

Phishing Domain Detection

Project By – Tejas Muley

Introduction

- Phishing is a type of fraud in which an attacker impersonates a reputable company or person in order to get sensitive information such as login credentials or account information via email or other communication channels.
- Phishing is popular among attackers because it is easier to persuade someone to click a malicious link that appears to be authentic than it is to break through a computer's protection measures.

Objective

- The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing.
- Try out different machine learning algorithms that's best fit for the this project case.
- The main goal is to predict whether the domains are real or malicious.
- Deployment of project on cloud.

Dataset Description

- Dataset Link :-
<https://data.mendeley.com/datasets/72ptz43s9v/1>
- Dataset Description :-
<https://www.sciencedirect.com/science/article/pii/S2352340920313202>
- 57000 + records and 112 attributes.

Process

- **Data Export From CSV :**

Loading CSV data using python pandas and extracting all the data into dataframe in python file.

- **Data Pre-processing :**

Performing EDA to get insight of data like identifying distribution , outliers Treatment, trend among data, Check for null values in the columns. If present impute the null values.

- **Feature Selection :**

Chose only those features which are helpful for output.

- **Train and Test Split :**

Train data is 70% of whole dataset.

Test data is 30% of remaining dataset.

Data is randomly choose in train and test.

There is only train and test data available there is no validation data

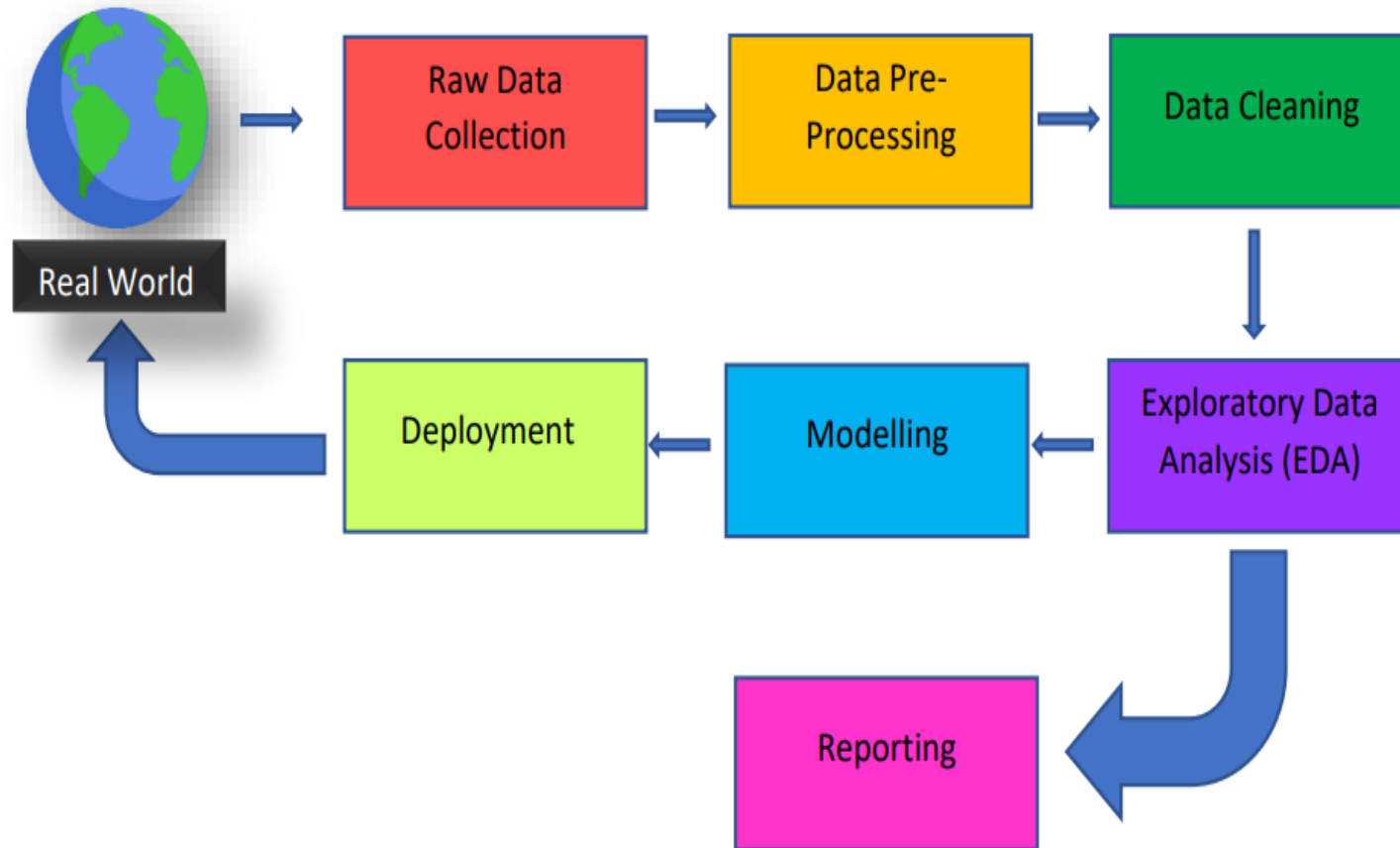
- **Model building :**

10 + Classification Machine Learning Algorithms applied for prediction and chose the best on amongst them.

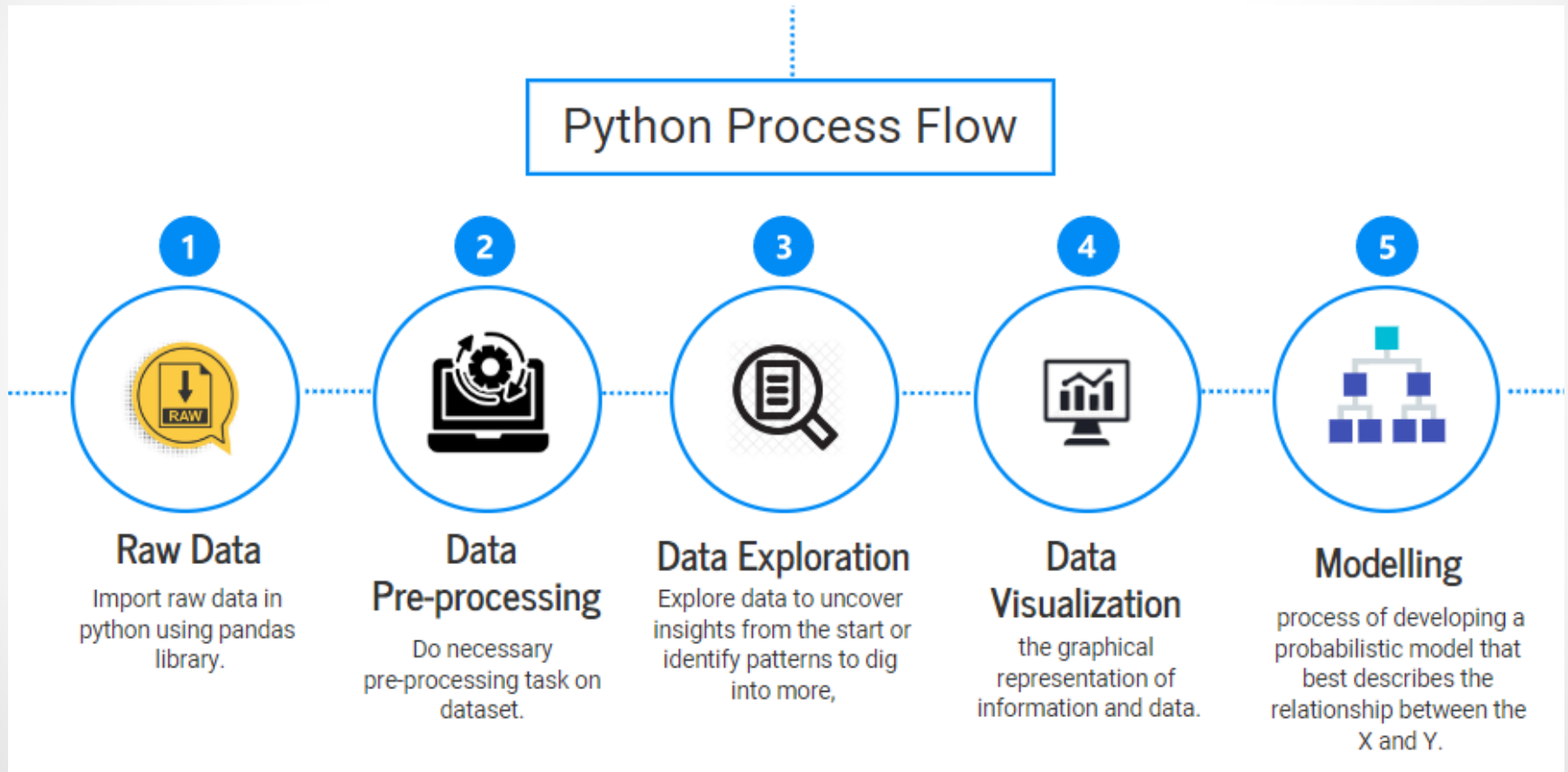
- **Pickle Model :**

Dumped the best model for future use and deployment.

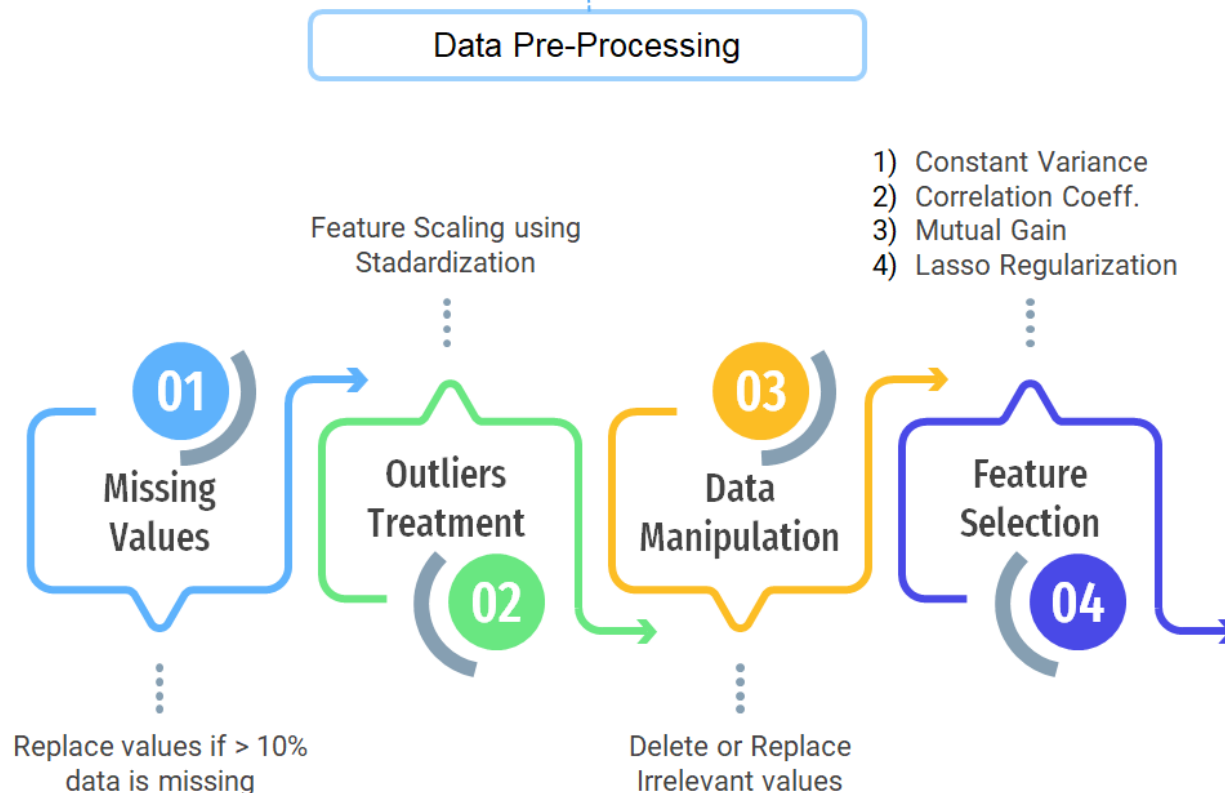
Architecture



Python Process Flow



Data Pre-processing

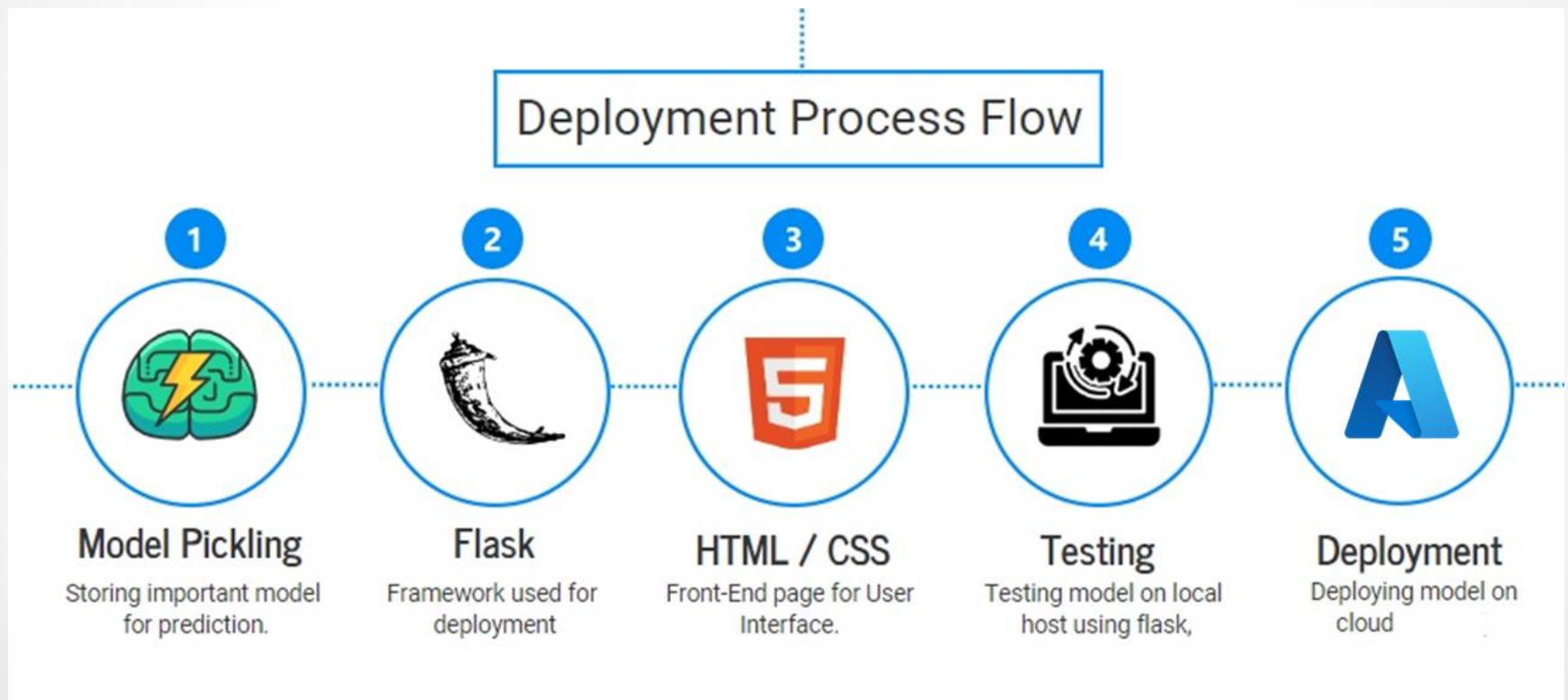


Model Building

Model Building



Deployment Process Flow



Tools Used in Project

Tools Used



Python

Python is used as a programming Language



Flask

Flask is used as a web framework to build a web application.



Database

Cassandra is used to store user inputs and prediction.



HTML

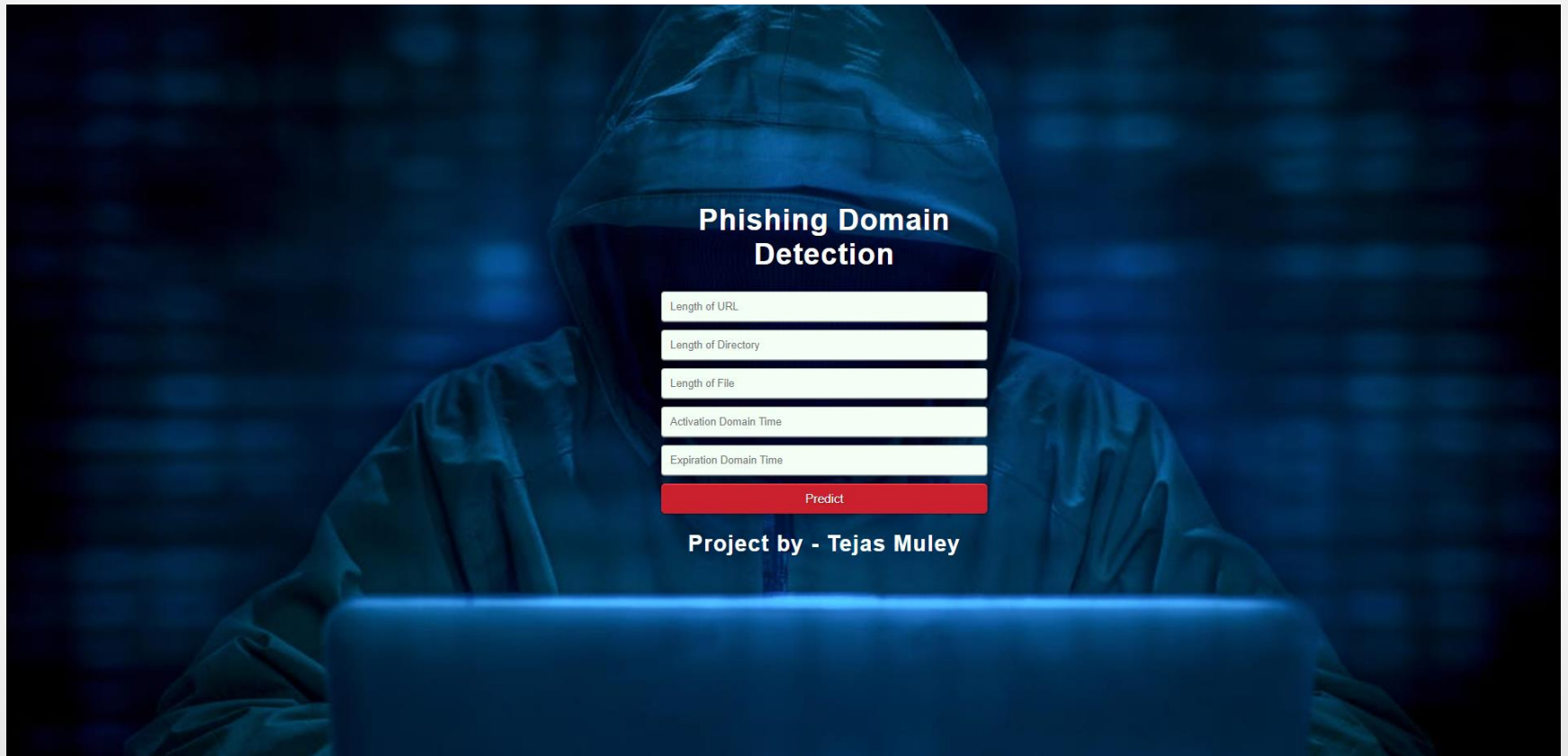
HTML is used for front-end we application.



Azure

Azure is used for deploying model on cloud.

HTML Home Page

The background image shows a person wearing a dark hoodie, sitting at a desk and using a laptop. The scene is dimly lit with a blue tint. Overlaid on the person's face is a web form titled "Phishing Domain Detection". The form contains five input fields for "Length of URL", "Length of Directory", "Length of File", "Activation Domain Time", and "Expiration Domain Time". Below these fields is a red button labeled "Predict". At the bottom of the form, it says "Project by - Tejas Muley".

Phishing Domain Detection

Length of URL

Length of Directory

Length of File

Activation Domain Time

Expiration Domain Time

Predict

Project by - Tejas Muley

HTML Output Page

This Site is Malicious !!!
Be Aware !!!



Conclusion

- This Phishing Domain Detection is used in-order to avoid phishing. According to the different parameters by the URL, we as an users should have an idea that how our personal information can leaked through phishing. This system helps to check whether the site is real or malicious, so that we can avoid and get alert either to provide information on that site or not.