**RSIP Career Plus ML 044**

**PROJECT TITLE**

**Detecting Phishing Website Using Machine Learning**

**Category: Machine Learning**

**Skills Required:**  
Python ,Python Web Frame Works, Python For Data Analysis, Python For Data Visualization,DataPreprocessingTechniques,MachineLearning,Classification Algorithms

**TEAM MEMBERS:**

  RAGUL .R

KAVIN KUMAR.P

YUVAANESH.VB

LOGESH.VS

**Project Description:**

In order to detect and predict phishing website, we proposed an intelligent, flexible and effective system that is based on using classification Data mining algorithm. System uses machine learning technique to add new keywords into database.



**INTRODUTION:**

Phishing is a malicious attack in online theft to steal the user's private information. That is a kind of scam in which unauthorized user tries to gain user private data and thus user falls into such traps. The motive of our project is to propose a structure that is safe for identifying phishing websites in less time with high accuracy. Currently, people accomplish most online business, transferring money, bill payments i.e. all the things are carried out using websites or applications. Therefore, finding website phishing is an enormously important thing in our day to day life. Identifying phishing websites is a tough task. After a detailed survey on this problem, we found the list-based anti- phishing approaches(blacklist or whitelist) which store URLs in the database. This approach compares the URL entered by users in browsers with URLs that are put in the database. Using these approaches the newly build phishing URLs fail to detect which are not being included in the database. A phishing attack occurs when an unauthorized person tries to send an email or the URL to get sensitive information of users for misuse. The victim registers the details unknowingly which mainly consist of password or username, credit card numbers they are likely to accept.

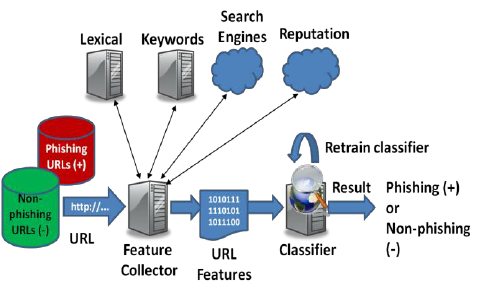
**What is Phishing Website?**

Phishing is the fraudulent attempt to obtain [sensitive information](https://en.wikipedia.org/wiki/Sensitive_information) or data, such as usernames, passwords and [credit card](https://en.wikipedia.org/wiki/Credit_card) details, by disguising oneself as a trustworthy entity in an [electronic communication](https://en.wikipedia.org/wiki/Electronic_communication). Typically carried out by [email spoofing](https://en.wikipedia.org/wiki/Email_spoofing), [instant messaging](https://en.wikipedia.org/wiki/Instant_messaging), and text messaging, phishing often directs users to enter [personal information](https://en.wikipedia.org/wiki/Personal_information) at a fake website which matches the [look and feel](https://en.wikipedia.org/wiki/Look_and_feel) of the legitimate site.

**The Facts About Phishing Website:**

* **Phishing** Scams Rely on Brand Name Recognition.
* Social Networking **Sites** are at Risk.
* Beware of Misspelled Domain Names.
* Cybercriminals Can Hide **Website** Addresses.
* Criminals Use Legitimate **URL**.
* You Can't Always Spot a Scam At First Sight.
* **Phishing Sites** Can Hide from Search Engines.

**Flow chart of Phishing Website**

****

**Experiment :**

**Proposed Solution:**

The main objective of this research is to use classification algorithms to identify the liver patients from healthy individuals. This project also aims to compare the classification algorithms based on their performance factors.

To detect the given website is phishing website or not.

The following algorithms are trained for this purpose, to choose the best performing classifier.

* Logistic Regression
* K-Nearest Neighbors

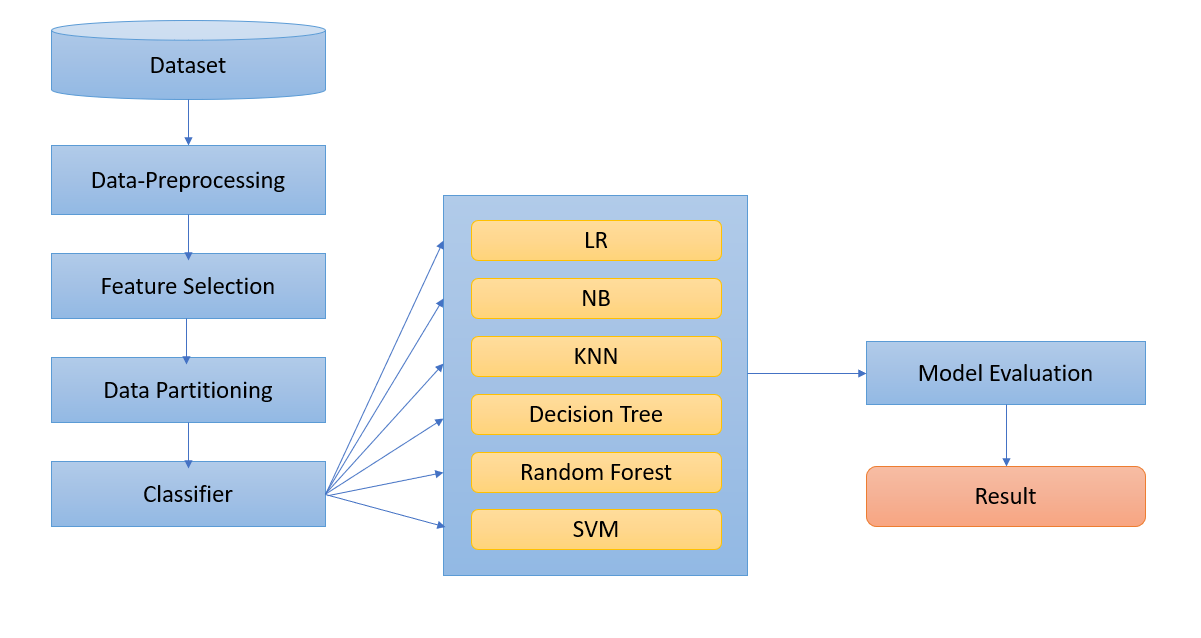
The results of each of the classification algorithm is summarized in the table shown below.

|  |  |
| --- | --- |
| **Model** | **Accuracy Score** |
| Logistic Regression | 0.912 |
| KNN | 0.935 |

As clearly summarized in the table, Logistic Regression gave the best result.

**THEORITICAL ANALYSIS:**

**BLOCK DIAGRAM**



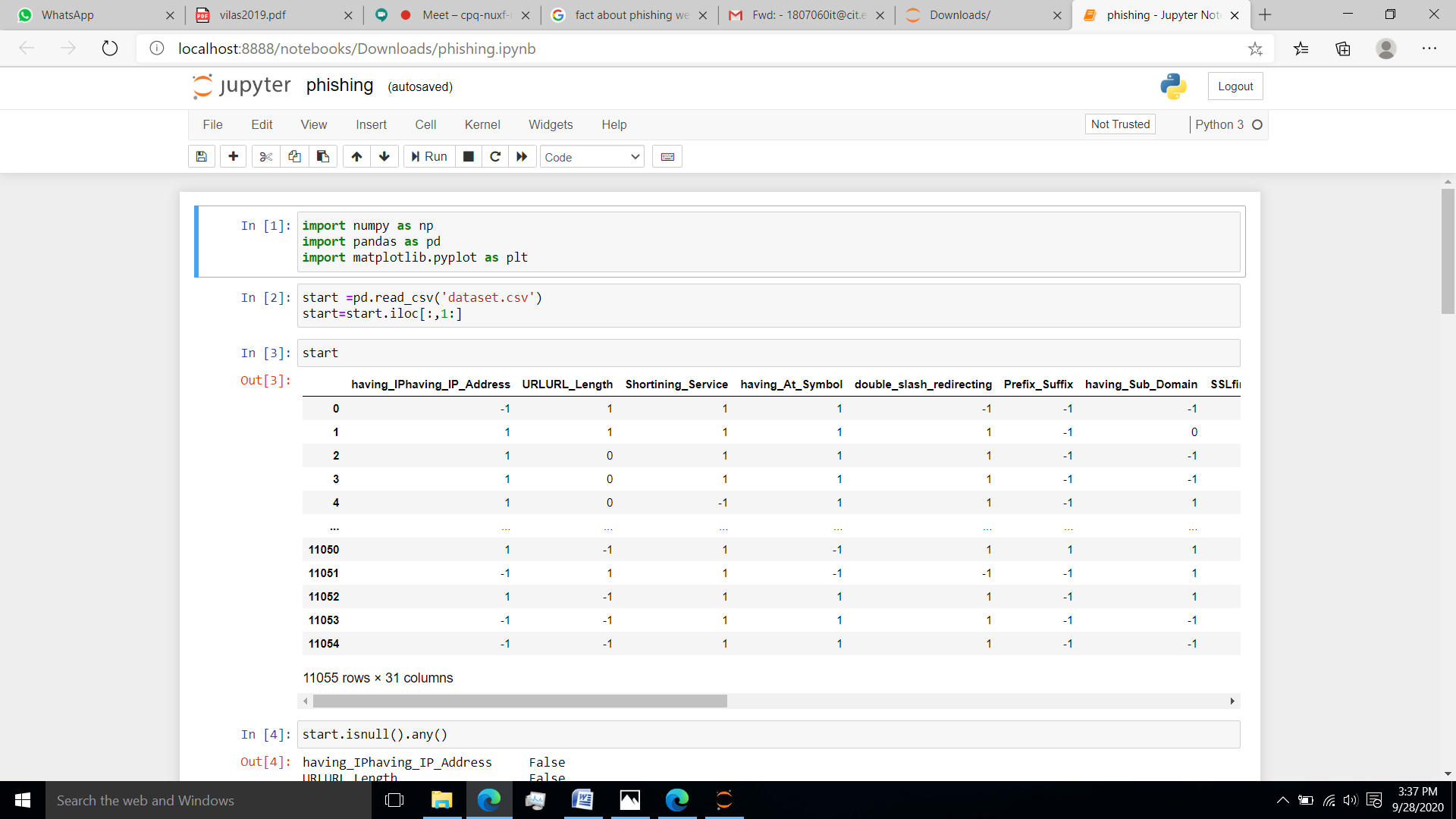
**HARDWARE/SOFTWARE DESIGNING:**

The steps followed in developing the model:

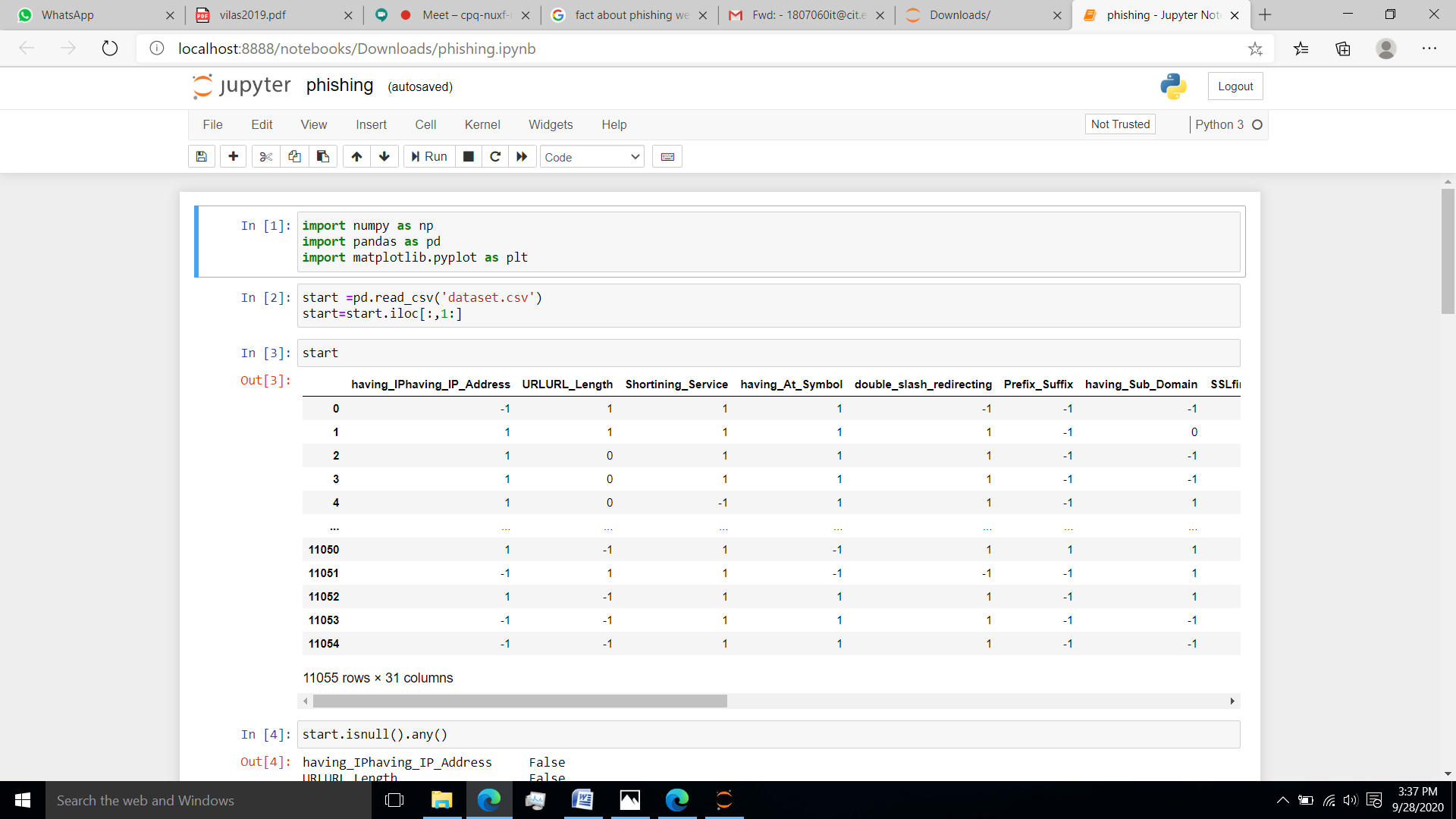
* Data Collection: The dataset was downloaded from the UCI ML Repository.
* Data Analysis: Evaluating cleanliness of the dataset by looking for any irrelevant data and handling missing data.
* Search for any trends, relations and correlations.
* Developing a model where the website can be identified to be phishing website or not.

**Importing the Libraries:**

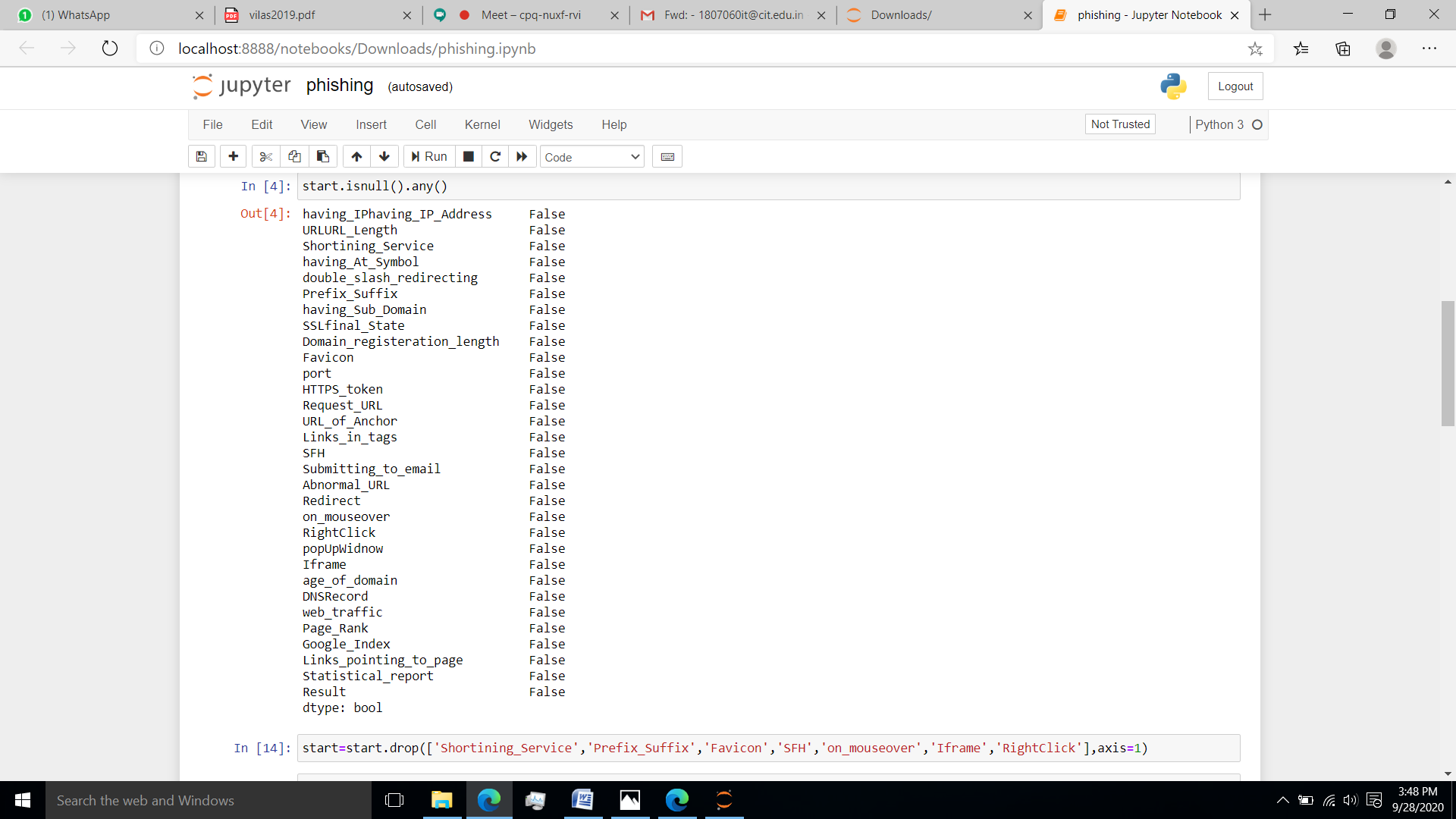
      Pandas,Numpy,Matplotlib



**Reading the .csv file:**



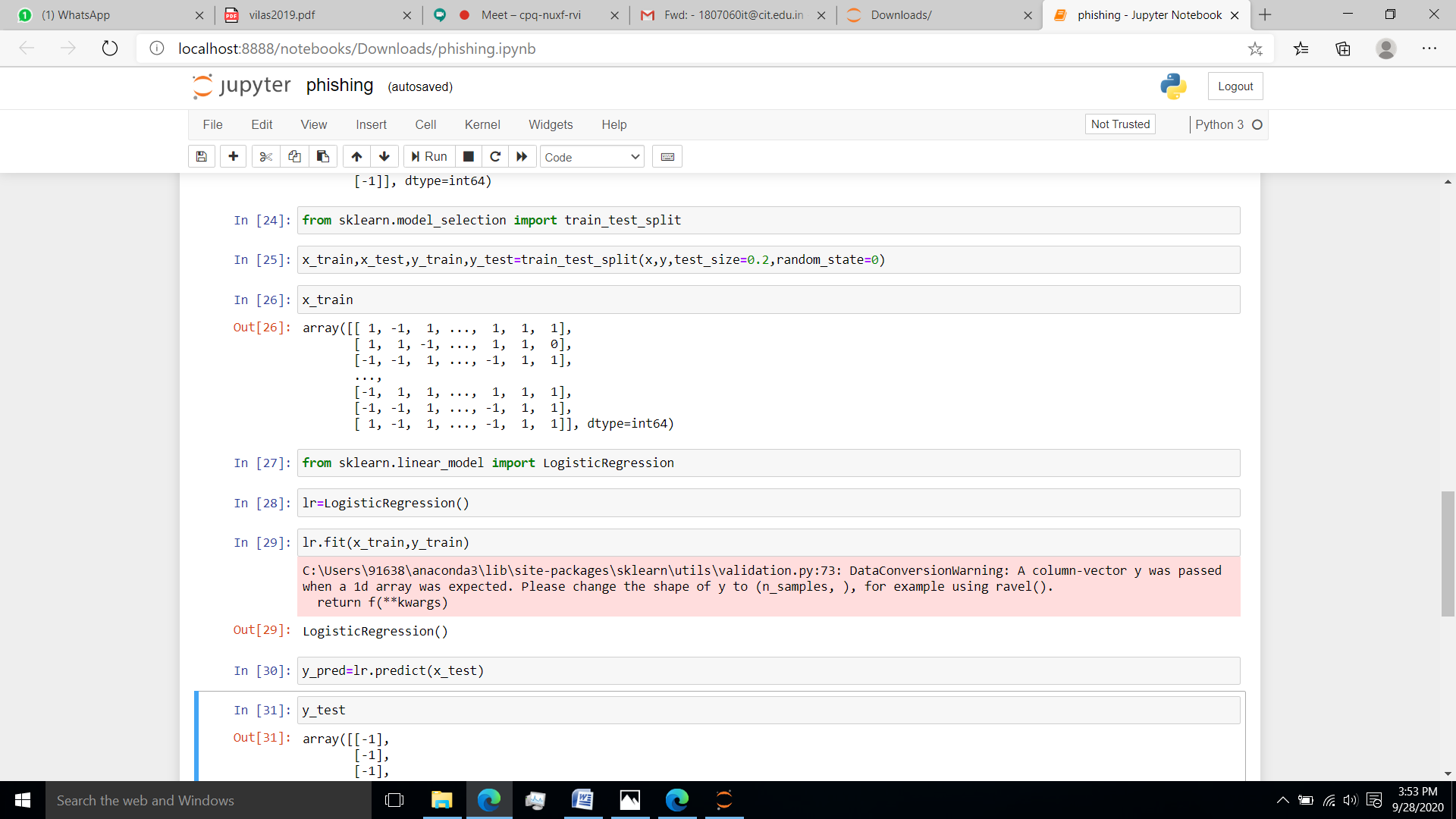
**Checking The Null Values:**

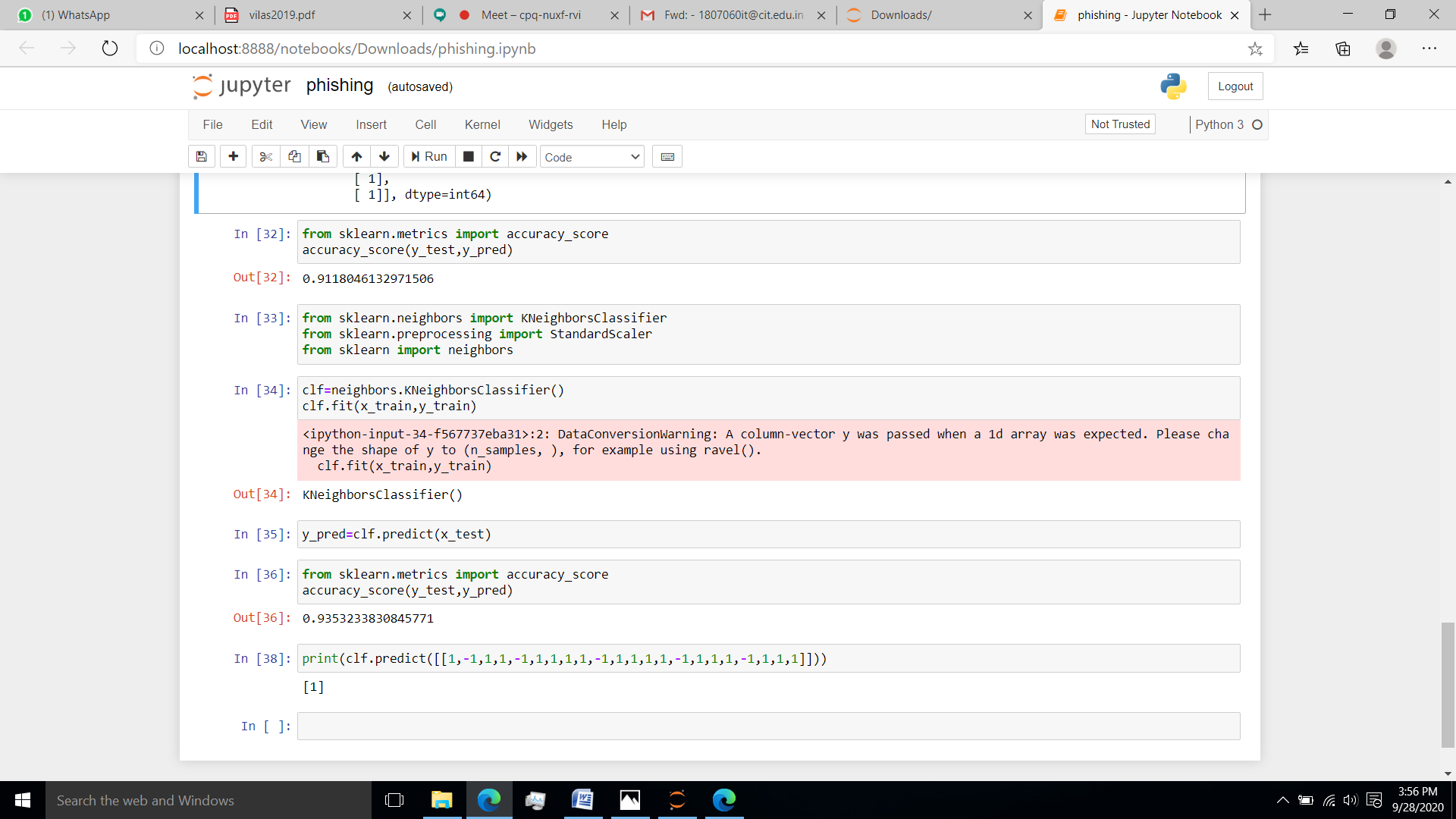


**Building a Model:**

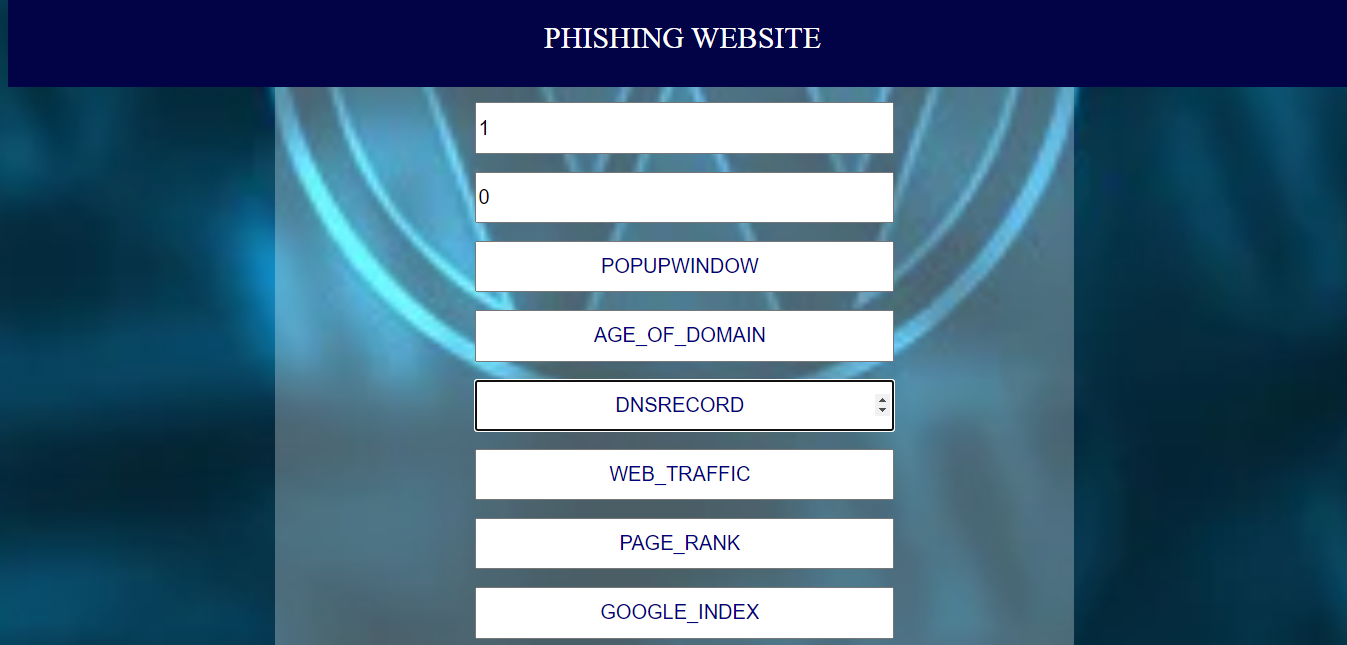
Since the outcome is binary and we have a reasonable number of examples at our disposal compared to number of features, this approach seems suitable. Since for this data, it already knows the output beforehand, it continuously adjusts the weights such that when these weights summed up with their features are introduced in the logistic function, the results are as near as possible to the actual ones.

Once presented with a test value, it again inserts the value into our logistic function and returns the output as a number between 0 and 2, which represents the probability of that test value being in a particular class.





Webpage:





**ADVANTAGES AND DISADVANTAGES:**

The benefits of this model are:

* Easy interface
* Straight forward results
* Accurate performance calculations

**Disadvantages:**

As our dataset is small, it’s training dataset is similar to test dataset. So it is difficult for the model to predict accurately for larger dataset.

**APPLICATIONS:**

This project makes it easier to predict whether the website is Phishing website or not.

**Future Scope:**

Database should be expanded on which the system will be tested much better. Also, the model requires further improvement mostly regarding feature selection of dectecting the phishing website into multiple components.

**Conclusion**:

In this project, we have proposed methods for detecting the website

whether is a phishing website or not using machine learning techniques. The two machine learning techniques that were used include Logistic Regression and KNN. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metrics. KNN was the model that resulted in the highest accuracy with an accuracy of 94%.

**Reference:**

1.Coresh J, Selvin E, Stevens L, et al. Prevalence of chronic kidney disease in the United States. JAMA. 2007;298:2038-2047.

**Result:**

Machine learning and artificial intelligence are powerful tools with the ability to improve understanding of predictive metrics in detecting the phishing website. We successfully trained and tested our model. This model was able to predict 0.935% of readmission rates accurately.