שמות המגישים:
מתן אביב

רוי ווסקר

השאלה עליה נרצה לענות:
כיום על השאלה האם כתובת URL היא תקינה או זדונית ניתן לענות בקלות על די הקלדה של הכתובת ב virustotal . אשר מחזיק דאטה סט של אתרים לגיטימיים וסך הכל בודק האם הכתובת שהוזנה נמצאת בדאטה סט או לא. אנחנו רוצים לענות על השאלה הזאת מזווית של למידת מכונה. היתרון של התעסקות בשאלה שעבורה הפתרון שרוב החברות מציעות הוא החזקה של מאגרי מידע גדולים הוא שיש לנו דאטה סט מאוד גדול מה שכמובן יתרום לאחוז דיוק גבוה ואפשרות להגיע לתוצאה מצוינת. הדאטה סט המצורף שמורכב מ2 מקורות מכיל 2 מיליון כתובת URL אשר חצי מתוכם לגיטימיים וחצי זדוניים. הלגיטימיים לקוחים מ Alexa's top 1 million most visited websites והזדוניים לקוחים מ Phishtank .

הכלים בהם נשתמש:

Logistic regression

Naïve base

Random forest

KNN