**Systemia nx ML Internship**

**Assignment 1** NIKILHARI.R

22CSR132

List of terminologies:-

Sample data about whether the loan has been granted or not.



* **Feature:-** The inputs for a machine learning modelor attributes of the data sets. In this dataset, features include Gender, Married, Dependents, Education, etc.
* **Label:**- The attribute which represents the output of the machine learning algorithm or which we will predict is called label.Here Loan status is label.
* **Prediction:-** What we are going to predict using our model .Here it is Loan status
* **Outlier:**- In the data set the co applicant income is 0 for someone and 10,986 for some others the significant change is called outlier.
* **Test Data:**- A subset of the data used to test the model. This data is not used during the training phase.
* **Training Data:-** The data used to train the model.
* X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, random\_state=0, train\_size = .75)
* In the above code 0.75 indicates that 75% data is used for training and remaining 25% is used for testing
* **Model:-** An algorithm or mathematical representation used to predict the output.
* **Overfitting:-** When our model works well for the training data(given data) but performs poorly for new datas is called overfitting
* **Underfitting:-** When our model performs poor even in training and test data then it is called overfitting.
* **Validation Data:-** Subset of data other than training and test data which is used to tune our model to prevent overfitting.
* **Hyperparameter:-** Parameters that are explicitily given by us to tune the model in the above mentioned example train\_test\_split train size is a hyperparameter.
* **Epoch:-** Like loop , an hyperparameter that determines number of times the training data undergoes through the learning algorithm.
* **Loss Function:**- A function that calculates the difference between the prediction and actual true value. Loss function should be minimum
* **Learning Rate:**- In order to get the minimum loss function value we will iterate through a loop the length of the step in each iteration is given as a learning rate as a huperparameter.
* **Regularization:-** Techniques used to prevent overfitting
* **Cross-Validation:** the data is split into k equal parts where 1 part is used as validation data and other k-1 part is used for training ,this will run for k times.
* **Feature Engineering:-** The process of creating new feature from the existing one for example in the above data set from applicant income and co applicant income we can form a new feature total income such that  
   TOTAL\_INCOME= APPLICANT INCOME + CO APPLICANT INCOME
* **Dimensionality Reduction:-** Techniques used to reduce the number of features in a dataset. If two features are similar their impact will on the outcome will also be same hence we can avoid one of the feature.
* **Bias:** When the model is so simple that it does not include all the features, which causes poor performance on even training data results in underfitting
* **Variance:** When the model is so complex it reacts to even the small variations which causes overfitting
* Ideal model has low bias and low variance.