**Week 2: Delving more into the statistical insights that you can draw from the data, feature engineering and feature selection**

1. Import the necessary libraries.
2. Run the Jupyter notebook from week 1 as a module and use its code in the new Jupyter notebook.

**Part A**: **Statistical Insights**: Statistical insights are important because at the time of validation and refining the machine learning models, it quantifies the performance of model and avoid overfitting.

Steps to be followed while carrying out the statistical insights from the train and test datasets:

1. Extracting all the column names as a list.
2. Get the descriptive statistics for both the datasets.
3. We break down the descriptive statistics by calculating each concept separately like Max, Mean, Standard Deviation, Min, Quantile, Mode, Median, Variance
4. Obtaining the position where the largest value of a particular column exists.
5. Summarize the datasets.
6. Finding the skewness and kurtosis of both the datasets.

* Skewness is the degree of symmetry, or more precisely, the degree of lack of symmetry.
* Kurtosis is the proportion of data that is heavy-tailed or light-tailed in comparison with a normal distribution.

**Part B**: **Feature Engineering**:

* It is the pre-processing step which is used to transform raw data into features that can be used for creating a predictive model using Machine Learning or Statistical Modelling. Different steps are carried out while performing feature engineering.
* Feature engineering is important since it influences how machine learning models perform and help to figure out the hidden patterns in the data.

1. Data Transformation:

* Data Transformation is done because it can be used as a classifier for a simple binary model that identifies any attack.
* Since it is binary classification problem, perform transformation on class label and encode 'normal' as 0 and 'anomaly' as 1.
* This step will help us to identify intrusion detection.
* In this step, a new feature class\_transform is derived from the feature class.
* After generating the new feature, we compute the cross tabulation on different categorical variables.
* Cross tabulation shelp to display the frequency distribution of categorical variables.
* It helps in exploring relationships and patterns between two or more categorical variables.
* It also provides valuable insights for making data-driven decisions.

1. One Hot Encoding:

### One Hot Encoding can be defined as a process of transforming categorical variables into numerical features that can be used as input to Machine Learning.

### By using one-hot encoding we can reduce bias, improve accuracy and interpretability.

### It is converting categorical variables into dummy variables.

### Here we check for the categorical features in train and test dataset.

### Train dataset has more categorical features than the test dataset, so we calculate the difference between both.

### Generate a new dataframe after finding the difference and reorder the columns.

### Append the new columns to the test encoded dataframe and generate the final test data.

1. Label Encoding:

* **Label Encoding** is a technique that is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which only take numerical data.
* It is used since it simplifies the data, reduces the memory usage, makes it flexible during data preprocessing.
* Here categorical features like 'flag', 'service', 'protocol\_type', 'class', 'land', 'logged\_in', 'is\_host\_login', 'is\_guest\_login', 'class' are converted into numerical variables.

1. Feature Scaling/Standardization and Normalization:

* Variables present in the dataset are measure at different scales and do not contribute equally to the fit of the model. This may create a bias. So, to deal with this problem standard scalar is used.
* Feature scaling is a data preprocessing technique used to transform the values of features in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.
* Normalization is a data preprocessing technique used to adjust the values of features in a dataset to a common scale. This is done to reduce the impact of different scales on the accuracy of machine learning models. Min-max normalization preserves the relationships among the original data values.
* Standardization is where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

**Part C**: **Feature Selection**

* Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features.
* It is required to make the process more accurate, increase the prediction by selecting most crucial variables and removing the redundant ones.

1. **Filter Methods**: In Filter Method, features are selected based on statistics measures.
2. **Numerical Input, Categorical Output**:

* Feature selection classification problem with numerical input variables using a correlation coefficient, with the categorical target.

### Pearson's correlation coefficient:

### The Pearson correlation coefficient can be used to summarize the strength of the linear relationship between two data samples.

### It is used because we can test whether there is a significant relationship between two variables.

### Pearson’s Correlation Coefficient is a measure of quantifying the association between the two continuous variables and the direction of the relationship with its values ranging from -1 to 1.

### In the following problem, we only use the numerical variables as input and keep class\_transform as target to compute the correlation.

### Kendall's correlation:

### It is a statistic used to measure the ordinal association between two measured quantities.

### It is used when data doesn’t meet one of the requirements of Pearson’s correlation.

### Kendall correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e., relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully different for a correlation of −1) rank between the two variables.

### Here also, we only use the numerical variables as input and keep class\_transform as target to compute the correlation.

## **Categorical Input, Categorical Output**:

## It is a classification predictive modeling problem with categorical input variables.

## We can use a correlation coefficient, such as Chi-Squared test (contingency tables) or Mutual Information.

1. Chi-squared test:

* The chi-square test is a statistical method that can be used for feature selection in machine learning.
* It is used because it checks how well the observed values for a given distribution fit with it when the variables are independent.
* It is used to determine whether there is a significant association between two categorical variables.
* In this problem, we select we only consider categorical variables as input and keep class\_transform as target.
* Then we use SelectKBest algorithm with number of top features to select.
* It finds the mask of the selected features and the respective scores.

1. Information Gain/Mutual Information:

* Information Gain measures the reduction in entropy or surprise by splitting a dataset according to a given value of a random variable.
* It helps to decide whether a feature should be used to split a node or not.
* Mutual information is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable.
* Here we use mutual\_info\_classif to generate feature importance for all the input variables which we selected.
* Plot the feature importance using bar plot.

1. **Wrapper methods**: In wrapper methodology, selection of features is done by considering it as a search problem, in which different combinations are made, evaluated, and compared with other combinations.
2. Forward Feature Selection:

* Forward selection is an iterative process, which begins with an empty set of features.
* It is used to identify the most relevant variables for predictive modeling.
* After each iteration, it keeps adding on a feature and evaluates the performance to check whether it is improving the performance or not.
* The process continues until the addition of a new variable/feature does not improve the performance of the model.
* Here we use the train dataset for specifying input and target variables.
* Random Forest Classifier is used for feature selection.
* Sequential Feature Selector is used with parameters like forward=True, scoring=’accuracy’, cv=5, verbose=2, k\_features=5, etc.

1. Backward Elimination:

* It is the opposite of forward selection. It finds the best subset of features from a given set of features.
* This technique begins the process by considering all the features and removes the least significant feature.
* This elimination process continues until removing the features does not improve the performance of the model.
* Here, we use the train dataset for specifying input and target variables.
* Random Forest Classifier is used for feature selection.
* Sequential Feature Selector is used with parameters like forward=False, scoring=’accuracy’, cv=5, verbose=2, k\_features=5, etc.

1. Recursive feature elimination:

* Recursive feature elimination (RFE) is a recursive greedy optimization approach, where features are selected by recursively taking a smaller and smaller subset of features.
* RFE avoids overfitting by cross-validation during the feature selection process.
* Here we use the train dataset for specifying input and target variables.
* Random Forest Classifier is used for feature selection.
* RFE is used with parameters like estimator, n\_features\_to\_select=2, step=1, etc.

1. Exhaustive feature selection:

* The performance of a machine learning algorithm is evaluated against all possible combinations of the features in the dataset.
* The feature subset that yields best performance is selected.
* The exhaustive search algorithm is the greedy algorithm of all the wrapper methods because it tries all the combination of features and selects the best.
* It is used because it can go over all the possible feature combinations and pick up the model with the highest accuracy.
* Here we use we use the numerical and categorical train dataset for specifying input and target variables.
* ExhaustiveFeatureSelector () method is used with Random Forest Classifier model and parameters like min\_features =1, max\_features =2, cv=2, scoring=’roc\_auc’, etc.
* Then we filter the best features and calculate the best score.

1. **Embedded methods:** It has the benefits of both the wrapper and filter methods by including interactions of features but also maintain reasonable computational costs.
   1. Lasso Regularization (L1):

* It adds a penalty to the different parameters of the machine learning model to reduce the freedom of the model.
* It shrinks some of the coefficients to zero. Then that feature can be removed from the model thereby avoiding the overfitting on the training data.
* It is preferred when we have many features as it provides sparse solutions.
* We use the logistic regression model with penalty=’l1’, solver=’liblinear’
* Next, we choose SelectFromModel as a part of feature selection and include logistic regression.
* Then we transform the input variables. It retains certain features after re`moving the features with values as zeros.
  1. Random Forest Importance:
* It is tree-based strategy used by random forests rank by how well they improve the purity of the node.
* It is reducing overfitting by averaging multiple decision trees and is less sensitive to noise and outliers in the data.
* Here, we use the Random Forest classifier model and fit it on the input and target variables of the train dataset.
* Then we find the feature importance and visualize them in a bar plot.