Loan prediction Project

- **TA : Dr . Verena Nashaat**
- **Team members**

Section no.	ID	Names
Section 16	**20201700449**	**عبدالرحمن محمد نجيب محي الدين**
Section 4	20201700102	أدهم مصطفى محسن محمد
Section 31	20201700909	نانيس عادل شحاته حسن
Section 7	20201700176	باسم عاطف السيد محمد
Section 4	20201700091	أحمد وائل صلاح الدين حسن
Section 34	202017001019	يوسف عبد الرؤوف أمين محمود
Section 6	20201701060	أندرو عصام جورجي عطا

The project process

First:

We started by importing the libraries as (pandas and sklearn)
then the dataset loading and displaying the head of data

Head

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y

then the description of it (count , std and mean)

Shape : (614, 13)

Describe	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878

Then we counted null values

NULLS	Loan_ID	0
Gender		13
Married		3
Dependents		15
Education		0
Self_Employed		32
ApplicantIncome		0
CoapplicantIncome		0
LoanAmount		22
Loan_Amount_Term		14
Credit_History		50
Property_Area		0
Loan_Status		0

Then we removed the null values from the dataset.

Second:

We made label encoding to the data by changing the data which has string values by giving it an integer value then We did the data scaling and displayed the new data description after the scaling.

Data after preprocess

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
count	480.000000	480.000000	480.000000	480.000000	480.000000	480.000000	480.000000	480.000000	480.000000	480.000000	480.000000	480.000000
mean	239.500000	0.820833	0.647917	0.777083	0.202083	0.137500	5364.231250	1581.093583	144.735417	342.050000	0.854167	1.022917
std	138.708327	0.383892	0.478118	1.020815	0.401973	0.344734	5668.251251	2617.692267	80.508164	65.212401	0.353307	0.776411

Then we split data to x and y where x means the data which the prediction depends on it so it contained the data set except Loan id and Loan status and y means the predicted data which is loan status

Then we split data to 80% for train and 20% for test

So data become as the following

```
x shape , x tarin , y test  (480, 11) (384, 11) (96, 11)
y shape , y tarin , y test  (480,) (384,) (96,)
```

Third:

We made algorithms to calculate the accuracy of data like and print the accuracy before and after scaling:

support-vector machine (SVM).

SVM algorithm accuracy is equals to:

before scaling= 0.5833333333333334.

after scaling = 0.7291666666666666.

“Decision Tree Classifier “we made the decision tree to classify the data and calculate the accuracy of these data which is

before scaling=0.6666666666666666.

after scaling=0.7604166666666666.

“Logistic Regression” we made logistic regression to return the probability value by sigmoid function which gave us an accuracy equals to:

before scaling=0.6666666666666666.

after scaling=0.7604166666666666.

Fourth:

extracted a new feature principal component analysis (PCA).