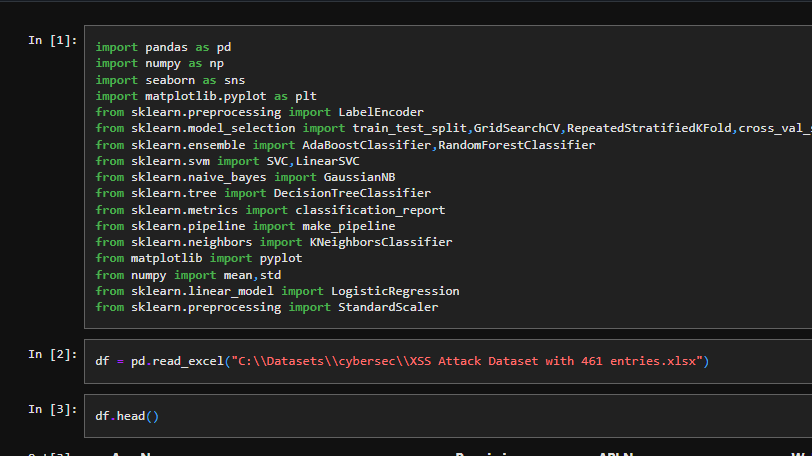
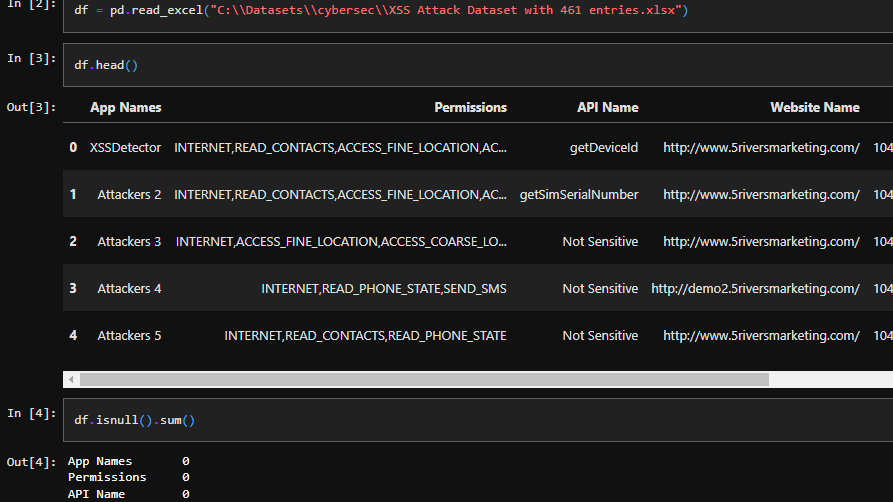
**Prediction of Cyber Attacks: Choosing the Best Model**

A cyber-attack is a malicious attempt by an individual to breach the information system of another individual or organization. Such attacks include Phising Attacks, Malware attacks and Dos Attacks.

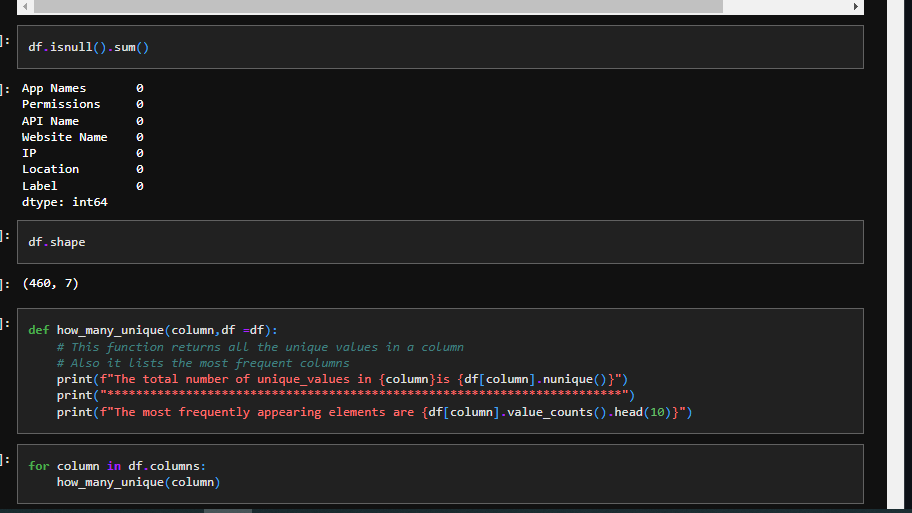
Given the adverse effects that this poses to any company system our task is to create a model that predicts cyber-attack cases using Artificial Intelligence. Choosing the best model is another challenge, the question is which one is the best? How can we tell? For this project I selected three models and used the power of model evaluation to solve this problem

Importing the necessary dependencies  
First look at the data



Unique columns

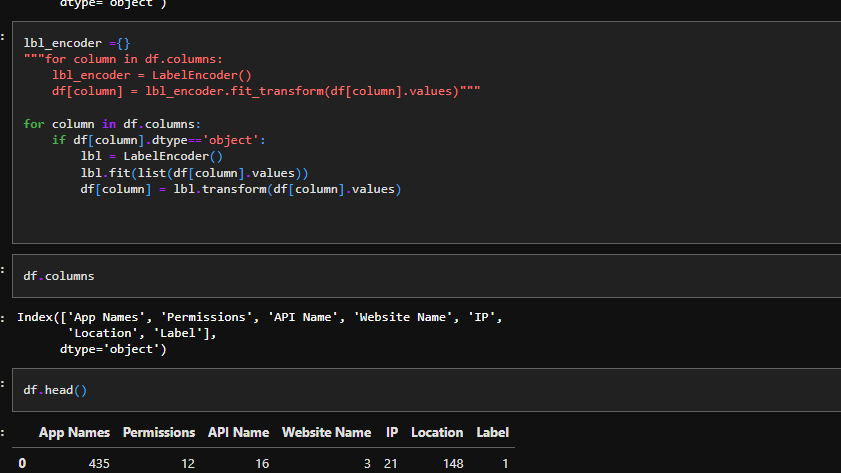
I created a custom function that helped me to get information on my data specifically the number of unique values in each column and the frequency of elements in my columns



Label Encoding

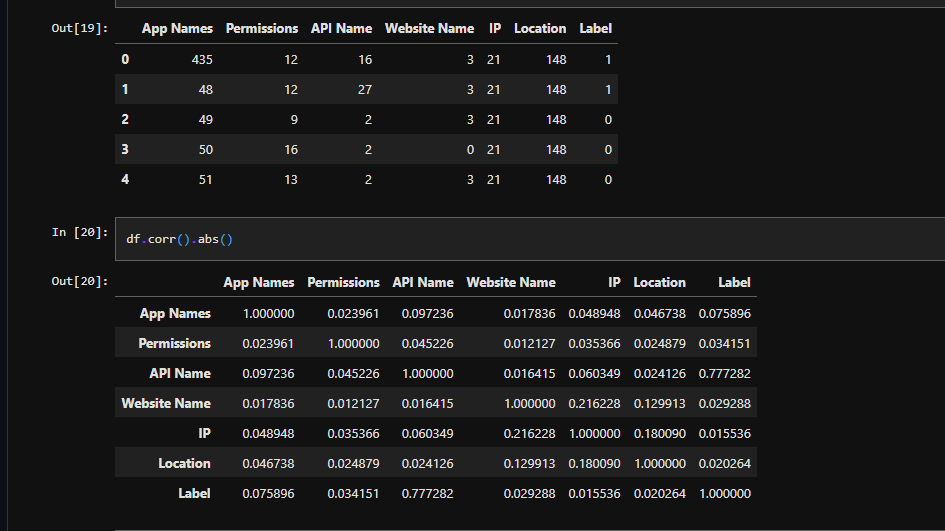
Since I had categorical data, I had to convert it into Numerical data, I did that using the Label Encoding

Technique, thankfully sklearn has such an algorithm, most machine learning models operate on numerical data.

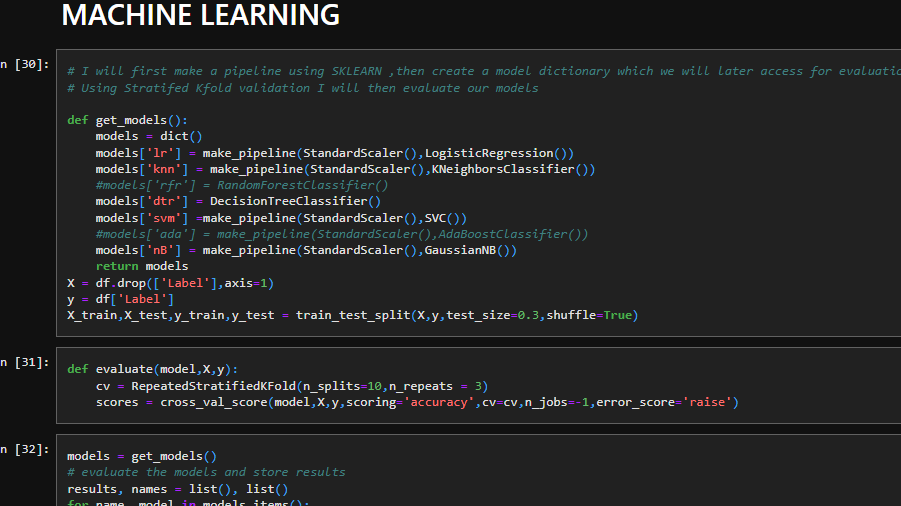


Checking correlations

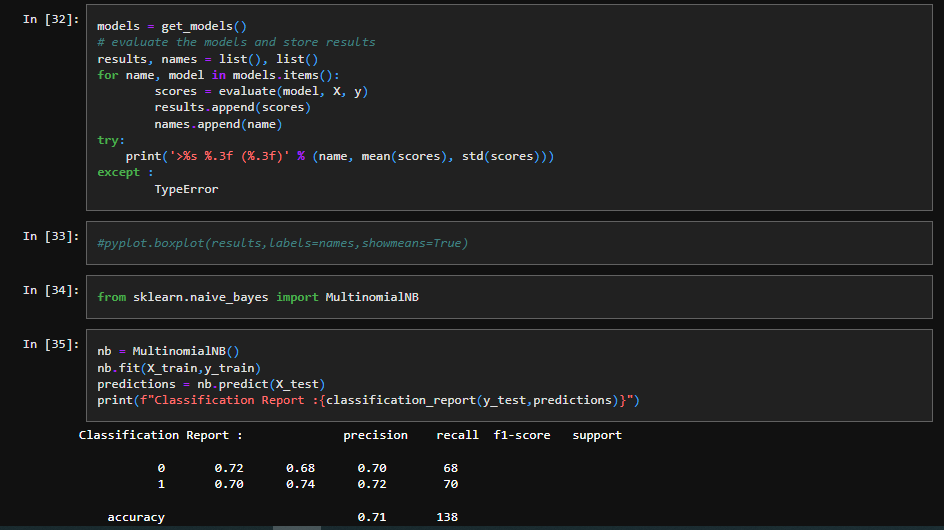
Before feeding my data into my machine learning models I checked for correlation between the variables .Correlation is a method of assessing a possible two-way linear association between two continuous variables it is measured by a statistic called the correlation coefficient, which represents the strength of the putative linear association between the variables in question. The value of the correlation coefficient lies between 0 and 1 with values closer to 1 indicating a greater linear association and closer to 0 representing a weak association.



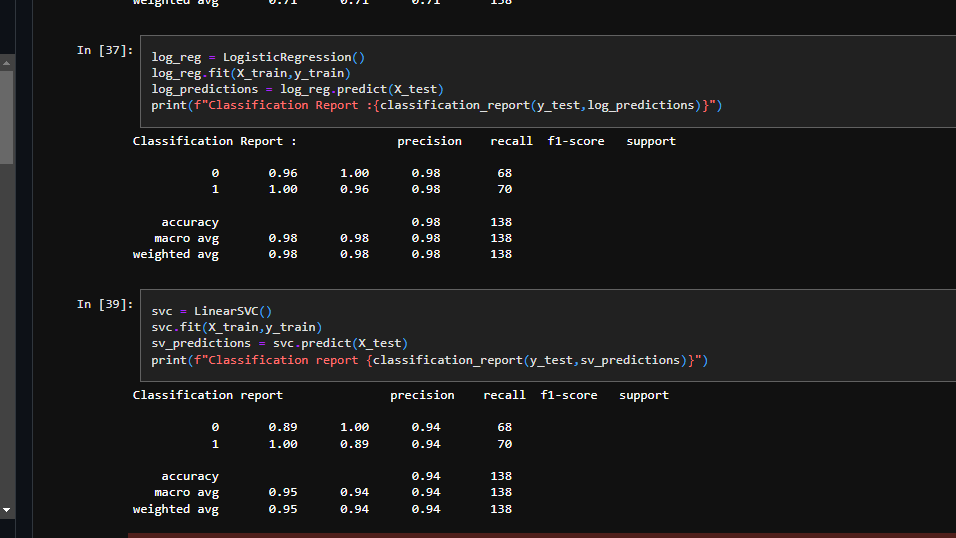
**Machine Learning**

Machine Learning entails computers learning from data, these algorithms improve by iterations through data without the need to be explicitly programmed. I first created a pipeline using Sklearn then proceeded to instantiate my models. Before doing machine learning it is a rule of thumb to split data into training and testing set. It is the testing data that we use to evaluate our models since this data hasn’t been seen yet.

I used Stratified K Fold sampling to mitigate problems caused by Random Sampling .In stratified K Fold ,the percentage of samples for each class are preserved, this ensures that each subgroup within the population receives proper representation within the sample and hence a better coverage of the population.



Logistic Regression and Naïve Bayes

Logistic Regression is a regression technique which has a binary outcome and has a logit applied such that the outcomes can either be 1 or 0. Naïve Bayes is a probabilistic algorithm that assumes independence between features .Naive Bayes is based on the Bayes Theorem and can be applied in a variety of problems such as Spam Detection .Linear Support Vector Machine (Linear SVC) is an algorithm that attempts to find a hyper plane to maximize the distance between classified samples. It then feeds the classifier some features to see what the prediction is 

Here were the results.

Logistic Regression had the highest accuracy 98% followed by Linear SVC at 95% .Multinomial Naïve Bayes performed the least with 78%.accuracy.