COM7039M Machine Learning

## Student Name

## Student ID

Table of Contents

[Student Name 1](#_Toc122272957)

[Student ID 1](#_Toc122272958)

[Data Collection 3](#_Toc122272959)

[Data Preparation 3](#_Toc122272960)

[Data pre processing 4](#_Toc122272961)

[Data Visualization 6](#_Toc122272962)

[Processing Data for Machine Learning 9](#_Toc122272963)

[Applying the ML model 11](#_Toc122272964)

[Figure 1 Dataset information 3](#_Toc122272990)

[Figure 2 Dataset information 4](#_Toc122272991)

[Figure 3 Null Values in each columns 4](#_Toc122272992)

[Figure 4 Segrigating data into numeric and categorical 5](#_Toc122272993)

[Figure 5 Filling null value with mean as per each columns 5](#_Toc122272994)

[Figure 6 filling cat columns with most occurring data i.e mode 6](#_Toc122272995)

[Figure 7 Checking skewness of each columns 7](#_Toc122272996)

[Figure 8 Corelation plot 8](#_Toc122272997)

[Figure 9 Box plot for checking outliers 9](#_Toc122272998)

[Figure 10 Label Encodeing 10](#_Toc122272999)

[Figure 11 Merging the data into single dataset 10](#_Toc122273000)

[Figure 12 Splitting the data into training and testing 10](#_Toc122273001)

[Figure 13 Applying Decesion Tree 11](#_Toc122273002)

[Figure 14 Parameters tuning for achieving more accuracy 11](#_Toc122273003)

# Data Collection

The dataset is of unknown loan provider bank or financial firm which is havin columns related to the particular loan application and their applicant information like credit history mortgage land area type and the application amount for the particular loan application the idea here is to analize the dataset to make the prediction weather the loan will going to be approved or not as per the provided information of the particular loan application.

# Data Preparation

The data set here consist of the columns like loan id Gender etc each column gives the information of the applicant and the last column Loan\_Status is the approval of the loan weather yes or not approved.

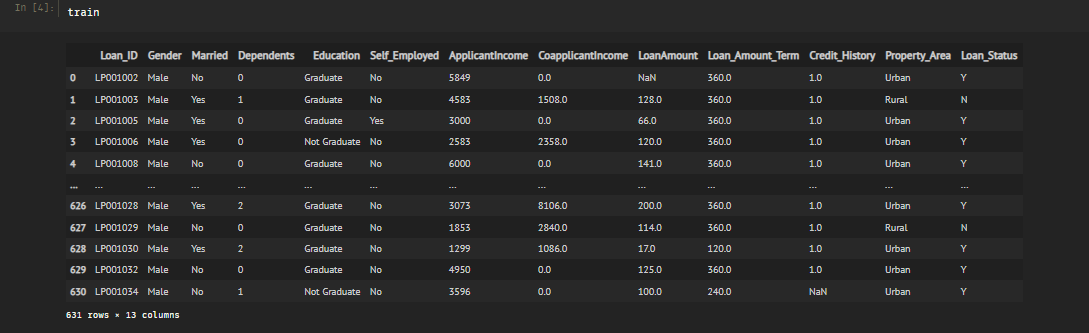


Figure Dataset information

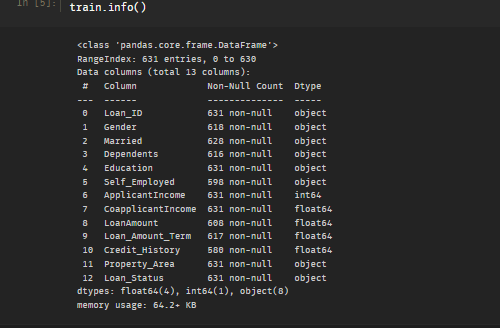


Figure Dataset information

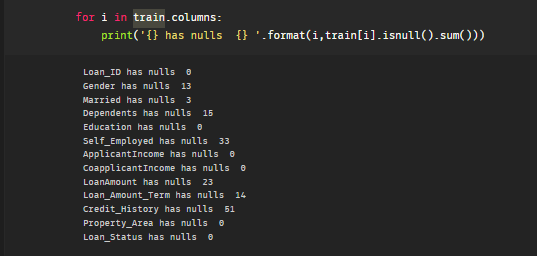


Figure Null Values in each columns

As per the above figure it is clearly be seen that the data is having different datatype columns and the non null values present in the dataset gives the insight that some of the columns are having nulls present in the dataset.

# Data pre processing

The Data as discussed above is having nulls value present in the dataset so to deal with it there are two catagories and for each category the null value treatement will be different as per this dataset the collected data are already limited so the removing of the null values will not be the best approach.

For treating the null values the best approach will be to split the dataset into categorical and numerical columns and treat their null values separately.

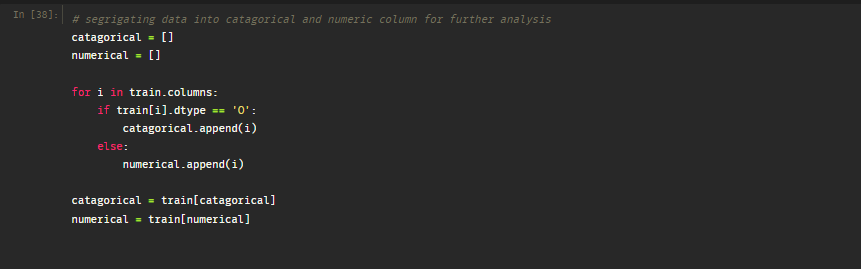


Figure Segrigating data into numeric and categorical

Now the dataset has been segeregated into the categorical and numerical columns the next thing will be to pre process the data and for the null value in the numerical column the best approach will be to fill them with mean as per each column and for the categorical column the null value will be filled with the model or max repeating value.

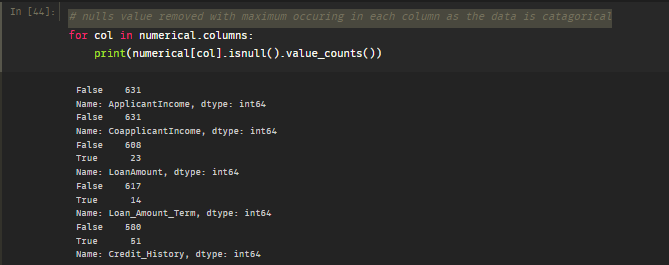


Figure Filling null value with mean as per each columns

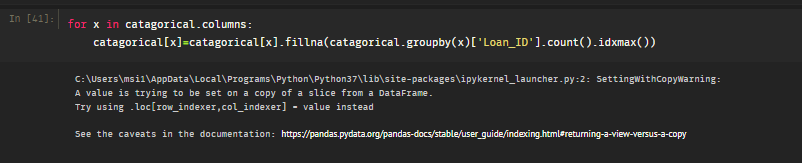


Figure filling cat columns with most occurring data i.e mode

# Data Visualization

As the dataset is finally had processed now its time to visualize the data and check the skewness and the co relation between each columns.

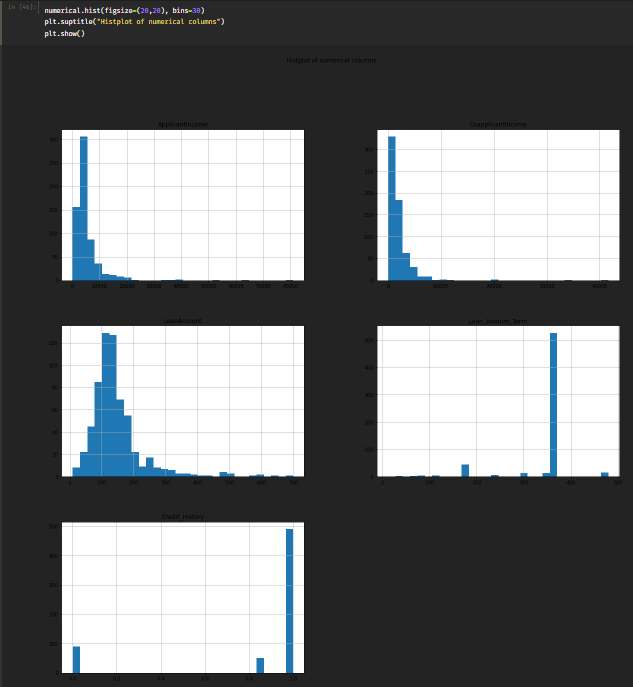


Figure Checking skewness of each columns

As per the above figure the dataset is not much skewed and is good for theto be applied to the machine learning algorithm and also it is segregated evenly.

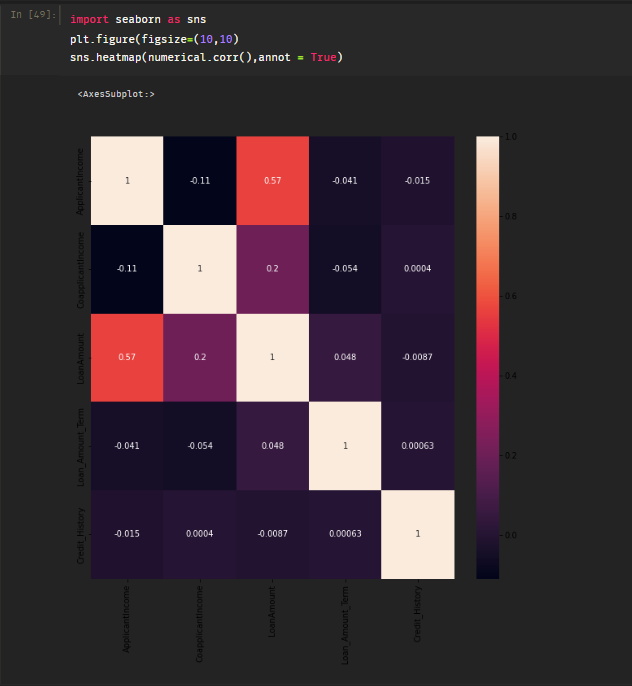


Figure Corelation plot

The above figure the co relation of each column with other columns and as here it is clearly be seen that the data is not having strong co relation with other columns except loan amount and applicant income and it is obvious that the applicant with the more income has more chances to take the big amount loan.

Now lets check the outliers for each numeric columns.

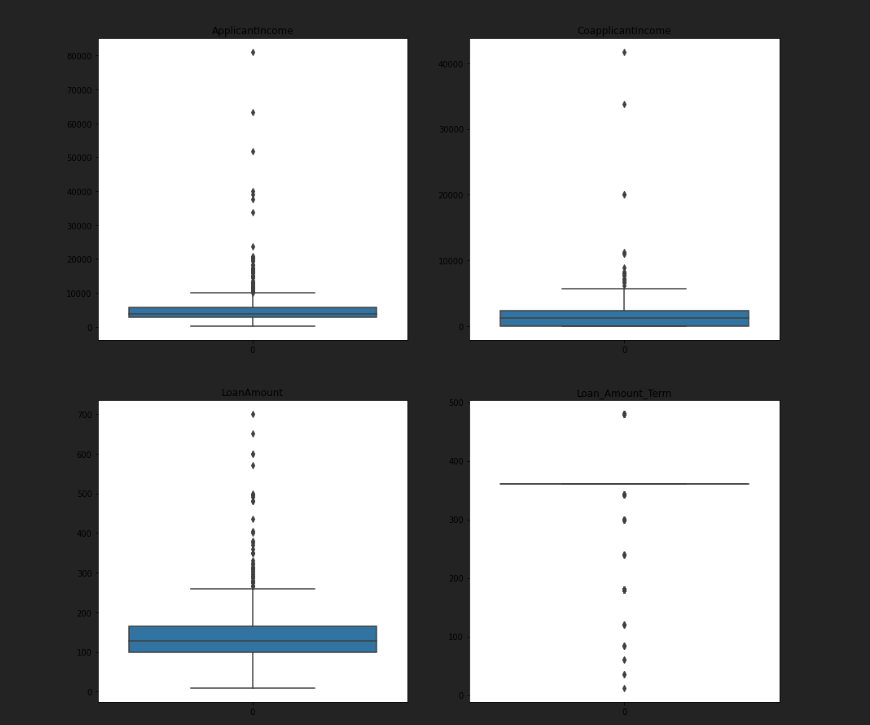


Figure Box plot for checking outliers

As per the above box plot the data is seems to have some outliers as seen the dots going too much away from the majority of the data rows but in this case the dataset can be have the outliers as the dataset is consist of different persons and loan amounts as per their applications so the outlier treatment here will not be the best approach here it may affect the model performance but if the data will be large these outliers will make more sense.

# Processing Data for Machine Learning

Noe the dataset has been visualized and initially processed it time to process the data so it will be ready for fitting the model.

The first theng will be need here is to convert the categorical column into the labeld data as the models can not take the strings or object type data into the model training.

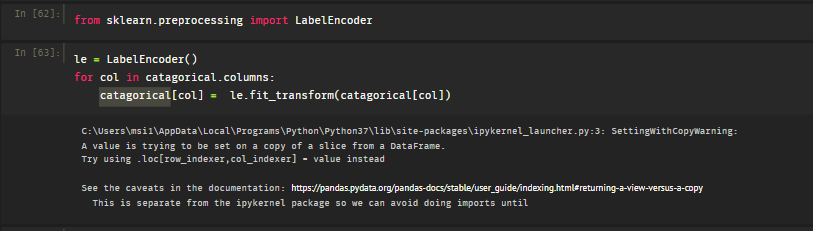


Figure Label Encodeing

the labels can be encoded either by one hot encoding or by using the pre built liberary and here in the above figure the pre-built liberary Label-Encoder has been used to convert the data into the labels.

As for the above analysis the data was segegrated into categorical and numeric columns now they needed to be merged again into the single dataset.

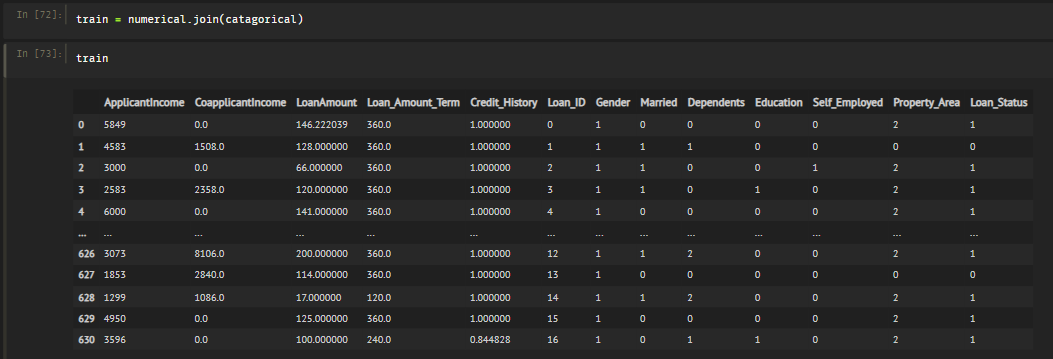


Figure Merging the data into single dataset

Now the next thing will be to again split the dataset into the training and testing dataset. The test size will be provided 20 percent of the whole dataset and 80 percent will be for training the model.

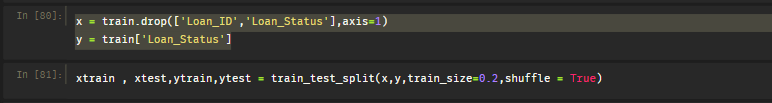


Figure Splitting the data into training and testing

# Applying the ML model

Now the dataset is ready to be applied for fitting the model and here the Decision tree classifier will be used

The decision tree classifier's key benefit is its capacity to use several feature subsets and decision rules at various stages of categorization. A generic decision tree has one root node, several internal and leaf nodes, and branches

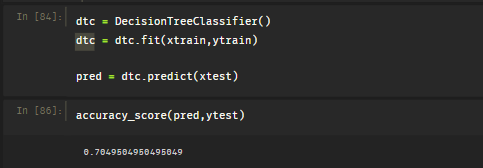


Figure Applying Decesion Tree

As per the above figure the decision tree which is applied initially is giving decent accuracy but it can be increased by using the hyper parameters and additional tuning to the dataset itself.

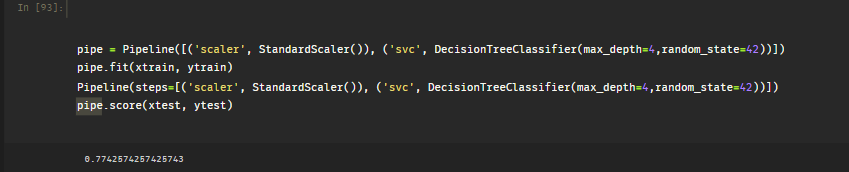


Figure Parameters tuning for achieving more accuracy

In the above figure the pipeline has been added so that there will be less leakage of the data while going through the model and standard scaler. Standard scaler used here so that the data will be scaled it usually increase the model run time and more accuracy to the predictions. In above figure after applying the pipeline and standard scaler and tuning the Decision tree with max depth 4 the accuracy is slightly increased with 7 percent approx.