Mata Kuliah - Penggalian Data

Nama Kelompok :

Anggota :

[202110370311222– Ibnu Fauzan Rachmadhanu]

[202110370311234 – Muhammad Wahyudi]

[202110370311241 – Abd Baasithur Rizqu]

Berikut ini merupakan update template laporan Mini Project kuliah Penggalian 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 phising. Analisis URL dengan menggunakan teknik Mutual Information dan logistic regression dalam penelitian. Idemtifikasi fitur fitur paling informatif untuk membedakan upaya phising, memungkinkan memperkuat pertahanan dan tetap beada di depan takrik phising 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	URLSimiliratyIndex	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 diteteksi 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 trhadap total karakter dalam URL.	Float

19	DegitRatioInURL	Metrik yang mengukur rasio digit trhadap total karakter dalam URL.	Integer
20	NoOfDegitsInURL	Metrik yang mengukur jumlah digit atau angka trhadap 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	NoOfOtherSpecialCharsIURL	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 mengukut 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 mengukurt 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 mentrasfer uang atau nilai keuangan lainnya dari sartu pihak ke pihak lain sebagai ganti jasa, barang aau 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 dikelompokan sebelumnya. Pada tahapan clasifikasi, 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 NoOfOtherSpecialChars	InURL 22333 non-null float64
24 SpacialCharRatioInURL	22333 non-null float64

22333 non-null float64

22333 non-null float64

27 LargestLineLength 22333 non-null float64

25 IsHTTPS

26 LineOfCode

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

Pengelompokann:

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. 'TLDLegitimateProb'
- 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 isi 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

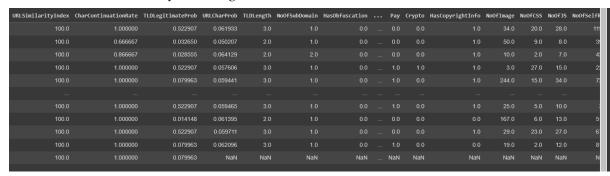
Tahap 3-4 (poin: 25): Data Pre-processing & Transformation

Link Code: https://colab.research.google.com/drive/10Y1hH-bUW0l3T1h16vGQ8TI_0Yx-v64d?usp=sharing

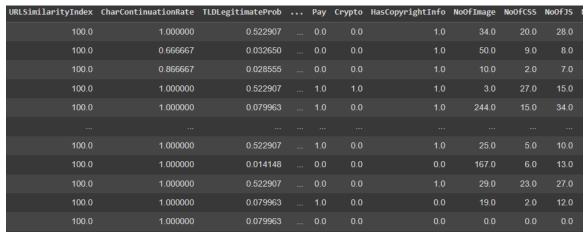
Beberapa teknik yang bisa digunakan yaitu (tentu sesuai kondisi dan kebutuhan):

- 1. Data Cleaning (outliers, missing values)
 - missing values adalah proses pembersihan dan perbaikan data, dimana data sebelumnya tidak terstruktur atau tidak rapi, data bervalue NaN. Tujuannya adalah memastikan kualitas data dengan baik sehingga hasil analisis atau pemodelan yang dihasilkan akurat.

data sebelum dilakukannya missing values:



data sesudah dilakukannya missing values:



 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	Ø	
URLCharProb	0	
TLDLength	0	
NoOfSubDomain	ø	
HasObfuscation	0	
NoOfObfuscatedChar	0	
ObfuscationRatio	0	
NoOfLettersInURL	0	
LetterRatioInURL	ø	
NoOfDegitsInURL	8	
DegitRatioInURL	0	
NoOfEqualsInURL	0	
NoOfQMarkInURL	0	
NoOfAmpersandInURL	0	
NoOfOtherSpecialCharsInURL	0	
SpacialCharRatioInURL	0	
ISHTTPS	0	
LineOfCode	0	
LargestLineLength	0	
HasTitle	9	

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 digubnakan 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)

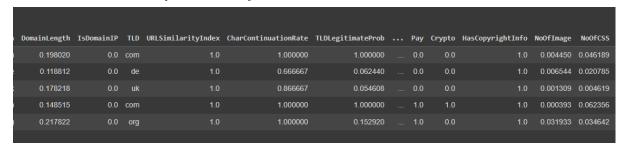
Data Duplikat:
Empty DataFrame
Columns: [FILENAME, URL, URLLength, Domain, DomainLength, IsDomainIP, Index: []
[0 rows x 56 columns]
```

3. Data Transformation (normalization minmax)

data sebelum dilakukannya data transformasii minmax:

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

data setelah dilakukannya data transformasi minmax:

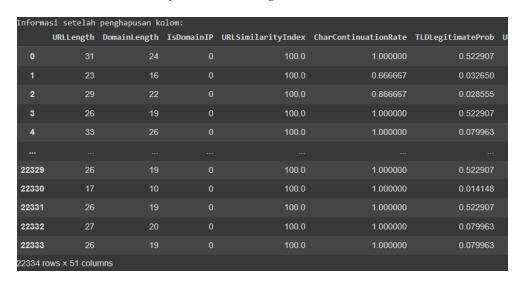


4. Data Reduction (penghapusan kolom)

data sebelum dilakukannya data cleaning:



data setelah dilakukannya data cleaning:



Tahap 5 (poin: 25): Data Mining

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

Tahap 6 (poin: 20): Knowledge Interpretation

· Pola-pola *useful* yang telah ditemukan.

Tahap 7 (poin: 15): Reporting

- · Simple academic Poster.
- · Jupiter Notebook (Python)