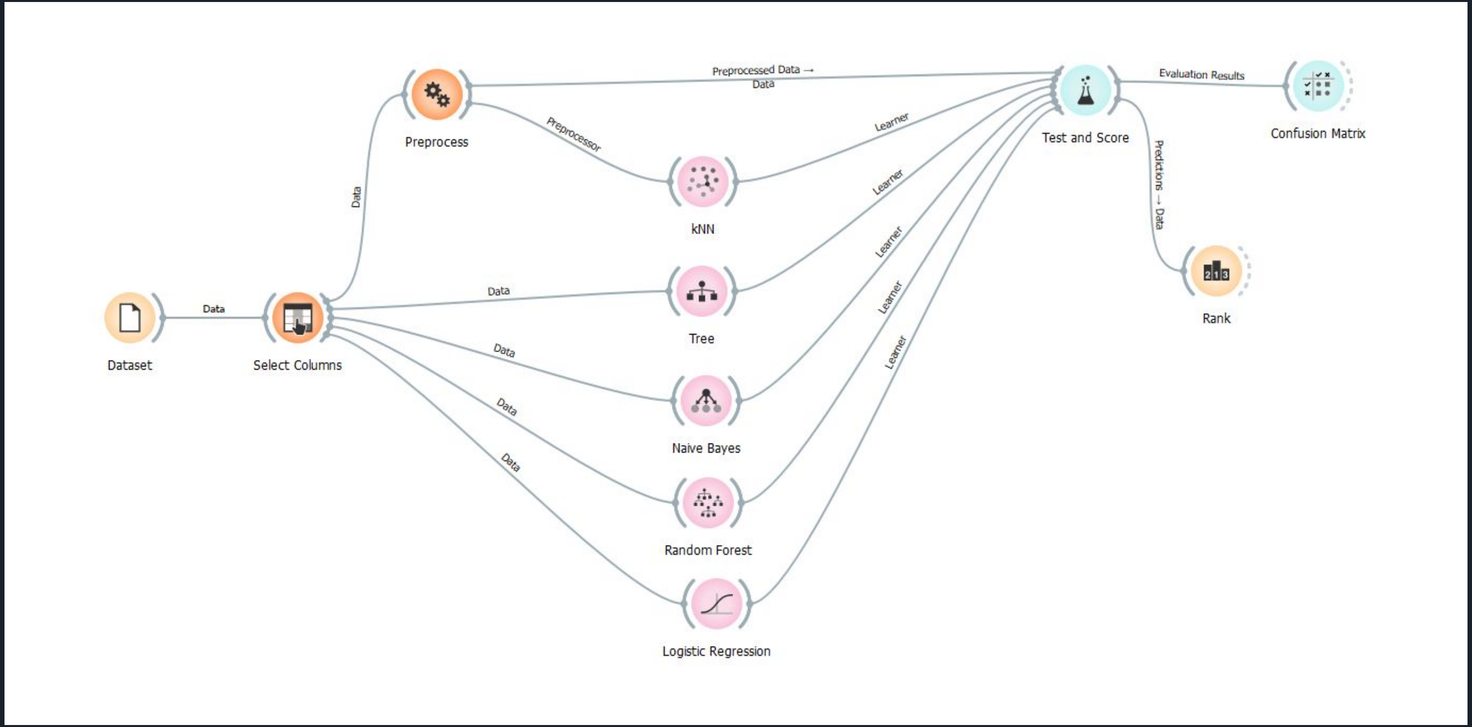
Com Pré-Processamento

Model	AUC	CA	F1	Prec	Recall	MCC
kNN	0.973	0.932	0.932	0.932	0.932	0.863

Predicted

	0	1	Σ
0	92.8 %	6.5 %	56000
1	7.2 %	93.5 %	61290
Σ	56085	61205	117290

Orange


Preprocessed Data → Data
Evaluation Results
Preprocess
Test and Score
Confusion Matrix
Preprocessor
Learner
Learner
Learner
Learner
Learner
Data
kNN
Predictions → Data
Data
Data
Tree
Rank
Dataset
Select Columns
Data
Data
Naive Bayes
Data
Random Forest
Logistic Regression

Dataset 1

Resultados

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	1.000	1.000	1.000	1.000	1.000	1.000
Logistic Regression	1.000	1.000	1.000	1.000	1.000	1.000
Naive Bayes	1.000	0.999	0.999	0.999	0.999	0.998
kNN	1.000	0.999	0.999	0.999	0.999	0.998
Tree	0.996	0.997	0.997	0.997	0.997	0.993

Dataset 2

Resultados

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	1.000	1.000	1.000	1.000	1.000	1.000
Random Forest	1.000	0.999	0.999	0.999	0.999	0.998
Naive Bayes	1.000	0.998	0.998	0.998	0.998	0.996
Logistic Regression	1.000	0.996	0.996	0.996	0.996	0.993
kNN	0.996	0.982	0.982	0.982	0.982	0.965

Dataset 3

Resultados

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.988	0.952	0.952	0.952	0.952	0.904
Tree	0.927	0.936	0.936	0.936	0.936	0.871
kNN	0.973	0.932	0.932	0.932	0.932	0.863
Logistic Regression	0.961	0.901	0.901	0.901	0.901	0.801
Naive Bayes	0.924	0.797	0.789	0.838	0.797	0.629

Dataset 4

Resultados

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.989	0.958	0.958	0.958	0.958	0.915
Logistic Regression	0.985	0.945	0.945	0.945	0.945	0.891
kNN	0.979	0.942	0.942	0.943	0.942	0.885
Tree	0.928	0.939	0.939	0.939	0.939	0.878
Naive Bayes	0.959	0.894	0.894	0.894	0.894	0.788

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	1.000	1.000	1.000	1.000	1.000	1.000
Logistic Regression	1.000	1.000	1.000	1.000	1.000	1.000
Naive Bayes	1.000	0.999	0.999	0.999	0.999	0.998
kNN	1.000	0.999	0.999	0.999	0.999	0.998
Tree	0.996	0.997	0.997	0.997	0.997	0.993

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	1.000	1.000	1.000	1.000	1.000	1.000
Random Forest	1.000	0.999	0.999	0.999	0.999	0.998
Naive Bayes	1.000	0.998	0.998	0.998	0.998	0.996
Logistic Regression	1.000	0.996	0.996	0.996	0.996	0.993
kNN	0.996	0.982	0.982	0.982	0.982	0.965

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.988	0.952	0.952	0.952	0.952	0.904
Tree	0.927	0.936	0.936	0.936	0.936	0.871
kNN	0.973	0.932	0.932	0.932	0.932	0.863
Logistic Regression	0.961	0.901	0.901	0.901	0.901	0.801
Naive Bayes	0.924	0.797	0.789	0.838	0.797	0.629

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.989	0.958	0.958	0.958	0.958	0.915
Logistic Regression	0.985	0.945	0.945	0.945	0.945	0.891
kNN	0.979	0.942	0.942	0.943	0.942	0.885
Tree	0.928	0.939	0.939	0.939	0.939	0.878
Naive Bayes	0.959	0.894	0.894	0.894	0.894	0.788

Algorithm	Features	Precision	Sensitivity	F-Measure	Accuracy
Decision Tree	NLP Features	0.964	0.977	0.971	**97.02%**
	Word Vector	0.944	0.695	0.800	82.48%
	Hybrid	0.933	0.973	0.953	95.14%
Adaboost	NLP Features	0.908	0.963	0.935	**93.24%**
	Word Vector	0.936	0.536	0.682	74.74%
	Hybrid	0.915	0.940	0.927	92.53%
Kstar	NLP Features	0.936	0.936	0.936	93.56%
	Word Vector	0.845	0.811	0.806	81.05%
	Hybrid	0.953	0.953	0.953	**95.27%**
kNN ($k = 3$)	NLP Features	0.940	0.977	0.958	95.67%
	Word Vector	0.955	0.697	0.806	83.01%
	Hybrid	0.946	0.974	0.960	**95.86%**
Random Forest	NLP Features	0.970	0.990	0.980	**97.98%**
	Word Vector	0.958	0.697	0.807	83.14%
	Hybrid	0.953	0.976	0.964	96.36%
SMO	NLP Features	0.928	0.975	0.951	**94.92%**
	Word Vector	0.947	0.697	0.803	82.71%
	Hybrid	0.923	0.972	0.947	94.48%
Naive Bayes	NLP Features	0.940	0.977	0.958	95.67%
	Word Vector	0.955	0.697	0.806	83.01%
	Hybrid	0.946	0.974	0.960	**95.86%**

Dataset 1

Fatores Mais Importantes

		#	Gain ratio	Gini
1	N URLSimilarityIndex		0.683	0.483
2	C HasSocialNet	2	0.559	0.301
3	C HasCopyrightInfo	2	0.463	0.271
4	C IsHTTPS	2	0.433	0.181
5	C HasDescription	2	0.405	0.232

Dataset 2

Fatores Mais Importantes

		#	Gain ratio	Gini
1	N id		0.500	0.500
2	N PctExtNullSelfRedirectHyperlinksRT		0.231	0.175
3	C FrequentDomainNameMismatch	2	0.229	0.104
4	N PctExtHyperlinks		0.225	0.272
5	C SubmitInfoToEmail	2	0.202	0.064

Dataset 3

Fatores Mais Importantes

		#	Gain ratio	Gini
1	N qty_questionmark_directory		0.377	0.198
2	N qty_hashtag_directory		0.377	0.198
3	N qty_slash_file		0.377	0.198
4	N qty_questionmark_file		0.377	0.198
5	N qty_hashtag_file		0.377	0.198

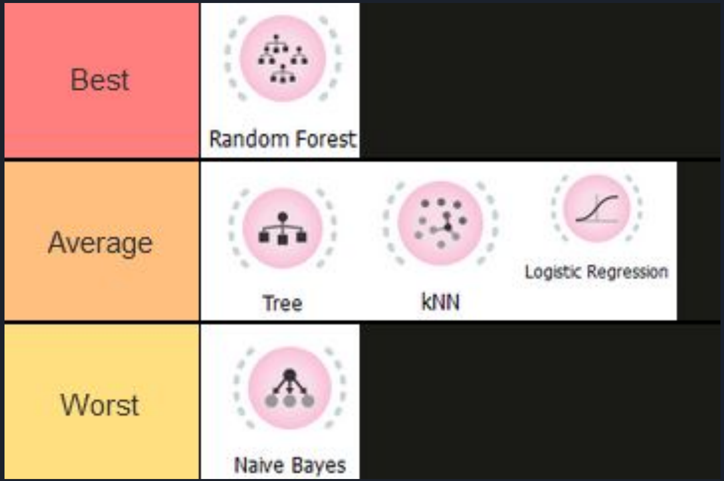
Dataset 4

Fatores Mais Importantes

		#	Gain ratio	Gini
1	C google_index	2	0.427	0.263
2	N nb_www		0.152	0.102
3	N nb_at		0.141	0.011
4	N page_rank		0.140	0.162
5	C ip	2	0.136	0.053

Melhores Algoritmos

1. Random Forest
2. Regressão Logística
3. kNN
4. Árvore de Decisão
5. Naive Bayes



Principais Fatores

1. Caracteres Especiais
2. Hyperlinks
3. Falta de HTTPS
4. Nome do Domínio

Referências

- Prasad, A., & Chandra, S. (2023). PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning. Computers & Security, 103545. doi: https://doi.org/10.1016/j.cose.2023.103545
- Tan, Choon Lin (2018), "Phishing Dataset for Machine Learning: Feature Evaluation", **Mendeley Data**, V1, doi: 10.17632/h3cgnj8hft.1
- G. Vrbančič, I. Jr. Fister, V. Podgorelec. Datasets for Phishing Websites Detection. Data in Brief, Vol. 33, 2020, DOI: 10.1016/j.dib.2020.106438
- Hannousse, Abdelhakim; Yahiouche, Salima (2021), "Web page phishing detection", **Mendeley Data**, V3, doi: 10.17632/c2gw7fy2j4.3
- VRBAN, Grega; FISTER JR, Iztok; PODGORELEC, Vili. Datasets for phishing websites detection. **ScienceDirect,** 2020. Disponível em: https://www.sciencedirect.com/science/article/pii/S2352340920313202. Acesso em: 03 dez. 2024.
- SALVIANO, Edgard Mesquita; SANTOS, João Pedro Ribeiro; SILVA, Matheus Almeida. Principais tipos de ataques Phishing e mecanismos de segurança. **Uniceplac,** 2021. Disponível em: https://dspace.uniceplac.edu.br/bitstream/123456789/1611/1/Edgard%20Mesquita%20Salviano_%20Jo%C3%A3o%20Pedro%20Ribeiro%20Santos_Matheus%20Almeida%20e%20Silva.pdf. Acesso em: 05 dez. 2024.
- ALEROUD, Ahmed; ZHOU, Lina. Phishing environments, techniques, and countermeasures: A survey. **ScienceDirect,** 2017. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0167404817300810. Acesso em: 05 dez. 2024.
- SAHINGOZ, Ozgur Koray; BUBER, Ebubekir; DEMIR, Onder; DIRI, Banu. Machine learning based phishing detection from URLs. **ScienceDirect,** 2019. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417418306067?casa_token=wm6Tsg9Go5UAAAAA:DEVAVOKRvEcDp0NruzpwMtKO3-eaasvILcPgCGMlHGnQQkeS2SXdC5P6T8hXqG3c9Cfu9Egn55U#tbl0005. Acesso em: 05 dez. 2024.