

Mata Kuliah - Penggalan Data

Nama Kelompok :

Anggota :

[202110370311222– Ibnu Fauzan Rachmadhanu]

[202110370311234 – Muhammad Wahyudi]

[202110370311241 – Abd Baasithur Rizqu]

Berikut ini merupakan update template laporan Mini Project kuliah Penggalan Data.

Nilai Total: 120 poin

Tahap 0 (poin: 25): Business Objective

Meningkatkan pertahanan keamanan siber dengan memberikan wawasan yang membantu dalam deteksi dan mitigasi oleh ancaman phishing. Analisis URL dengan menggunakan teknik Mutual Information dan logistic regression dalam penelitian. Identifikasi fitur fitur paling informatif untuk membedakan upaya phishing, memungkinkan memperkuat pertahanan dan tetap beada di depan takrik phishing yang tersu berkembang.

Tahap 1 (poin: 25): Original Data

- PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning.
- Data yang digunakan.
 - o Deskripsi singkat.

Serangan phishing melalui URL yang menipu menjadi masalah yang signifikan dalam lingkungan digital saat ini. Untuk mengatasi hal ini, diperlukan kerangka kerja pendeteksian yang efektif dan efisien. Artikel ini memperkenalkan PhiUSIIL, kerangka kerja pendeteksian URL phishing berdasarkan Indeks Kemiripan dan Pembelajaran Inkremental.

- o Sebutkan dan jelaskan atribut pada data tersebut.

No	Fitur	Deskripsi	TypeData
1	FILENAME	Nama pemberian pada sebuah file untuk mengidentifikasi dan membedakannya dari file lainnya.	Objek
2	URL	Referensi resource web yang diatur oleh jaringan komputer.	Objek
3	URL LENGTH	URL phishing sering kali menunjukkan kecenderungan untuk lebih panjang dibandingkan dengan URL yang sah.	Integer
4	Domain	Alamat yang perlu diakses untuk membuka dan mengakses website.	Objek
5	Domain Length	Jumlah karakter dalam sebuah domain internet, khususnya bagian nama di antara "www." dan ekstensi domain seperti ".com" atau ".org".	Integer
6	IsDomainIP	URL atau alamat IP digunakan sebagai nama domain.	Integer
7	TLD	Bagian akhir dari nama domain	Objek
8	URLSimilaratyIndex	Indeks untuk engukur kemiripan antara URL.	Float
9	CharContinuationRate	Indeks untuk mengukur kemiripan antara kontinuitas pendidikan atau keluhan beberapa kriteria.	Float
10	TLDLegitimateProb	Indeks untuk mengukur kemiripan antara legitimasinya dari top level Domain (TLD).	Float

11	URLCharProb	Indeks untuk mengukur kemiripan tingkat legitimasinya dari top level domain (TLD) yang digunakan dalam suatu website.	Float
12	TLDLength	Indeks untuk mengukur TLD dalam URL.	Integer
13	NoOfSubDomain	Peretas sering menggunakan teknik kemiripan visual untuk menipu pengguna. Mereka membuat subdomain yang sangat mirip dengan situs web yang sah.	Integer
14	HasObfuscation	Teknik untuk menyembunyikan atau mengaburkan kode, data, atau informasi lainnya agar sulit dipahami atau dideteksi oleh pihak yang tidak berwenang.	Integer
15	NoOfObfuscatedChar	Menampilkan jumlah karakter yang dikaburkan dalam URL.	Integer
16	ObfuscationRatio	Metrik untuk mengukur tingkat pengaburan atau penyembunyian kode atau informasi dalam suatu program.	Float
17	NoOfLettersInURL	Metrik pengukur jumlah huruf yang terkandung dalam URL.	Integer
18	LetterRatioInURL	Metrik yang mengukur rasio huruf terhadap total karakter dalam URL.	Float

19	DegitRatioInURL	Metrik yang mengukur rasio digit terhadap total karakter dalam URL.	Integer
20	NoOfDegitsInURL	Metrik yang mengukur jumlah digit atau angka terhadap total karakter dalam URL.	Integer
21	NoOfEqualsInURL	Matrik untuk mengukur jumlah sama dengan “=” dalam URL.	Integer
22	NoOfQMarkInURL	Matrik untuk mengukur jumlah tanda tanya “?” dalam URL.	Integer
23	NoOfAmpersandInURL	Matrik untuk mengukur jumlah tanda ampersand “&” dalam URL.	Integer
24	NoOfOtherSpecialCharsInURL	Matrik untuk mengukur jumlah jumlah karakter khusus selain tanda tanya “?” dan tanda ampersand “&” dalam URL.	Integer
25	SpacialCharRatioInURL	Matrik untuk mengukur rasio karakter khusus terhadap total karakter dalam URL.	Float
26	IsHTTPS	HTTP aman	Integer
27	LineOfCode	Matrik untuk mengukur jumlah total baris kode dalam suatu program.	Integer
28	LargestLineLength	Kode mungkin lebih panjang, teknik yang digunakan oleh peretas untuk menyembunyikan kegiatan mereka	Integer

29	HasTitle	Sebagian besar situs web asli menyediakan judul halaman.	Integer
30	Title	Istilah yang merujuk pada judul.	Objek
31	DomainTitleMatchScore	Metrik yang digunakan untuk mengukur sejauh mana domain cocok atau relevan.	Float
32	URLTitleMatchScore	Metrik yang digunakan untuk mengukur sejauh mana URL cocok atau relevan.	Float
33	HasFavicon	Sebagian besar situs web asli menyertakan logo situs web mereka dalam tag favicon.	Integer
34	Robots	Merujuk pada file teks khusus yang disebut "robots.txt"	Integer
35	IsResponsive	Situs web dirancang untuk menjadi responsif.	Integer
36	NoOfURLRedirect	Situs phishing dapat mengarahkan pengguna ke halaman yang berbeda.	Integer
37	NoOfSelfRedirect	Matrix yang digunakan dalam analisis web untuk mengukur jumlah redirect.	Integer
38	HasDescription	Situs web terkemuka menggabungkan deskripsi halaman dengan menggunakan nama meta 'deskripsi' untuk setiap halaman mereka.	Integer
39	NoOfPopup	pop-up atau iframe dapat digunakan dalam situs web phishing.	Integer
40	NoOfiFrame	pop-up atau iframe dapat digunakan dalam situs web phishing.	Integer

41	HasExternalFormSubmit	Situs phishing sering kali menggunakan formulir HTML untuk mengumpulkan informasi pengguna	Integer
42	HasSocialNet	informasi hak cipta dan tautan ke profil jejaring sosial dapat disertakan dalam situs web yang sebenarnya.	Integer
43	HasSubmitButton	menentukan halaman memiliki tombol kirim (submit).	Integer
44	HasPasswordField	analisis web untuk menentukan apakah halaman memiliki kolom masukan (input) untuk kata sandi (password) atau tidak.	Integer
45	HasHiddenFields	menentukan halaman memiliki kolom masukan hidden fields atau tidak.	Integer
46	Bank	layanan untuk menyimpan, memberikan pinjaman, dan transaksi.	Integer
47	Pay	proses mentransfer uang atau nilai keuangan lainnya dari satu pihak ke pihak lain sebagai ganti jasa, barang atau kewajiban lainnya.	Integer
48	Crypto	pengamanan untuk melindungi data serta komunikasi dalam sistem komputer dan jaringan.	Integer
49	HasCopyrightInfo	informasi hak cipta dan tautan ke profil jejaring sosial dapat disertakan dalam situs web yang sebenarnya.	Integer
50	NoOfImage	Jumlah total gambar atau grafik yang terdapat dalam suatu konteks tertentu.	Integer
51	NoOfCSS	Jumlah total (Cascading Style Sheets CSS) yang digunakan dalam pengembangan web atau aplikasi.	Integer

52	NoOfJS	Mengacu pada total file JavaScript yang digunakan.	Integer
53	NoOfSelfRef	Jumlah referensi atau rujukan ke entitas itu sendiri pada dataset.	Integer
54	NoOfEmptyRef	Jumlah referensi yang tidak memiliki nilai atau tidak merujuk ke entitas atau data yang konkret.	Integer
55	NoOfExternalRef	Merujuk pada jumlah total referensi atau koneksi yang mengarah ke entitas atau sumber daya diluar sistem, dokumen atau dataset.	Integer
56	Label	Penanda yang digunakan untuk mengidentifikasi atau mengkategorikan sesuatu.	Integer

o Jelaskan data mining task yang akan digunakan (*classification, clustering, regression, association rule mining, anomaly detection, dsb.*).

1. Classification. Dipilih sebagai tugas untuk data mining, digunakan yang nantinya dibuat pengelompokan class atau kategori dari atribut yang dipilih. kemudian dari hasil classifikasi dapat dipilih apakah sebuah website termasuk dalam kategori phishing atau bukan melalui kolom yang dikelompokkan sebelumnya. Pada tahapan klasifikasi, nantinya lebih fiutamakan data type integer atau float.

Dataframe awal berisi sebagai berikut:

#	Column	Non-Null Count	Dtype
0	FILENAME	22334 non-null	object
1	URL	22334 non-null	object
2	URLLength	22334 non-null	int64

3	Domain	22334 non-null object
4	DomainLength	22334 non-null int64
5	IsDomainIP	22334 non-null int64
6	TLD	22334 non-null object
7	URLSimilarityIndex	22334 non-null float64
8	CharContinuationRate	22334 non-null float64
9	TLDLegitimateProb	22334 non-null float64
10	URLCharProb	22333 non-null float64
11	TLDLength	22333 non-null float64
12	NoOfSubDomain	22333 non-null float64
13	HasObfuscation	22333 non-null float64
14	NoOfObfuscatedChar	22333 non-null float64
15	ObfuscationRatio	22333 non-null float64
16	NoOfLettersInURL	22333 non-null float64
17	LetterRatioInURL	22333 non-null float64
18	NoOfDegitsInURL	22333 non-null float64
19	DegitRatioInURL	22333 non-null float64
20	NoOfEqualsInURL	22333 non-null float64
21	NoOfQMarkInURL	22333 non-null float64
22	NoOfAmpersandInURL	22333 non-null float64
23	NoOfOtherSpecialCharsInURL	22333 non-null float64
24	SpacialCharRatioInURL	22333 non-null float64
25	IsHTTPS	22333 non-null float64
26	LineOfCode	22333 non-null float64
27	LargestLineLength	22333 non-null float64

28	HasTitle	22333 non-null float64
29	Title	22333 non-null object
30	DomainTitleMatchScore	22333 non-null float64
31	URLTitleMatchScore	22333 non-null float64
32	HasFavicon	22333 non-null float64
33	Robots	22333 non-null float64
34	IsResponsive	22333 non-null float64
35	NoOfURLRedirect	22333 non-null float64
36	NoOfSelfRedirect	22333 non-null float64
37	HasDescription	22333 non-null float64
38	NoOfPopup	22333 non-null float64
39	NoOfiFrame	22333 non-null float64
40	HasExternalFormSubmit	22333 non-null float64
41	HasSocialNet	22333 non-null float64
42	HasSubmitButton	22333 non-null float64
43	HasHiddenFields	22333 non-null float64
44	HasPasswordField	22333 non-null float64
45	Bank	22333 non-null float64
46	Pay	22333 non-null float64
47	Crypto	22333 non-null float64
48	HasCopyrightInfo	22333 non-null float64
49	NoOfImage	22333 non-null float64
50	NoOfCSS	22333 non-null float64
51	NoOfJS	22333 non-null float64
52	NoOfSelfRef	22333 non-null float64

53	NoOfEmptyRef	22333	non-null	float64
54	NoOfExternalRef	22333	non-null	float64
55	label	22333	non-null	float64

Pengelompokan:

Subset 1:

1. 'URLSimilarityIndex'
2. 'LineOfCode'
3. 'NoOfExternalRef'
4. 'NoOfImage'
5. 'NoOfSelfRef'

Subset 2:

1. 'URLSimilarityIndex'
2. 'LineOfCode'
3. 'NoOfExternalRef'
4. 'NoOfImage'
5. 'NoOfSelfRef'
6. 'NoOfJS'
7. 'LargestLineLength'
8. 'NoOfCSS'
9. 'HasSocialNet'
10. 'LetterRatioInURL'

Subset 3:

1. 'URLSimilarityIndex'
2. 'LineOfCode'
3. 'NoOfExternalRef'

4. 'NoOfImage'
5. 'NoOfSelfRef'
6. 'NoOfJS'
7. 'LargestLineLength'
8. 'NoOfCSS'
9. 'HasSocialNet'
10. 'LetterRatioInURL'
11. 'HasCopyrightInfo'
12. 'HasDescription'
13. 'IsHTTPS'
14. 'NoOfOtherSpecialCharsInURL'
15. 'DomainTitleMatchScore'

Subset 4:

1. 'URLSimilarityIndex'
2. 'LineOfCode'
3. 'NoOfExternalRef'
4. 'NoOfImage'
5. 'NoOfSelfRef'
6. 'NoOfJS'
7. 'LargestLineLength'
8. 'NoOfCSS'
9. 'HasSocialNet'
10. 'LetterRatioInURL'
11. 'HasCopyrightInfo'
12. 'HasDescription'

13. 'IsHTTPS'
14. 'NoOfOtherSpecialCharsInURL'
15. 'DomainTitleMatchScore'
16. 'HasSubmitButton'
17. 'SpacialCharRatioInURL'
18. 'TLDLegitimateProb'
19. 'URLTitleMatchScore'
20. 'IsResponsive'

Subset 5:

1. 'URLSimilarityIndex'
2. 'LineOfCode'
3. 'NoOfExternalRef'
4. 'NoOfImage'
5. 'NoOfSelfRef'
6. 'NoOfJS'
7. 'LargestLineLength'
8. 'NoOfCSS'
9. 'HasSocialNet'
10. 'LetterRatioInURL'
11. 'HasCopyrightInfo'
12. 'HasDescription'
13. 'IsHTTPS'
14. 'NoOfOtherSpecialCharsInURL'
15. 'DomainTitleMatchScore'
16. 'HasSubmitButton'

17. 'SpacialCharRatioInURL'
18. 'TLDDegitimateProb'
19. 'URLTitleMatchScore'
20. 'IsResponsive'
21. 'DegitRatioInURL'
22. 'NoOfDegitsInURL'
23. 'CharContinuationRate'
24. 'NoOfiFrame'
25. 'NoOfEmptyRef'

Sumber data (paper utama dari dataset ini adalah <https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset>).

Tahap 2 (poin: 10): Target Data (Optional)

- Poin ini digunakan ketika tidak semua atribut (pada data yang dipilih) digunakan.

Atribut yang tidak dipakai antara lain:

1. FILENAME
2. URL
3. Domain
4. TLD
5. Title

Tahap 3-4 (poin: 25): Data Pre-processing & Transformation

Link Code: https://colab.research.google.com/drive/1OY1hH-bUW0l3T1h16vGQ8TI_0Yx-v64d?usp=sharing

Beberapa teknik yang bisa digunakan yaitu (tentu sesuai kondisi dan kebutuhan):

1. Data Cleaning (*outliers, missing values*)

- missing values adalah proses pembersihan dan perbaikan data, dimana data sebelumnya tidak terstruktur atau tidak rapi, data bernilai NaN. Tujuannya adalah memastikan kualitas data dengan baik sehingga hasil analisis atau pemodelan yang dihasilkan akurat.

data sebelum dilakukannya missing values:

URLSimilarityIndex	CharContinuationRate	TLDLegitimateProb	URLCharProb	TLDLength	NoOfSubDomain	HasObfuscation	...	Pay	Crypto	HasCopyrightInfo	NoOfImage	NoOfCSS	NoOfJS	NoOfSelf
100.0	1.000000	0.522907	0.061933	3.0	1.0	0.0	...	0.0	0.0	1.0	34.0	20.0	28.0	11.0
100.0	0.666667	0.032650	0.050207	2.0	1.0	0.0	...	0.0	0.0	1.0	50.0	9.0	8.0	3.0
100.0	0.866667	0.028555	0.064129	2.0	2.0	0.0	...	0.0	0.0	1.0	10.0	2.0	7.0	4.0
100.0	1.000000	0.522907	0.057606	3.0	1.0	0.0	...	1.0	1.0	1.0	3.0	27.0	15.0	2.0
100.0	1.000000	0.079963	0.059441	3.0	1.0	0.0	...	1.0	0.0	1.0	244.0	15.0	34.0	7.0
...
100.0	1.000000	0.522907	0.059465	3.0	1.0	0.0	...	1.0	0.0	1.0	25.0	5.0	10.0	3.0
100.0	1.000000	0.014148	0.061395	2.0	1.0	0.0	...	0.0	0.0	0.0	167.0	6.0	13.0	5.0
100.0	1.000000	0.522907	0.059711	3.0	1.0	0.0	...	0.0	0.0	1.0	29.0	23.0	27.0	6.0
100.0	1.000000	0.079963	0.062096	3.0	1.0	0.0	...	1.0	0.0	0.0	19.0	2.0	12.0	8.0
100.0	1.000000	0.079963	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN

data sesudah dilakukannya missing values:

URLSimilarityIndex	CharContinuationRate	TLDLegitimateProb	...	Pay	Crypto	HasCopyrightInfo	NoOfImage	NoOfCSS	NoOfJS	NoOfSelf
100.0	1.000000	0.522907	...	0.0	0.0	1.0	34.0	20.0	28.0	11.0
100.0	0.666667	0.032650	...	0.0	0.0	1.0	50.0	9.0	8.0	3.0
100.0	0.866667	0.028555	...	0.0	0.0	1.0	10.0	2.0	7.0	4.0
100.0	1.000000	0.522907	...	1.0	1.0	1.0	3.0	27.0	15.0	2.0
100.0	1.000000	0.079963	...	1.0	0.0	1.0	244.0	15.0	34.0	7.0
...
100.0	1.000000	0.522907	...	1.0	0.0	1.0	25.0	5.0	10.0	3.0
100.0	1.000000	0.014148	...	0.0	0.0	0.0	167.0	6.0	13.0	5.0
100.0	1.000000	0.522907	...	0.0	0.0	1.0	29.0	23.0	27.0	6.0
100.0	1.000000	0.079963	...	1.0	0.0	0.0	19.0	2.0	12.0	8.0
100.0	1.000000	0.079963	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0

- outliers bertujuan untuk mengidentifikasi observasi yang tidak biasa dalam data. Outliers dapat mengganggu analisis statistik dan machine learning dengan menyebabkan bias dalam estimasi dan pengujian.

output dilakukannya outliers:

URLLength	119
DomainLength	417
IsDomainIP	33
URLSimilarityIndex	25
CharContinuationRate	175
TLDLegitimateProb	0
URLCharProb	0
TLDLength	0
NoOfSubDomain	0
HasObfuscation	0
NoOfObfuscatedChar	0
ObfuscationRatio	0
NoOfLettersInURL	0
LetterRatioInURL	0
NoOfDegitsInURL	0
DegitRatioInURL	0
NoOfEqualsInURL	0
NoOfQMarkInURL	0
NoOfAmpersandInURL	0
NoOfOtherSpecialCharsInURL	0
SpacialCharRatioInURL	0
IsHTTPS	0
LineOfCode	0
LargestLineLength	0
HasTitle	0

DomainTitleMatchScore	0
URLTitleMatchScore	0
HasFavicon	0
Robots	0
IsResponsive	0
NoOfURLRedirect	0
NoOfSelfRedirect	0
HasDescription	0
NoOfPopup	0
NoOfiFrame	0
HasExternalFormSubmit	0
HasSocialNet	0
HasSubmitButton	0
HasHiddenFields	0
HasPasswordField	0
Bank	0
Pay	0
Crypto	0
HasCopyrightInfo	0
NoOfImage	0

2. Data Integration (*data duplicate*)

Pada tahap integration simulasi yang cocok digunakan adalah data duplicate, yang bertujuan untuk mengatasi duplikasi. Tujuan utamanya: Meningkatkan Kualitas Data, Meningkatkan Kinerja Analisis, Mencegah Bias, Meningkatkan Interpretasi, dan Meningkatkan Efisiensi Penyimpanan. Dengan demikian, data yang digunakan dalam analisis dan pengambilan keputusan adalah data yang berkualitas, akurat, dan bebas dari redundansi.

output dilakukannya data duplicate:

```
▶ duplicate_rows = df[df.duplicated()]

# Menampilkan data duplikat
print("Data Duplikat:")
print(duplicate_rows)

Data Duplikat:
Empty DataFrame
Columns: [FILENAME, URL, URLLength, Domain, DomainLength, IsDomainIP, T
Index: []

[0 rows x 56 columns]
```

3. Data Transformation (*normalization minmax*)

data sebelum dilakukannya data transformasi minmax:

DomainLength	IsDomainIP	TLD	URLSimilarityIndex	CharContinuationRate	TLDLegitimateProb	...	Pay	Crypto	HasCopyrightInfo	NoOfImage	NoOfCSS
24	0	com	100.0	1.000000	0.522907	...	0.0	0.0	1.0	34.0	20.0
16	0	de	100.0	0.666667	0.032650	...	0.0	0.0	1.0	50.0	9.0
22	0	uk	100.0	0.866667	0.028555	...	0.0	0.0	1.0	10.0	2.0
19	0	com	100.0	1.000000	0.522907	...	1.0	1.0	1.0	3.0	27.0
26	0	org	100.0	1.000000	0.079963	...	1.0	0.0	1.0	244.0	15.0

data setelah dilakukannya data transformasi minmax:

DomainLength	IsDomainIP	TLD	URLSimilarityIndex	CharContinuationRate	TLDLegitimateProb	...	Pay	Crypto	HasCopyrightInfo	NoOfImage	NoOfCSS
0.198020	0.0	com	1.0	1.000000	1.000000	...	0.0	0.0	1.0	0.004450	0.046189
0.118812	0.0	de	1.0	0.666667	0.062440	...	0.0	0.0	1.0	0.006544	0.020785
0.178218	0.0	uk	1.0	0.866667	0.054608	...	0.0	0.0	1.0	0.001309	0.004619
0.148515	0.0	com	1.0	1.000000	1.000000	...	1.0	1.0	1.0	0.000393	0.062356
0.217822	0.0	org	1.0	1.000000	0.152920	...	1.0	0.0	1.0	0.031933	0.034642

4. Data Reduction (*penghapusan kolom*)

data sebelum dilakukannya data cleaning:

	FILENAME	URL	URLLength	Domain	DomainLength	IsDomainIP	TLD	URLSimilarityIndex
0	521848.txt	https://www.southbankmosaics.com	31	www.southbankmosaics.com	24	0	com	100.0
1	31372.txt	https://www.uni-mainz.de	23	www.uni-mainz.de	16	0	de	100.0
2	597387.txt	https://www.voicefmradio.co.uk	29	www.voicefmradio.co.uk	22	0	uk	100.0
3	554095.txt	https://www.sfnmjournai.com	26	www.sfnmjournai.com	19	0	com	100.0
4	151578.txt	https://www.rewildingargentina.org	33	www.rewildingargentina.org	26	0	org	100.0

5 rows x 56 columns

data setelah dilakukannya data cleaning:

Informasi setelah penghapusan kolom:							
	URLLength	DomainLength	IsDomainIP	URLSimilarityIndex	CharContinuationRate	TLDLegitimateProb	U
0	31	24	0	100.0	1.000000	0.522907	
1	23	16	0	100.0	0.666667	0.032650	
2	29	22	0	100.0	0.866667	0.028555	
3	26	19	0	100.0	1.000000	0.522907	
4	33	26	0	100.0	1.000000	0.079963	
...
22329	26	19	0	100.0	1.000000	0.522907	
22330	17	10	0	100.0	1.000000	0.014148	
22331	26	19	0	100.0	1.000000	0.522907	
22332	27	20	0	100.0	1.000000	0.079963	
22333	26	19	0	100.0	1.000000	0.079963	

22334 rows x 51 columns

Tahap 5 (poin: 25): Data Mining

- Algoritma data mining yang digunakan (sesuai data mining task).
- Skenario eksperimen sederhana.

Tahap 6 (poin: 20): Knowledge Interpretation

- Pola-pola *useful* yang telah ditemukan.

Tahap 7 (poin: 15): Reporting

- Simple academic Poster.
- Jupiter Notebook (Python)

