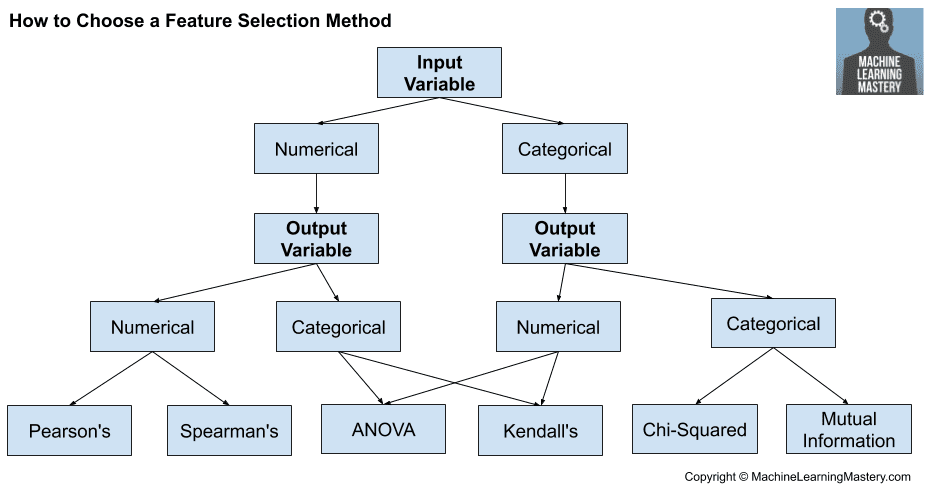
ARYA.AI ASSIGNMENT

Aim : Binary Classification - Do an exploratory analysis of the dataset provided, decide on feature selection, preprocessing before training a model to classify as class ‘0’ or class ‘1’.

Steps : The dataset consists of 2 dataset (training set and test set) training data has a dimension of 3910x58 including 1 dependent parameter. Initially export the dataset to jupyter notebook and start some basic EDA for analyzing. Though the dataset is rich with numeric data.

Step 2 : Feature Selection –



a) we need to find the relationship between 57 columns w.r.t to Initially I used correlation methods to find that relationship and from there some columns are removed to make out points to fall more quickly to global minima.

b) try out some feature importance to remove constant columns which possess no relation to y field

c ) Mutual Information is carried out to each feature and plot the graph.

Step 3 : Splitting – then the refined data got splitted for making model predictions, I have used random forest classifier to make predictions, They are so successful because they provide in general a good predictive performance, low overfitting and easy interpretability. This interpretability is given by the fact that it is straightforward to derive the importance of each variable on the tree decision. In other words, it is easy to compute how much each variable is contributing to the decision.

Accuracy :

Model got trained with random forest classifier and consider 2 other sersions too,each version is depends upon the deatures which got selected. Accuracy = 95%

Why Feature Selection is Needed ?

Feature selection is **the process of reducing the number of input variables when developing a predictive model**. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.It enables ML Algorithm to train faster,reduces complexity of model and makes it easier to interpret.If the **right subset** is chosen it improves the **Accuracy** too.

Why Random Forest Classifier Used ?

One of the biggest advantages of random forest is its versatility. It can be **used for both regression and classification tasks**, and it's also easy to view the relative importance it assigns to the input features. If there are enough trees in the forest, the classifier won't overfit the model.It can handle large datasets efficiently it uses ensemble methodologies.