מגיש: יוסי גוטליב 021767959

דוח מחקר – זיהוי כתובות פישינג

מחיפוש באינטרנט מצאתי ששיטות הנפוצות לזיהוי כתובות פישינג משתמשות בבניית מילון של כל המילים הנמצאות בכתובות ,בניית מטריצה של שכיחות המילים(-הפיצ'רים) בכל הכתובות ושימוש במודל ML כלשהו לניבוי.

יש שהשתמשו כפיצ'רים, לא במילים שבכתובות אלא: length of URL , length of Hostname , count of Letters, count of '%', use of IP or not, וכד' אבל לא ראיתי שהגיעו לתוצאות טובות יותר.

הDataset – שהשתמשתי בו – URLDATA מ-<https://www.kaggle.com/siddharthkumar25/malicious-and-benign-urls>

השתמשתי להשוואה גם בdataset\_small.csv- מ-

<https://data.mendeley.com/datasets/72ptz43s9v/1> וקיבלתי תוצאות דומות מאוד.

בניית מטריצת שכיחות נעשתה ע"י פונ' CountVectorizer ב- Sklearn.

לכאורה פונ' TfidVectorizer צריכה לתת תוצאה טובה יותר כי המישקול בה מתוחכם יותר ומתחשב בשכיחות המילה בכל הכתובות אך בהשוואה שעשיתי קיבלתי תוצאות טיפה יותר טובות דווקא ב- CountVectorizer אולי בגלל ששכיחות המילה בכל הכתובות כוללת גם את הכתובות הטובות וגם את הרעות.

בדקתי מודלים של LogisticRegrassion , DecisionTree,AdaBoost וקיבלתי תוצאות מדויקות מאוד בשלושתם:

ב- DecisionTree 0.998

ב- AdaBoost 0.997

ב- LogisticRegrassion 0.995

ויתרתי על בדיקת מודלים אחרים של ML ומשחק בפרמטרים ב DecisionTreeכיוון שהדיוק שקיבלתי כבר גבוה מאוד ושמרתי את מודל DecisionTree .