Name: Sultan Abdullah Bin Mahfouz Tuwaiq\_Four\_Week\_Data\_Science\_Bootcamp

Phishing Dataset

DATA DISCRIPTION

The dataset is providing information about phishing emails that happened in 2017. This data provides different features such as number of dots, URL length, number of percentages, frequent domain name mismatch, right click availability and other more features. The data set is considered to be a practical example to be taken into consideration to implement the different methods and machine learning models that have been learned during the bootcamp provided by Sdaia and Metis. The dimension of the dataset is approximately 10,000x50 which indicates that the problem would be beneficial to solve as well as interesting to deal with, and it has been taken from Kaggle website. It is estimated to be working with features such as the number of dashes in hostname, missing title and submit info to email features in order to predict if the email is considered to be a spam or not. In terms of modeling, it is intended to implement classification from supervised learning to estimate either the email is a spam or not.

PLAN & TOOLS

First of all, the objective of this proposal is to introduce the instructor with the intended project that will be implemented and submitted using GitHub. The main goal of the project is to utilize the enriching knowledge that have been taught during the bootcamp which includes learning fundamental concepts and the use of essential Python library such as NumPy, sklearn and panda. Therefore, the plan is to analyze the obtained dataset from Kaggle to understand the data much more clearly and implement what we have learned such as finding the mean, median and standard deviation of some features in the dataset. Additionally, it is planned to describe the general data and use informative built-in functions in NumPy library. Moreover, it is not limited to obtaining useful information but also visualize the dataset by using sklearn and matplotlib useful libraries which indicate a better understanding of the dataset. Finally, this approach might be extended to implement a famous algorithm, classification, in order to predict an output such as spam email or safe email. This project will illustrate and obtain the main objective which is analyzing the dataset and predicting one of the main features and explain step by step the progress of analyzing the data including data cleaning. Thus, this model can be beneficial for all email users for the sake of privacy and protection.

CONCLUSION

To sum up, this proposal illustrate to analyze the dataset obtained from Kaggle in order to utilize the enriching information learned during the bootcamp such as using tools in Python and also express the knowledge of using machine learning to predict an email is a spam or safe by training the model with existing data of shape 10,000x50 roughly.

Appendix

https://www.kaggle.com/shashwatwork/phishing-dataset-for-machine-learning