פיצ'רים מיותרים או לא רלוונטיים בקובץ האימון

להלן סיכום הפיצ'רים שמומלץ להסיר מקובץ האימון (phishing\_url\_train.csv), בהתבסס על ניתוח שונות וקורלציה גבוהה בין פיצ'רים:

# 🔻 פיצ'רים עם שונות אפס

|  |  |
| --- | --- |
| שם הפיצ'ר | הסבר |
| nb\_or | מספר סימני '|' – מופיע תמיד כ־0, ולכן חסר תועלת. |
| ratio\_nullHyperlinks | אחוז קישורים ריקים (href="#") – קבוע בכל הדוגמאות. |
| ratio\_intRedirection | יחס הפניות פנימיות – לא משתנה כלל. |
| ratio\_intErrors | שגיאות בקישורים פנימיים – תמיד 0. |
| submit\_email | האם הטופס שולח לכתובת אימייל – לא קיים בדאטה. |
| sfh | בדיקת Server Form Handler – ערך קבוע, לא שימושי. |

# 🔁 פיצ'רים מאוד דומים (קורלציה גבוהה)

|  |  |  |
| --- | --- | --- |
| פיצ'ר 1 | פיצ'ר 2 | הסבר |
| longest\_words\_raw | longest\_word\_path | שני הפיצ'רים כמעט תמיד זהים. מומלץ להשאיר רק אחד מהם (למשל את longest\_words\_raw). |

# ✅ סיכום: פיצ'רים שמומלץ להסיר

- nb\_or  
- ratio\_nullHyperlinks  
- ratio\_intRedirection  
- ratio\_intErrors  
- submit\_email  
- sfh  
- longest\_word\_path