**Loan Application Status Prediction**

**Problem Statement:-**

**There is dataset in which details of applicants who have applied for loan .It includes details like**

**Credit history**

**Loan amount**

**Their income**

**Dependents**

**In this dataset , there are independent as well as dependent variables.**

**Independent Variables:-**

**Loan ID**

**Gender**

**Married**

**Dependent**

**Education**

**Self employed**

**Applicant income**

**Co-applicant income**

**Loan amount**

**Loan amount term**

**Credit history**

**Property area**

**Dependent Variable:-**

**Loan status**

**We have to build a model that can predict whether the loan of the applicant will be approved or not on the basis of details provided in dataset.**

**Data Analysis:-**

**It is a process of applying logical techniques to describe and evaluate data.**

**We used data.info( ) in which 613 rows and 13 columns are present .**

**We also used data.head( ) to evaluate the data in vast form**

**EDA:-**

**EDA means Exploratory Data Analysis. It is an approach for data analysis that employs a variety of techniques that is mostly graphical.**

**Firstly import all the necessary libraries and store the data in our system for analysis.**

**import pandas as pd**

**import numpy as np**

**import matplotlib. pyplot as plt**

**import seaborn as sns**

**from sklearn.ensemble import Gradient Boosting Classifier**

**from sklearn.ensemble import Random Forest Classifier**

**from sklearn.model\_selection import Cross\_val\_score**

**from sklearn.tree import Decision Tree Classifier**

**from sklearn.neighbors import Kneighbors Classifier**

**from sklearn.import svm**

**If there are missing values in the dataset before doing any statistical analysis,we need to handle those missing values by using:-**

**data.isnull( ).sum( )**

**As we can see seven columns contain missing values.**

**Then we can print**

**percent of missing Gender record**

**percent of missing married record**

**percent of missing dependents record**

**percent of missing self employed record**

**percent of missing loan amount record**

**percent of missing credit history record.**

**Now,lets plot the count plot for 6 columns. We will use seaborn library for plotting count plot Columns like:-**

**Gender**

**Married**

**Dependents**

**Self employed**

**Loan amount**

**Loan amount term**

**Credit history**

**Then handle the missing values of 7 columns we firstly use data.copy( ) then fillna. So there is no missing value present in dataset.**

**Pre-processing pipeline:-**

**Pre-processing pipeline are a simple way to keep your data pre-processing and modeling code organized. We used encoding on gender\_stat,yes\_no\_stat,dependents\_stat,education\_stat,property\_stat.**

**Then the train data gender is replace by gender\_stat**

**married is replace by yes\_no\_stat**

**dependents is replace by dependents\_stat**

**education is replace by education\_stat**

**self employed is replce by yes\_no\_stat**

**property area replace by property\_stat**

**we can separate features and target by using x and y**

**Building machine learning model:-**

**The initial step in building a machine learning model to understand the need for it.**

**import Gradient Boosting Classifier**

**import Random Forest Classifier**

**import Decision Tree Classifier**

**import Knearest neighbor Classifier**

**Firstly we used**

**Gradient Boosting Classifier:-**

**It is a group of machine learning algorithms that combine many weak learning models together to create strong predictive model.**

**So the outcome accuracy is 78.01**

**Random Forest Classifier :-**

**It contains a number of decision tree on various subsets of the given data set and improve the predictive accuracy of the data set.**

**So the outcome accuracy is 74.75**

**Decision Tree Classifier:-**

**It is a type of supervised learning algorithm that can be used for both regression and classification problems. It can be represented by a tree structure. Like any other tree representation it has a root node, internal nodes, and leaf nodes.**

**So the outcome accuracy is 71.82**

**K-nearest neighbors:**

**It is a supervised learning algorithms and determines how many nearest neighbors you want your datapoint look at .**

**So the outcome accuracy is 61.4**

**Lastly we plot the result of classification based on performance.**

**So the best classification model is Gradient Boosting Classifier having 78.01 accuracy .**

**Concluding Remarks:-**

**In loan approval prediction dataset we predict how many loan applicant will approved or not . So firstly we import all the necessary libraries and analysis the data . then check the information of the data and missing values of data. To handle the missing values plotting the count plot for dataset . then apply encoding technique . Lastly , the classification model is build we used different different model to find out the best classification model whose accuracy is best . we find out that the gradient boosting classifier having 78.01is best model that we used in this loan application status prediction project.**