תיאור כל הפיצ'רים בקובץ phishing\_url\_test.csv

|  |  |
| --- | --- |
| שם הפיצ'ר | הסבר בעברית |
| url | כתובת האתר המלאה. משמשת לזיהוי בלבד, לרוב לא תורמת ישירות למודל. |
| length\_url | אורך מלא של כתובת האתר. כתובות פישינג לרוב ארוכות כדי להסתיר מידע. |
| length\_hostname | אורך שם המארח (hostname) בתוך ה-URL. מארחים ארוכים עלולים להעיד על פישינג. |
| ip | האם נעשה שימוש בכתובת IP במקום דומיין (1 = כן). שימוש כזה נפוץ בפישינג. |
| nb\_dots | מספר הנקודות (.) ב-URL. ריבוי נקודות עשוי להצביע על סאב-דומיינים מזויפים. |
| nb\_hyphens | מספר המקפים (-) בכתובת. מקפים רבים יכולים להצביע על התחזות. |
| nb\_at | מספר סימני '@' בכתובת. משמשים להסוואה ובלבול המשתמש. |
| nb\_qm | מספר סימני שאלה (?) – שימוש בפרמטרים רבים עלול להעיד על ניסיון התחזות. |
| nb\_and | מספר הסימנים '&' – שימוש בפרמטרים מרובים. |
| nb\_or | מספר הסימנים '|' – נדיר בכתובות לגיטימיות. |
| nb\_eq | מספר הסימנים '=' – שימוש בפרמטרים ב-URL. |
| nb\_underscore | מספר הקווים התחתונים (\_) בכתובת. |
| nb\_tilde | מספר הסימנים '~' – שימוש לא שגרתי. |
| nb\_percent | מספר סימני '%' – קידוד כתובת או ניסיון הסוואה. |
| nb\_slash | מספר הלוכסנים (/) – מצביע על עומק בתיקיות. |
| nb\_star | מספר הכוכביות (\*) – שימוש נדיר שמעלה חשד. |
| nb\_colon | מספר הנקודתיים (:) – ריבוי לא תקני. |
| nb\_comma | מספר הפסיקים (,) – אינו רגיל בכתובות לגיטימיות. |
| nb\_semicolumn | מספר נקודה-פסיק (;) – שימוש חריג. |
| nb\_dollar | מספר סימני דולר ($) – לא מקובל בכתובת רגילה. |
| nb\_space | מספר הרווחים או קידוד רווח (%20). |
| nb\_www | מספר הופעות של 'www' – יותר מאחת עשויה להצביע על התחזות. |
| nb\_com | מספר הופעות של '.com' – ריבוי עלול להצביע על זיוף. |
| nb\_dslash | מספר הופעות של '//' – ריבוי מעבר לשני מקובל מעלה חשד. |
| http\_in\_path | האם נמצא המחרוזת 'http' בתוך הנתיב של ה-URL (לא בתחילתו). שימוש כזה עלול לבלבל את המשתמש לחשוב שמדובר בכתובת מאובטחת. |
| https\_token | האם 'https' מופיע בתוך שם הדומיין – מבלבל את המשתמש לחשוב שזה אתר מאובטח. |
| ratio\_digits\_url | יחס הספרות לאורך ה-URL – מספרים רבים יכולים להעיד על אוטומציה או התחזות. |
| ratio\_digits\_host | יחס הספרות בדומיין – מעלה חשד כאשר גבוה מדי. |
| punycode | האם נעשה שימוש בקידוד פוני (Punycode) לדומיין בינלאומי – משמש לפישינג. |
| port | האם פורט מצוין בכתובת – לרוב לא מצוין באתרים לגיטימיים. |
| tld\_in\_path | האם סיומת דומיין (למשל '.com') מופיעה בנתיב – עשוי להיות ניסיון להסוואה. |
| tld\_in\_subdomain | האם סיומת דומיין מופיעה בסאב-דומיין – רמז לזיוף. |
| abnormal\_subdomain | האם הסאב-דומיין חריג (למשל עם מספרים או דפוסים מוזרים). |
| nb\_subdomains | מספר הסאב-דומיינים – מספר גבוה מדי עשוי להיות התחזות. |
| prefix\_suffix | האם שם הדומיין מכיל מקף – עשוי להיות ניסיון התחזות למותג. |
| random\_domain | האם הדומיין מורכב מתווים אקראיים – אינדיקציה לפישינג. |
| shortening\_service | האם נעשה שימוש בשירות קיצור כתובות (bit.ly וכו'). |
| path\_extension | האם הנתיב כולל סיומות חשודות כמו '.exe'. |
| nb\_redirection | מספר ההפניות – ריבוי הפניות הוא סימן חשוד. |
| nb\_external\_redirection | מספר הפניות לדומיינים חיצוניים – עשוי להיות מסוכן. |
| length\_words\_raw | מספר המילים (תווים מופרדים במפרידים כמו /, ., -) בכתובת האתר. |
| char\_repeat | מספר תווים החוזרים על עצמם ברצף (כמו aaa, 1111). תבניות חוזרות יכולות להעיד על כתובת אוטומטית או מזויפת. |
| shortest\_words\_raw | אורך המילה הקצרה ביותר בכתובת. |
| shortest\_word\_host | אורך המילה הקצרה ביותר בשם הדומיין. |
| shortest\_word\_path | אורך המילה הקצרה ביותר בנתיב (path) של הכתובת. |
| longest\_words\_raw | אורך המילה הארוכה ביותר בכתובת כולה. |
| longest\_word\_host | אורך המילה הארוכה ביותר בדומיין. |
| longest\_word\_path | אורך המילה הארוכה ביותר בנתיב. |
| avg\_words\_raw | אורך ממוצע של המילים בכתובת האתר. |
| avg\_word\_host | אורך ממוצע של מילים בשם המארח. |
| avg\_word\_path | אורך ממוצע של מילים בנתיב. |
| phish\_hints | מספר רמזים לחשד לפישינג שנמצאו בתוכן העמוד (כמו טקסטים מחשידים). |
| domain\_in\_brand | האם הדומיין תואם למותג אמיתי. |
| brand\_in\_subdomain | האם שם מותג מופיע בסאב-דומיין – עלול להיות התחזות. |
| brand\_in\_path | האם שם מותג מופיע בנתיב – לעיתים מזויף. |
| suspecious\_tld | האם לסיומת הדומיין (TLD) יש מוניטין שליליים (כמו .tk, .xyz). |
| statistical\_report | האם קיימים דיווחים סטטיסטיים שדירגו את האתר כחשוד. |
| nb\_hyperlinks | מספר ההיפר-קישורים (a href) בעמוד. |
| ratio\_intHyperlinks | יחס הקישורים שמפנים פנימית בתוך האתר – ערך נמוך עלול להעיד על אתר מזויף. |
| ratio\_extHyperlinks | יחס הקישורים החיצוניים באתר. ערך גבוה עשוי להעיד על התחזות או גנבת זהות. |
| ratio\_nullHyperlinks | אחוז הקישורים שמובילים לכתובות ריקות (כמו href="#") – נפוץ בהתחזות. |
| nb\_extCSS | מספר קבצי CSS חיצוניים שנטענים לדף. ריבוי קבצים חיצוניים עלול להעיד על דף לא אמין. |
| ratio\_intRedirection | יחס ההפניות שמובילות פנימית – עשוי להעיד על מבנה תקין של האתר. |
| ratio\_extRedirection | יחס ההפניות (Redirects) שמפנות לדומיינים חיצוניים – אינדיקציה לסיכון. |
| ratio\_intErrors | אחוז שגיאות בקישורים פנימיים. |
| ratio\_extErrors | אחוז השגיאות בקישורים חיצוניים (404 וכד'). |
| login\_form | האם קיים טופס התחברות בעמוד. אתרי פישינג מחקים לרוב טפסים כאלה. |
| external\_favicon | האם ה-Favicon (האייקון שמופיע בטאב) נטען משרת חיצוני. אם כן – מעלה חשד. |
| links\_in\_tags | האם הקישורים מוטמעים בתגיות מטא, סקריפט או קישורים מוסתרים – סימן מחשיד. |
| submit\_email | האם טופס שליחת הנתונים משתמש בכתובת אימייל במקום שליחה לשרת – נפוץ בפישינג. |
| ratio\_intMedia | יחס המדיה שמוטענת מתוך הדומיין עצמו. ערך נמוך עלול להעיד שהעמוד תלוי במשאבים חיצוניים. |
| ratio\_extMedia | יחס המדיה (תמונות, וידאו וכו') שמוטענת ממקורות חיצוניים. ערך גבוה מעיד על תלות חיצונית – סימן לפישינג. |
| sfh | האם שדה טופס HTML שולח נתונים לכתובת חשודה או ריקה. (SFH = Server Form Handler). ערך 1 לרוב מעיד על פישינג. |
| iframe | האם העמוד כולל שימוש ב־iframe (מסגרות מוטמעות). שימוש כזה נפוץ בפישינג להסוואת תוכן. |
| popup\_window | האם הדף פותח חלון קופץ (pop-up). משמש לעיתים להונאה. |
| safe\_anchor | אחוז הקישורים בעמוד ש'נראים בטוחים' (למשל מכילים טקסט ברור). ערך נמוך עשוי להעיד על ניסיון התחזות. |
| onmouseover | האם נעשה שימוש באירוע JavaScript `onmouseover` (שינוי התנהגות בעכבר). נפוץ להטעיה. |
| right\_clic | האם יש חסימה של קליק ימני (מונע מהמשתמש לבדוק מקור הדף וכו'). |
| empty\_title | האם כותרת הדף ריקה – סימן לאתר לא מקצועי או חשוד. |
| domain\_in\_title | האם הדומיין מופיע בכותרת הדף – עשוי להעיד על שקיפות. |
| domain\_with\_copyright | האם הדומיין מופיע יחד עם סמל זכויות יוצרים (©) – ניסיון לשדר אמינות. |
| whois\_registered\_domain | האם הדומיין רשום לפי WHOIS. |
| domain\_registration\_length | משך זמן רישום הדומיין – קצר מאוד מעלה חשד. |
| domain\_age | גיל הדומיין – אתרי פישינג לרוב חדשים. |
| web\_traffic | מדד לתנועה באינטרנט – תנועה נמוכה יכולה להעיד על אתר פישינג. |
| dns\_record | האם קיימת רשומת DNS – אם לא, סימן לפישינג. |
| google\_index | האם האתר מופיע באינדקס של גוגל – אתרי פישינג לרוב לא מופיעים. |
| page\_rank | דירוג האתר לפי סמכות וקישורים חיצוניים – ערך נמוך מעלה חשד. |
| status | התווית: האם מדובר בדוגמה של פישינג (phishing) או אתר אמיתי (legitimate). זהו הפיצ'ר שמהווה את ה-label במודל. |