**INTRODUCTION**

* 1. **PROPOSED WORK**

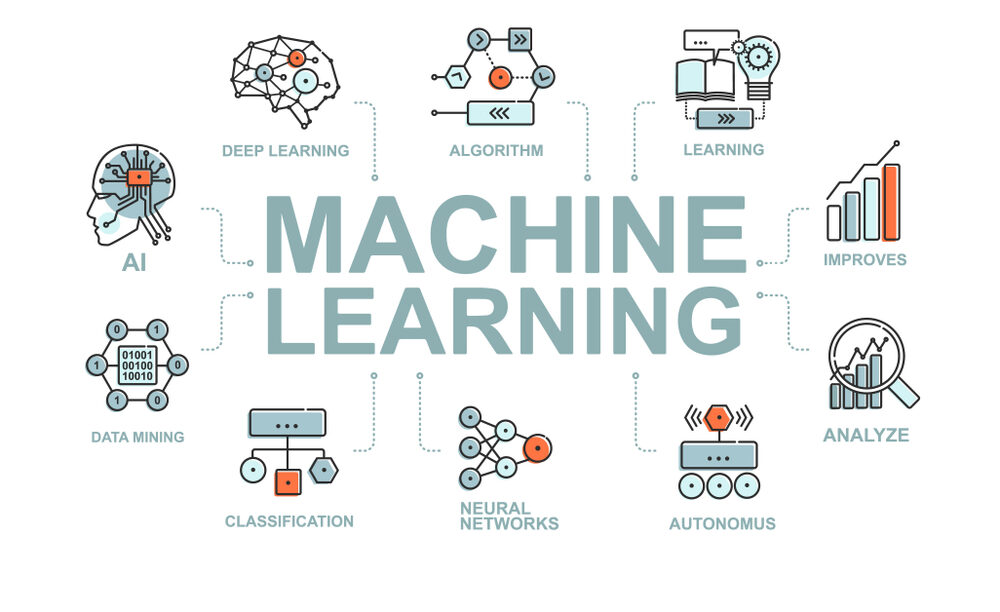
This project aims to develop a machine learning-based system for detecting and classifying malicious URLs. The system will use various models, including Logistic Regression, K-Nearest Neighbors, Naive Bayes, Random Forest, and Support Vector Classification, to identify and categorize URLs as phishing, spam, or malware. The system will be designed to detect disguised and hard-to-spot malicious URLs, which pose a significant threat to the digital world. The results of this project can contribute to improving online security and protecting users' personal and professional information from cybercrime.

* 1. **WHAT IS MACHINE LEARNING?**

Machine learning is a branch of artificial intelligence that focuses on the development of algorithms and statistical models that allow computer systems to learn and improve from experience, without being explicitly programmed. It is a data-driven approach where the computer system learns patterns and relationships from data, and uses that knowledge to make predictions or decisions on new, unseen data.

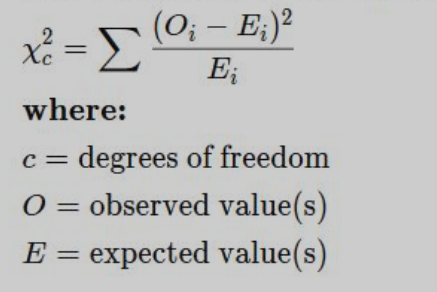
Machine learning algorithms can be classified into three categories: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the algorithm is trained using labeled data, where the desired output is already known. In unsupervised learning, the algorithm is trained on unlabeled data, and it tries to find patterns and relationships on its own. In reinforcement learning, the algorithm learns by trial and error, and it receives feedback in the form of rewards or penalties based on its actions.

Machine learning has many applications in various industries, including healthcare, finance, retail, and marketing. It is used for tasks such as image and speech recognition, natural language processing, recommendation systems, fraud detection, and predictive analytics. Machine learning has the potential to revolutionize many fields and make our lives easier and more efficient.



* 1. **DATA ANALYSIS**
     1. **Chi square analysis:**

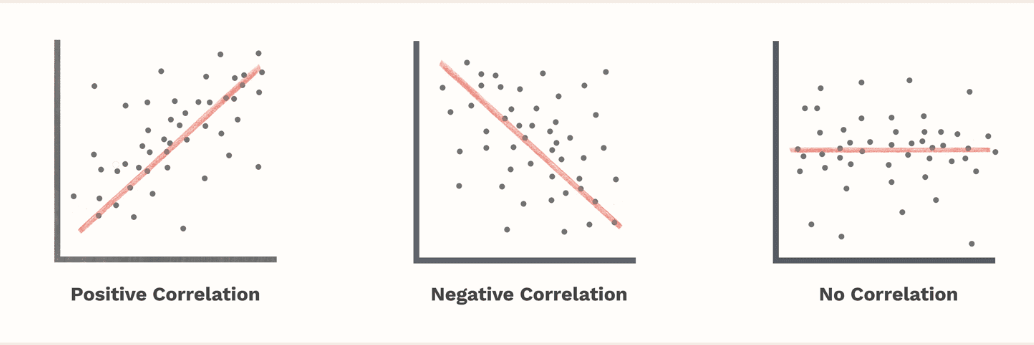
Chi-square analysis is a statistical method that determines whether there is a significant association or dependency between two categorical variables. It calculates a chi-square statistic based on the observed and expected frequency distribution of the variables. The result is a p-value, which represents the probability of obtaining the observed result by chance. If the p-value is less than the significance level, it indicates a significant association or dependency between the variables. This method is commonly used in social sciences, psychology, biology, and epidemiology to analyze categorical data and identify patterns and relationships.



* + 1. **Correlation Analysis:**

Correlation analysis is a statistical method used to measure the strength and direction of the relationship between two continuous variables. It determines how closely related the two variables are and whether they are positively or negatively related.

The correlation coefficient is a numerical value that ranges from -1 to 1 and represents the strength and direction of the relationship between the variables. A coefficient of 1 indicates a perfect positive correlation, where the two variables move in the same direction. A coefficient of -1 indicates a perfect negative correlation, where the two variables move in opposite directions. A coefficient of 0 indicates no correlation between the variables.



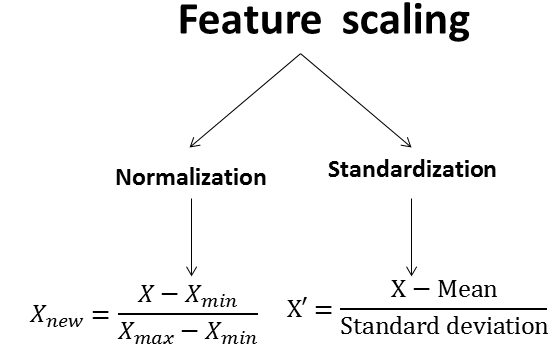
* 1. **FEATURE SCALING**

Feature scaling is a technique used in machine learning to normalize the range of values of independent variables or features of a dataset. It is done to ensure that no variable dominates the others based on their scale, and to improve the performance and accuracy of certain machine learning algorithms.

For example, if a dataset contains features with different scales, such as age and income, then the algorithm may give more weight to income than age, simply because the income values are larger. This can negatively affect the accuracy of the model.

There are several methods of feature scaling, including normalization, standardization, and scaling to unit length. Normalization scales the data so that it ranges between 0 and 1. Standardization scales the data so that it has a mean of 0 and a standard deviation of 1. Scaling to unit length scales the data so that the length of each sample vector is 1.

Overall, feature scaling is an important pre-processing step in many machine learning applications, particularly in cases where the features have different scales and units.



* 1. **MACHINE LEARNING MODELS**
     1. **Logistic Regression**

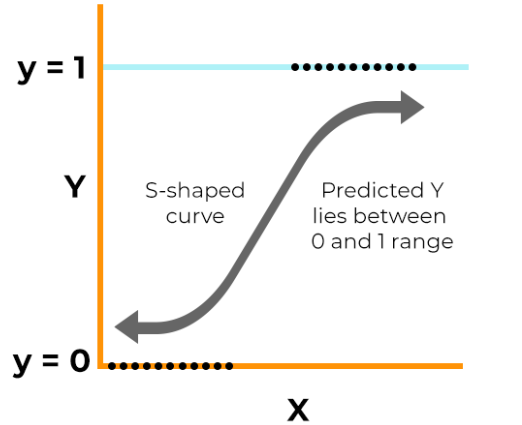
Logistic regression is a type of supervised machine learning algorithm that is used for binary classification problems, where the outcome variable is categorical and has only two possible outcomes, such as yes/no or true/false.

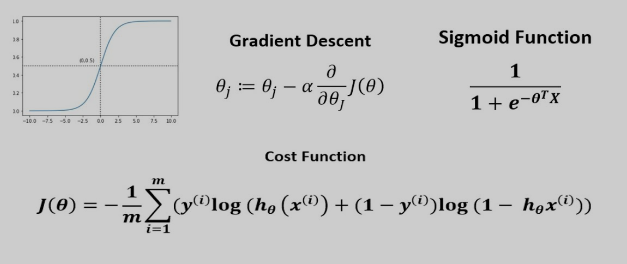
The goal of logistic regression is to predict the probability of an outcome based on one or more input variables or features. It uses a logistic function (also known as the sigmoid function) to transform the input variables into a probability value between 0

and 1, which represents the likelihood of the outcome being positive.

Logistic regression works by fitting a logistic regression model to the training data, which involves estimating the coefficients of the input variables that best fit the data. The model can then be used to make predictions on new data by applying the learned coefficients to the input variables and using the logistic function to calculate the probability of the outcome.

In summary, logistic regression is a simple yet powerful algorithm for binary classification problems that predicts the probability of an outcome based on input variables or features.





* + 1. **K Nearest Neighbors**

K-Nearest Neighbors (KNN) is a type of supervised machine learning algorithm used for classification and regression tasks. The algorithm is based on the idea that data points with similar attributes tend to be clustered together in space.

In KNN, the k nearest neighbors to a new data point are determined from the training data, based on their distance to the new data point. The distance can be calculated using various metrics such as Euclidean distance or Manhattan distance.

For classification tasks, the class label of the new data point is then determined by majority vote of the class labels of its k nearest neighbors. For regression tasks, the output value is the average of the values of the k nearest neighbors.

The choice of k is an important parameter in KNN, as it can significantly affect the performance of the algorithm. A smaller value of k can lead to overfitting and a larger value of k can lead to underfitting.

KNN is a simple yet effective algorithm that can be used for a wide range of classification and regression tasks. However, it can be computationally expensive and may not work well with high-dimensional data.



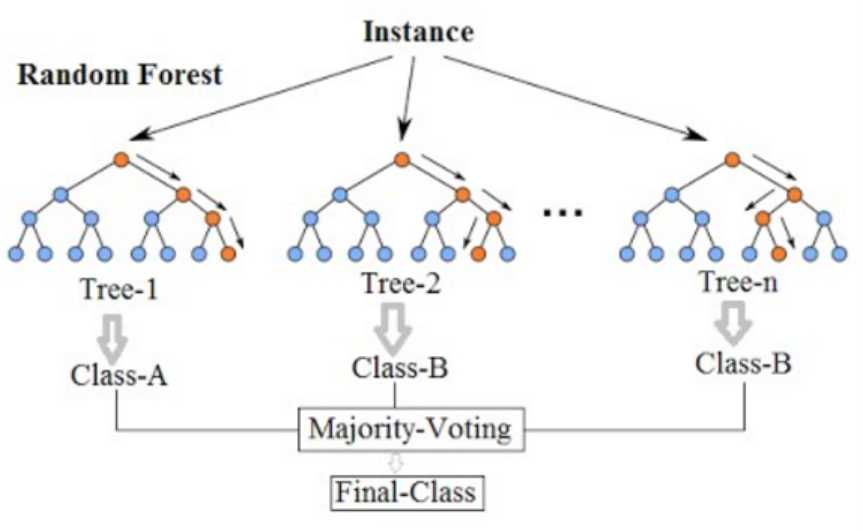
* + 1. **Random Forest**

Random forest is a type of supervised machine learning algorithm used for classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to make more accurate predictions.

In random forest, multiple decision trees are constructed using a random subset of the training data and a random subset of the input variables. The decision trees are then combined to make predictions, with each tree contributing a vote to the final prediction.

The randomization of the data and variables used in each decision tree helps to reduce overfitting and increase the accuracy of the model. Additionally, random forest can handle both numerical and categorical data and can also provide information on the importance of the input variables.

Random forest is a powerful algorithm that can be used for a wide range of classification and regression tasks. However, it can be computationally expensive and may not work well with high-dimensional data.



* + 1. **Support Vector Machine**

Support Vector Machines is a type of supervised machine learning algorithm used for classification and regression tasks. The goal of SVM is to find a hyperplane that separates the data into different classes with the largest margin possible.

In SVM, the hyperplane is determined by finding the support vectors, which are the data points closest to the hyperplane. The distance between the support vectors and the hyperplane is known as the margin, and the goal of SVM is to maximize the margin.

If the data is not linearly separable, SVM can still be used by transforming the data into a higher-dimensional space using a kernel function. This allows the data to be separated by a hyperplane in the higher-dimensional space, which corresponds to a nonlinear separation in the original feature space.

SVM is a powerful algorithm that can handle both linear and nonlinear classification tasks. However, it can be computationally expensive and sensitive to the choice of hyperparameters. Additionally, SVM can be affected by the presence of outliers in the data.



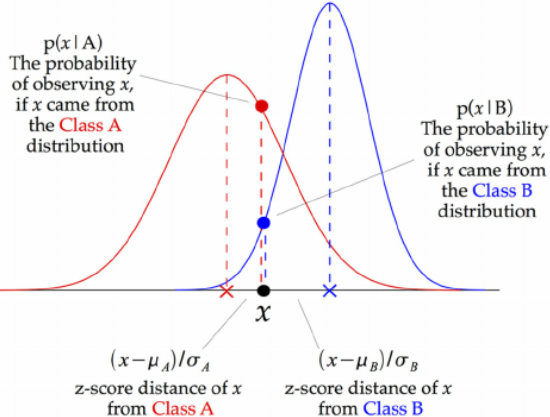
* + 1. **Gaussian Naïve Bayes**

Gaussian Naive Bayes is a type of supervised machine learning algorithm used for classification tasks. It is based on Bayes' theorem and assumes that the input variables are independent and normally distributed.

In Gaussian Naive Bayes, the algorithm first calculates the prior probability of each class based on the training data. Then, for a new data point, the algorithm calculates the conditional probability of each class given the input variables using the Gaussian probability density function.

The algorithm selects the class with the highest probability as the predicted class for the new data point.

Gaussian Naive Bayes is a simple and fast algorithm that can handle high-dimensional data and is relatively insensitive to irrelevant input variables. However, it assumes that the input variables are independent, which may not be true in practice.



* 1. **PROPOSED METHOD**

The proposed method consists of following phases 1. Feature Extraction 2. Feature reduction 3. Train the Model 4. Testing.

Feature extraction is the primary phase for machine learning techniques. UNB dataset 2016 is used for feature extraction and experiments with different machine learning algorithms.

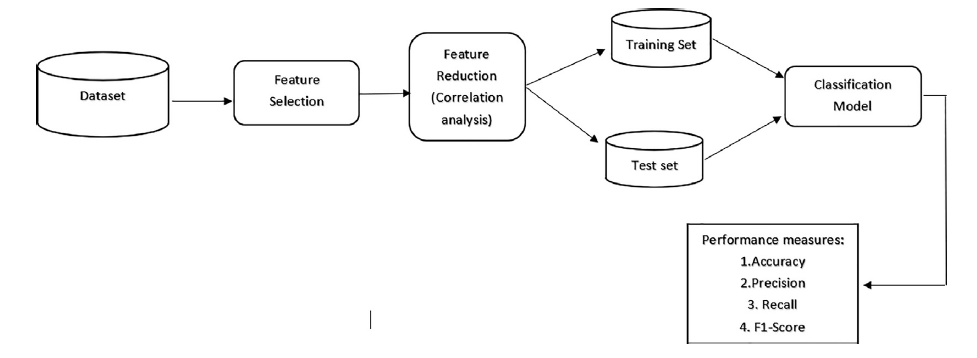
The UNB dataset includes benign, phishing, malware, spam url dataset.

Minimum set of optimal features which reduces the execution time and storage consumption. Initially identified 18 most common features and 9 newly identified features (\*) in the dataset.

All extract features are not always suitable for classification. So it requires feature reduction method to identify the fitness of the features. Correlation analysis help to find the relationship between target feature and other features which in turn help to reduce the feature. Features are removed if there is no significant correction is present.

Nearly 20 features are selected after correction analysis and 7 features are removed from features set (Underscores \_ in\_Domain, Double-slashes \_in\_URL, At(@)\_in\_URL, Hash(#) \_in\_URL, Https\_in\_URL, Numbers\_\_in\_URL,

Special\_char\_ratio\_in\_URL).

****