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 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.
    - 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.

- 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		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	Float

		level domain (TLD) yang digunakan dalam suatu website.	
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	Integer

		trhadap total karakter dalam URL.	
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 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		Float
32	URLTitleMatchScore		Float

33	HasFavicon	Sebagian besar situs web asli menyertakan logo situs web mereka dalam tag favicon.	Integer
34	Robots		Integer
35	IsResponsive	Situs web dirancang untuk menjadi responsif.	Integer
36	NoOfURLRedirect	Situs phishing dapat mengarahkan pengguna ke halaman yang berbeda.	Integer
37	NoOfSelfRedirect		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		Integer
44	HasPasswordField		Integer
45	HasHiddenFields		Integer
46	Bank		Integer

47	Pay		Integer
48	Crypto		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

    - Jelaskan data mining task yang akan digunakan (*classification, clustering, regression, association rule mining, anomaly detection,* dsb.).
        1. Classification. Pada dipilih sebagai tugas untuk data mining, untuk memprediksi apakah sebuah website termasuk dalam kategori phishing atau bukan.
- 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:

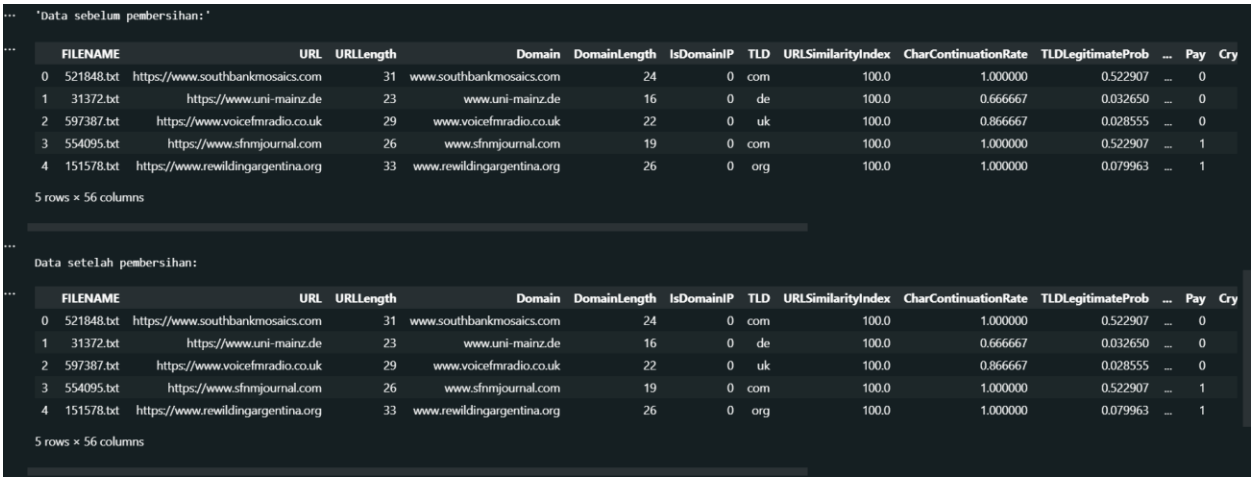
1. FILENAME
2. URL
3. URLLenght
4. Domain
5. DomainLenght
6. IsDomainIP
7. TLD
8. CharContinuationRate
9. TLDLegitimateProb
10. URLCharProb
11. TLDLength
12. NoOfSubDomain
13. HasObfuscation
14. NoOfObfuscatedChar
15. ObfuscationRatio
16. NoOfLettersInURL
17. LetterRatioInURL
18. NoOfDegitsInURL
19. DegitRatioInURL
20. NoOfEqualsInURL
21. NoOfQMarkInURL
22. NoOfAmpersandInURL
23. NoOfOtherSpecialCharsInURL
24. SpacialCharRatioInURL
25. IsHTTPS adalah HTTP aman.
26. LargestLineLength
27. HasTitle
28. Title
29. DomainTitleMatchScore
30. URLTitleMatchScore
31. HasFavicon
32. Robots
33. IsResponsive

34. NoOfURLRedirect
35. NoOfSelfRedirect
36. HasDescription
37. NoOfPopup
38. NoOfiFrame
39. HasExternalFormSubmit
40. HasSocialNet
41. HasSubmitButton
42. HasHiddenFields
43. HasPasswordField
44. Bank
45. Pay
46. Crypto
47. HasCopyrightInfo
48. NoOfCSS
49. NoOfJS
50. NoOfEmptyRef
51. Label

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

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

- Data Cleaning: 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 pembersihan:'

	FILENAME	URL	URLLength	Domain	DomainLength	IsDomainIP	TLD	URLSimilarityIndex	CharContinuationRate	TLDLegitimateProb	...	Pay	Cry
0	521848.txt	https://www.southbankmosaics.com	31	www.southbankmosaics.com	24	0	com	100.0	1.000000	0.522907	...	0	
1	31372.txt	https://www.uni-mainz.de	23	www.uni-mainz.de	16	0	de	100.0	0.666667	0.032650	...	0	
2	597387.txt	https://www.voicefmradio.co.uk	29	www.voicefmradio.co.uk	22	0	uk	100.0	0.866667	0.028555	...	0	
3	554095.txt	https://www.sfnmjournal.com	26	www.sfnmjournal.com	19	0	com	100.0	1.000000	0.522907	...	1	
4	151578.txt	https://www.rewildingargentina.org	33	www.rewildingargentina.org	26	0	org	100.0	1.000000	0.079963	...	1	

5 rows × 56 columns

Data setelah pembersihan:

	FILENAME	URL	URLLength	Domain	DomainLength	IsDomainIP	TLD	URLSimilarityIndex	CharContinuationRate	TLDLegitimateProb	...	Pay	Cry
0	521848.txt	https://www.southbankmosaics.com	31	www.southbankmosaics.com	24	0	com	100.0	1.000000	0.522907	...	0	
1	31372.txt	https://www.uni-mainz.de	23	www.uni-mainz.de	16	0	de	100.0	0.666667	0.032650	...	0	
2	597387.txt	https://www.voicefmradio.co.uk	29	www.voicefmradio.co.uk	22	0	uk	100.0	0.866667	0.028555	...	0	
3	554095.txt	https://www.sfnmjournal.com	26	www.sfnmjournal.com	19	0	com	100.0	1.000000	0.522907	...	1	
4	151578.txt	https://www.rewildingargentina.org	33	www.rewildingargentina.org	26	0	org	100.0	1.000000	0.079963	...	1	

5 rows × 56 columns

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)