iNeuron

Phishing Domain Detection (Classification [Machine Learning]).

Architecture Design:-

Project Member: Aman Gupta.

Introduction: - The software needs the architecture design to represent the design of software.

IEEE defines architectural design as “the process of defining a collection of hardware and software components and their interfaces to establish the framework for the development of a computer system.”

Each style will describe a system category that consists of :-

A set of components (e.g. a database, computational modules) that will perform a function required by the system.

The set of connectors will help in coordination, communication, and cooperation between the components.

Conditions that how components can be integrated to form the system.

Semantic models that help the designer to understand the overall properties of the system.

- **Scope:-**
  Architecture design (AD) is a component-level design process That follows a step-by-step refinements process. This process Can be used for designing data structures, required software, Architecture, source code, and ultimately, performance algorithms. Overall, the data organization may be defined during requirements analysis and then refined during data design work and the complete work flow.

- Constrains
  We predict the domain is a fake or not by taking input from the user.

Problem Statement:

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.

The mail goal is to predict whether the domains are real or malicious.

Approach:

The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing. Has been done on the project. Tried with different machine learning algorithms such as Logistic Regression, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, Adaptive Boosting, Gradient Boosted Tree.

Data-Set:

This data set consist of 88,647 websites labelled as legitimate or phishing and allow the researchers to train their classification models, build phishing detection systems.

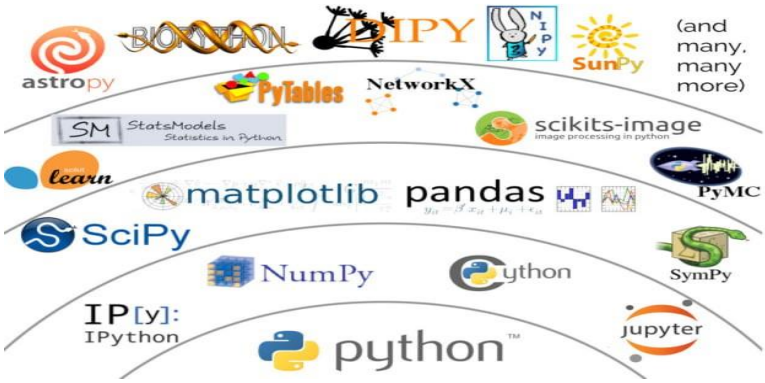
For more details of the dataset visit:

Datasets for phishing websites detection - ScienceDirect

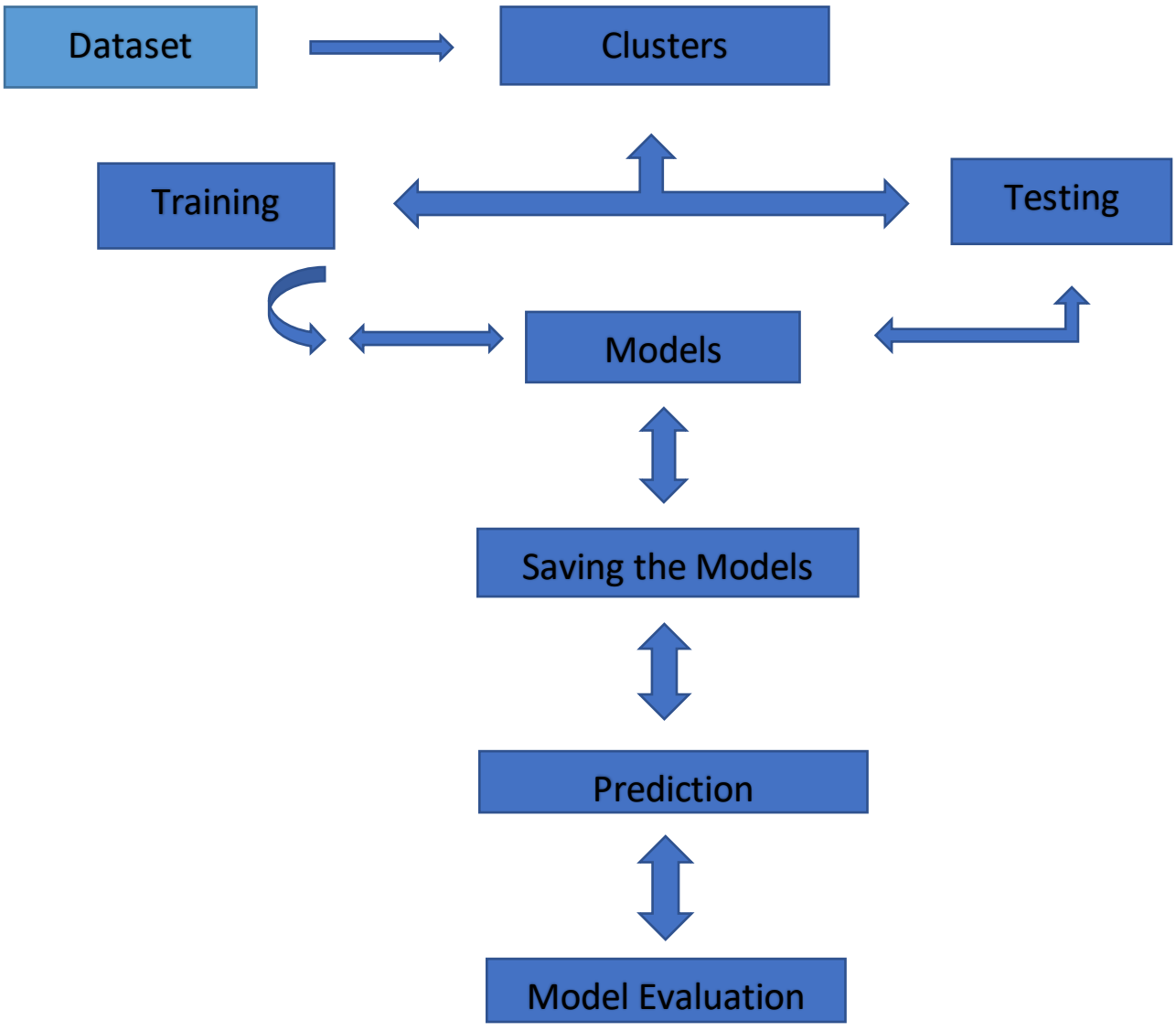
Datasets for phishing websites detection - ScienceDirect

Tools Used:

Python Programming language with some packages like NumPy, Pandas, Scikit learn, Pickle, Flask, HTML, CSS, JS



Architecture :-



- **User input / Output flow:-**

Conclusion :

It turns out model is performing well with a recall score of 78% in the cluster one and in the cluster two it is giving recall of 93% but there are some False Positives which will affect.