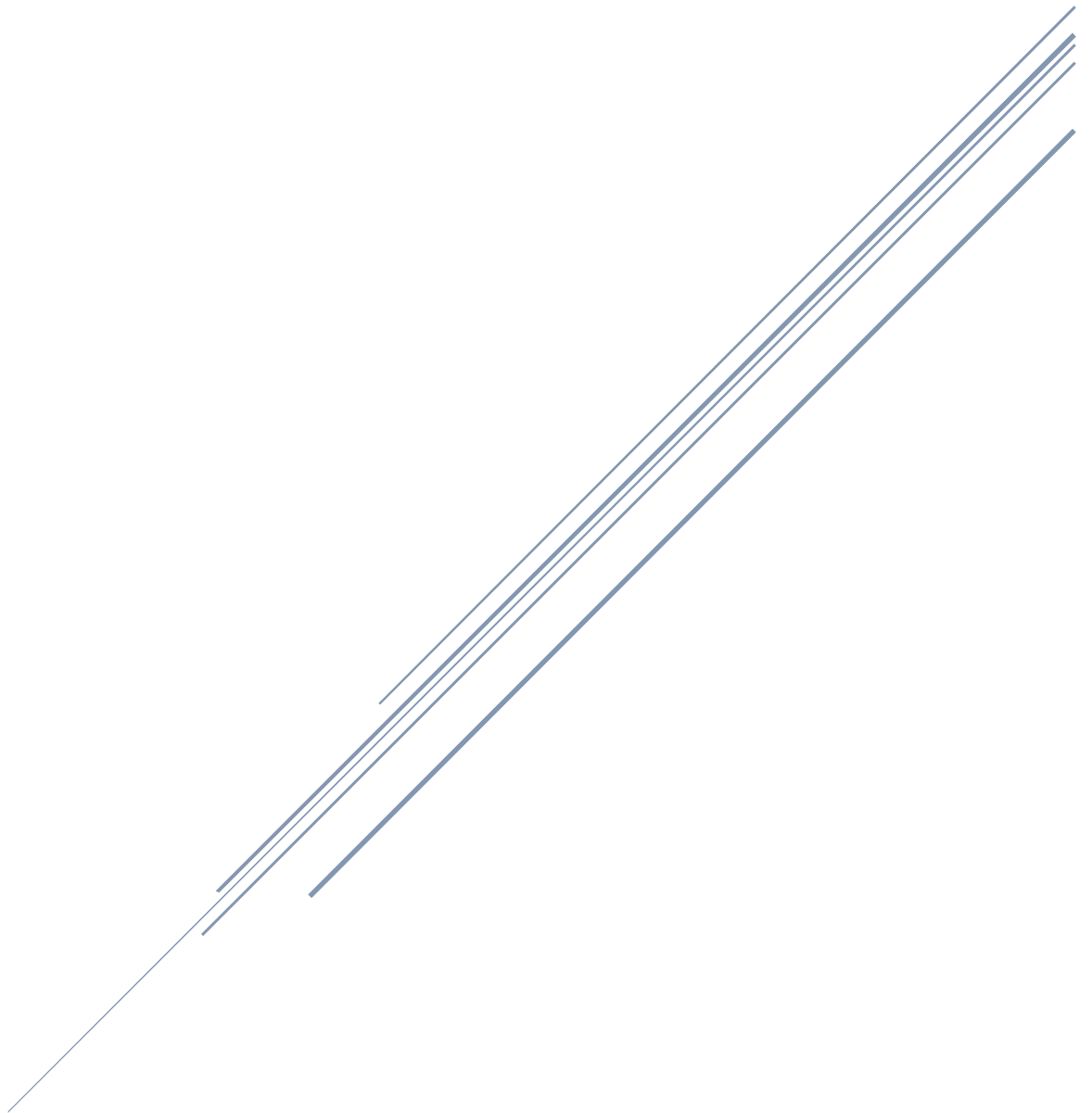


CLASSIFICATION PROJECT PROPOSAL

SDAIA T5 Data science boot camp

Norah Alqahtani, Albatoul Alosaimi, and Shroaq Almutairi



Introduction:

Malicious URL, a.k.a. malicious website, is a common and serious threat to cybersecurity. Malicious URLs host unsolicited content (spam, phishing, drive-by downloads, etc.) and lure unsuspecting users to become victims of scams (monetary loss, theft of private information, and malware installation), and cause losses of billions of dollars every year. It is imperative to detect and act on such threats in a timely manner. Traditionally, this detection is done mostly through the usage of blacklists. However, blacklists cannot be exhaustive, and lack the ability to detect newly generated malicious URLs. To improve the generality of malicious URL detectors, machine learning techniques have been explored with increasing attention in recent years. So, we aim in this Project to take advantage of Classification techniques to predict whether a specific URL malicious or benign.

Question/need:

- predicting whether, a specific URL malicious or benign.

Data Description:

We are planning to use an online available dataset

	whois_regDate	whois_expDate	whois_updatedDate	dot_count	url_len	digit_count	special_count	hyphen_count	double_slash	single_slash	at_the_rate	p
0	403	326	23	6	225	58	12	4	0	10	0	0
1	2727	194	168	7	177	47	0	1	0	11	0	0
2	5431	46	317	6	60	0	0	0	0	2	0	0
3	3643	374	5	1	116	21	1	1	1	10	0	0
4	-1	-1	-1	3	36	0	0	0	0	1	0	0
...
9995	8450	316	473	2	23	0	0	0	0	2	0	0
9996	8719	776	1109	3	28	0	0	0	0	1	0	0
9997	7481	553	112	3	43	1	0	0	0	3	0	0
9998	7658	376	3	3	38	0	0	1	0	1	0	0
9999	8443	689	824	3	32	0	0	0	0	1	0	0

10000 rows x 14 columns

Consist of URLs and its features: whois_regDate, whois_expDate, whois_updatedDate, dot_count, url_len, digit_count, special_count, hyphen_count, double_slash, single_slash, at_the_rate, protocol, and protocol_count. In addition, it has a label associated with each row. The target contain two labels (binary classification) where 0 means benign and 1 means malicious.

Algorithms

- We build Decision Tree, KNN, Random Forest, and Logistic Regression in python, and techniques such as regularization and polynomial features, adding interaction terms and dummy variables have been used.
- Rigorous model selection and evaluation has been used to select model between Decision Tree, KNN, Random Forest, and Logistic regression Regression

Tools

- Pandas and Numpy (Exploring the data)
- Matplotlib and Seaborn (Visualizing the data)
- Sci-kit Learn (linear Regression model and other models)