**Introduction:**

The first part of the project contains the classification section, whether a person is eligible for a loan i.e. would their application for loan be accepted or rejected.

**Preprocessing:**

Feature Selection and Engineering:

The features used for this project are:

1. Gender
2. Married
3. Dependents
4. Education
5. Self\_Employed
6. TotalIncome
7. Credit\_History
8. Property\_Area
9. Loan\_Status

No. 6 ‘TotalIncome’ eas engineered from two existing columns - ‘ApplicantIcome’ and ‘CoapplicantIncome’.

Label Encoding:

Most of the fields taken for training of the modeling (mentioned above), are classification type and LabelEncoder() from sklearn.preprocessing to convert them to numerical type(integer) to allow the model to process them.

**EDA:**

1. First we see the relation of the acceptance of loan application on the gender of the applicant.

What we see is that the acceptance rate for Males is around 70%, while it is about 65% for Females.

The dataset is not sizable enough to say if any discrimination exists, but it certainly points towards the existence of the possibility of discrimination.

1. Next we see if being a graduate brings a difference in an applicants chances of being accepted. We can see that graduates have about a 65% chance of being accepted while non-graduates have a 60% chance. It show an expected result that banks prefer applicants of higher education, but the fact there are almost 20% more undergraduates shows that, in terms of number rather than percentages, more undergraduates receive loans than graduates. This also shows that graduates are less likely to take loans, be it for financial necessity or trying to start a business.
2. Now we see the dependency of marital status on the probability of being accepted. We see that both status have around 60% acceptance rate but that unmarried applicants are more in number (about 60% more). This shows that unmarried people are more likely to be in a need of financial assistance or that they are more likely to try starting their own company.
3. In number (2) we saw that there were more undergraduate applicants for a loan, here we try to identify the reason for the same. We see that around 13% of graduates are self employed while around 15% of undergraduates are self-employed. Here the numbers don’t show much of a difference but due to the difference in the number of graduates and non-graduate, we see that of the self employed people 80% are undergraduates.
4. We finally see the relation of Income of the applicants to the loan status. What is interesting is that there seems to be no correlation between them. There are a lot more (about twice) accepted application than rejected ones. We see from the graph that most applicant are of relatively lower income and that most of the high income applicants were accepted. One outlier in this case is a single applicant, the one with the highest income ,is rejected. The reason for this could be a high number of dependents and/or property being of rural type.

**Project Summary:**

The project is rather simple. After the preprocessing, The data is split into training set and testing set including input(X) and output(Y) using sklearn’s inbuilt function: train\_test\_split().

A logistic regression model is made using another inbuilt feature of sklearn, which is then trained using the fitness function with the training data.

Y\_Pred is the predicted data obtained after running the test input (X\_Test) through the trained model.

The accuracy\_score() is used to determine the accuracy of the trained model.

Since the data-set is relatively small the accuracy comes between anywhere from 78% to 89%, with the median being 79%-83%. Since the dataset is quite small we see a large difference in the accuracy with each iteration as the change in test data becomes very volatile.