# Workshop Day 5

## Case Study 1

# Machine learning project in python to predict loan approval

We have the dataset with the loan applicant's data and whether the application was approved or not.

## Steps involved in this machine learning project:

Following are the steps involved in creating a well-defined ML project:

- Understand and define the problem
- Analyse and prepare the data
- Apply the algorithms
- Reduce the errors
- Predict the result

These 2 steps need to be performed in the day4 workshop itself

These 3 steps need to be performed in the day5 workshop

## Predict if the loan application will get approved

- We have the loan application information like the applicant's name, personal details, financial information and requested loan amount and related details and the outcome (whether the application was approved or rejected). Based on this we are going to train a model and predict if a loan will get approved or not.

# Splitting the Data set

As we have seen already, In Machine learning we have two kinds of datasets

- Training dataset - used to train our model
- Testing dataset - used to test if our model is making accurate predictions

Our dataset has 480 records. We are going to use 80% of it for training the model and 20% of the records to evaluate our model. copy paste the below commands to prepare our data sets

Though our dataset has lot of columns, we are only going to use the Income fields, loan amount, loan duration and credit history fields to train our model.

**Refer to the file Day 5 Session 1 Loan Approval Prediction Machine Learning Model.ipynb (demo given in Day5 session1)**

**Tabulate the accuracies of the following ML Models for Loan Approval Prediction ML algorithm by changing different parameters and percentages of training and testing data sets as shown below and conclude which ML model with percentages of training and testing percentages offers the best accuracy.**

**Model 1 (refer to datasets of training and testing in the demo pdf)**

| **ML Algorithm** | **Training dataset – 90%**<br>**Testing dataset – 10%** | **Training dataset – 70%**<br>**Testing dataset – 30%** | **Training dataset – 60%**<br>**Testing dataset – 40%** |
|---|---|---|---|
| Logistic Regression | 0.72916 | 0.79166 | 0.79687 |
| Decision tree | 0.70833 | 0.72222 | 0.72395 |
| Random forest | 0.72916 | 0.75694 | 0.77604 |

**Model 2 (refer to datasets of training and testing in the demo pdf)**

| ML Algorithm | Training dataset – 90%<br>Testing dataset – 10% | Training dataset – 70%<br>Testing dataset – 30% | Training dataset – 60%<br>Testing dataset – 40% |
|---|---|---|---|
| Logistic Regression | 0.68750 | 0.65277 | 0.63541 |
| Decision tree | 0.68750 | 0.63888 | 0.61979 |
| Random forest | 0.68750 | 0.63888 | 0.62500 |

**Model 3 (refer to datasets of training and testing in the demo pdf)**

| ML Algorithm | Training dataset – 90%<br>Testing dataset – 10% | Training dataset – 70%<br>Testing dataset – 30% | Training dataset – 60%<br>Testing dataset – 40% |
|---|---|---|---|
| Logistic Regression | 0.72916 | 0.79166 | 0.80208 |
| Decision tree | 0.60416 | 0.72222 | 0.69791 |
| Random forest | 0.64583 | 0.77083 | 0.76562 |

## Results:

**Model 1 is the best model as compared with Model 2 and Model 3. This is because it has more accuracy compared to two other models.**

**Logistic Regression is the best ML algorithm as compared with Decision tree and Random forest. This is because it has more accuracy compared to two other ML algorithms.**