

Master Big Data and Cloud Computing 2021-2023



# Loan Eligibility Prediction - Machine Learning

Realized by : Doha SASSA & Abir RAKRAK

Professeur de module : Said TKATEK

Chef de filière : Jaafar ABOUCHABAKA

Remerciement

Premièrement nous remercions Dieu source de toutes connaissances

Au terme de ce travail, nous tenons à remercier notre encadrant Said Tkatek, professeur du module Administration du base de données oracle. Et les cadres de département d’informatique, faculté des sciences Kénitra.

Nous voudrons également exprimer notre gratitude envers nos collègues.

Finalement, nous remercions nos parents et nos amis, pour leur soutien et encouragements.

Souhaitons que notre projet serait à la hauteur de vos attentes.



Table of contents:

I-Tools installation:………………………………….4

I-1) Oracle database………...................................4

I-2) Visual studio code…………………….........13

I-3) Python…………………………………..13-14

II-Introduction: …………………………………….15

II-1) What is Machine learning?..........................15

II-2) Why is Machine learning important?.........15

II-3) What are the different types of machine

learning? ………………………………15-16

II-4) What are the advantages and disadvantages

of machine learning?....................................17

II-5) How to choose the right machine learning

model……………………………………..18

III-Case study : Loan eligibility prediction………...19

III-1) Objectives of this case……………………19

III-2) Datasets………………………………..19-23

III-3) Plotting and analyzing data……………24-26

III-4) Using machine learning algorithms…...27-37

VI-Summary……………………………………….38

I-Tools installation

I-1) Oracle database

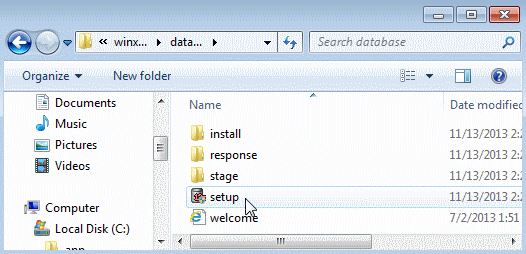
-Download OracleXE 11g express edition:

-Download Oracle 12c:

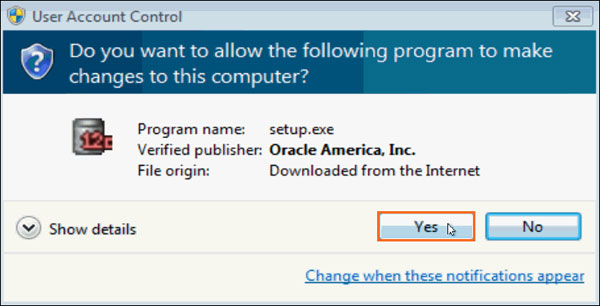
<https://drive.google.com/drive/folders/1Ix39NHU0rtLtxYs5_0-zSUXKqCvdkWe3?usp=sharing>

-Steps of oracle 12c installation:

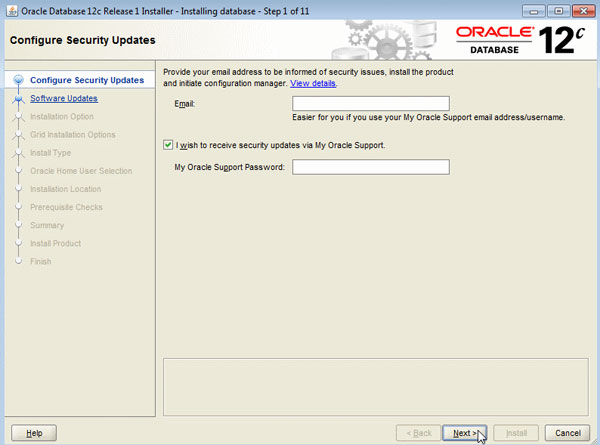
1. Expand the**database** folder that you extracted in the previous section. Double-click **setup**.

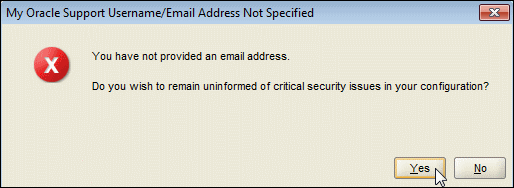


2.Click **Yes** in the User Account Control window to continue with the installation.



3. The **Configure Security Updates** window appears. Enter your email address and My Oracle Support password to receive security issue notifications via email. If you do not wish to receive notifications via email, deselect "**I wish to receive security updates via My Oracle Support**". Click **Next** to continue. Click "**Yes**" in the confirmation window to confirm your preference.

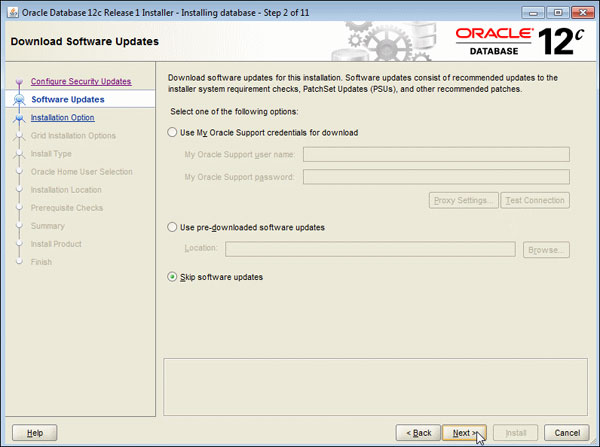




4. The **Download Software Updates** window appears with the following options:

* **Select "Use My Oracle Support credentials for download" to download and apply the latest software updates.**
* **Select "Use pre-downloaded software updates" to apply software updates that you previously downloaded.**
* **Select "Skip software updates" if do not want to apply any updates.**

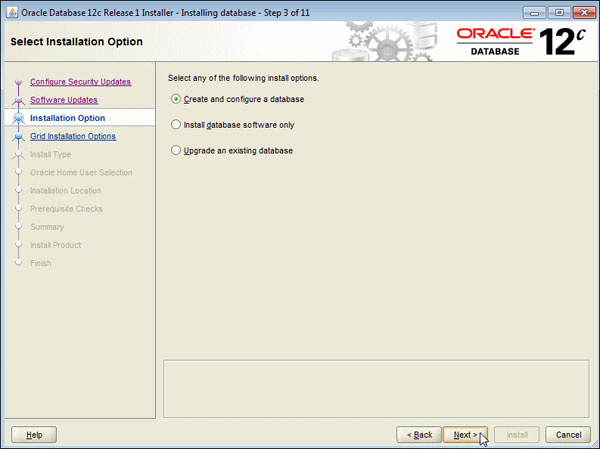
Accept the default and click **Next**.



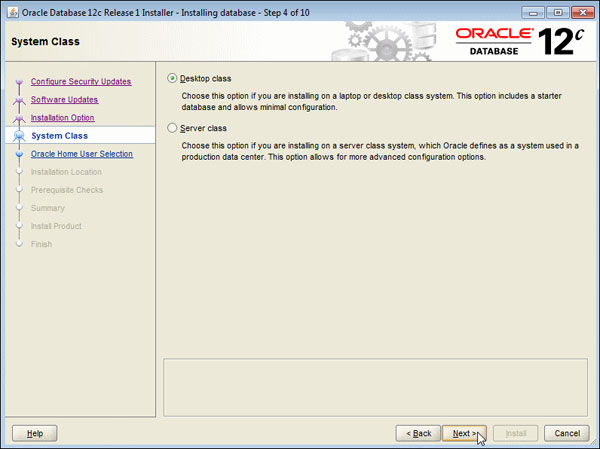
5. The **Select Installation Option** window appears with the following options:

* **Select "Create and configure a database" to install the database, create database instance and configure the database.**
* **Select "Install database software only" to only install the database software.**
* **Select "Upgrade an existing database" to upgrade the database that is already installed.**

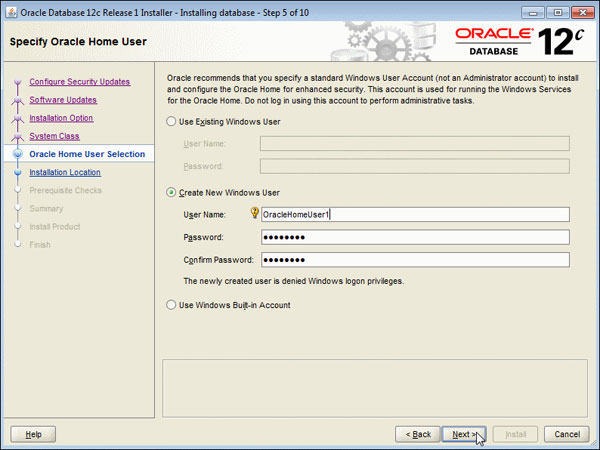
In this OBE, we create and configure the database. Select the **Create and configure a database** option and click **Next**.



6. The **System Class** window appears. Select Desktop Class or Server Class depending on the type of system you are using. In this OBE, we will perform the installation on a desktop/laptop. Select **Desktop class** and click **Next**.



7. Select the **Create New Windows User** option. Enter the user name as **OracleHomeUser1** and password  as **Welcome1**. Click **Next**.



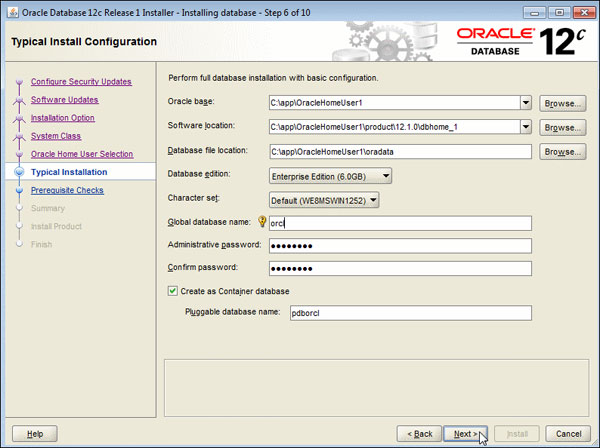
8. The **Typical Install Configuration** window appears. Click on a text field and then the balloon icon ( https://www.oracle.com/webfolder/technetwork/tutorials/obe/db/12c/r1/Windows_DB_Install_OBE/images/more.gif ) to

know more about the field. Note that by default, the installer creates a container database along with a pluggable

database called "pdborcl". The pluggable database contains the sample HR schema. Change the Global

database name to **orcl**. Enter the "Administrative password" as **Oracle\_1**. This password will be used later to

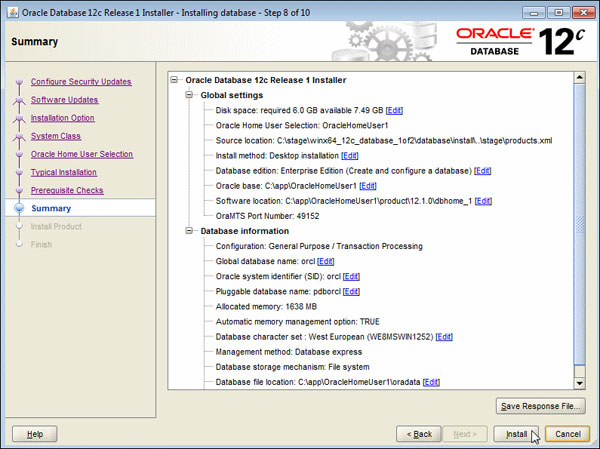
log into administrator accounts such as SYS and SYSTEM. Click **Next**.



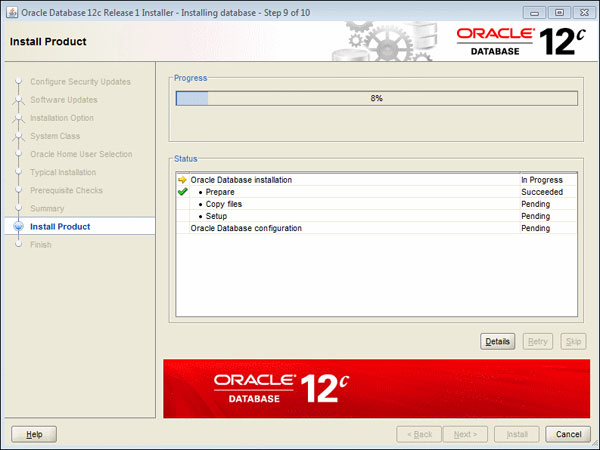
9. The prerequisite checks are performed and a **Summary**window appears. Review the settings and click **Install**.

**Note:**Depending on your firewall settings, you may need to grant permissions to allow java to access the

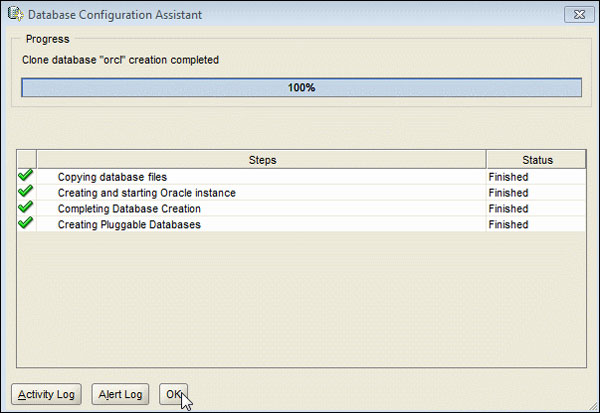
network.



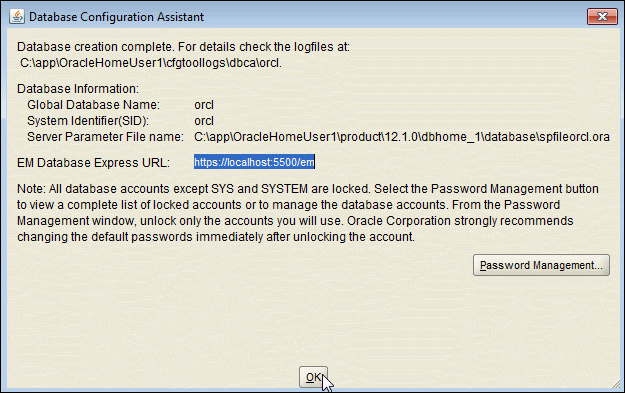
10. The progress window appears.



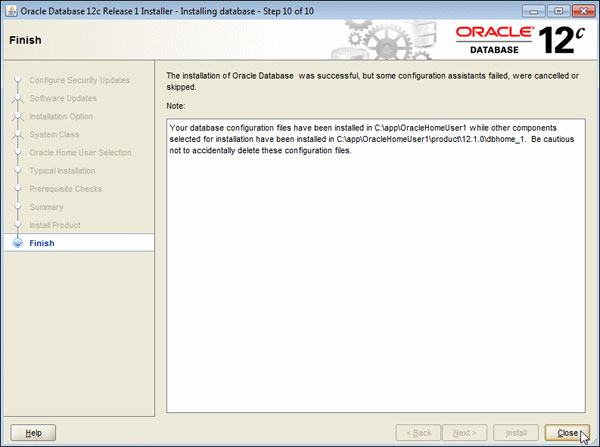
11. The Database Configuration Assistant creates the database.



12. After the Database Configuration Assistant creates the database, you can navigate to <https://localhost:5500/em> as a SYS user to manage the database using Enterprise Manager Database Express. You can click "Password Management..." to unlock accounts. Click **OK** to continue.



13. The **Finish**window appears. Click **Close** to exit the Oracle Universal Installer.



I-2)Visual studio code:



Code editor: Visual Studio Code: extensible code editor developed by Microsoft for Windows, Linux and MacOs. Supports languages: C, C++, Go, CSS, JavaScript, Type Script, C# and PHP etc…

Download link: <https://code.visualstudio.com/download>

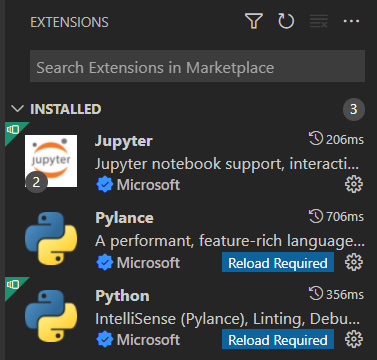
Language used: Python version 3.9.5

I-3) Python

Download python latest version on machine:

[**Download Python | Python.org**](https://www.python.org/downloads/)

Dowload python and Jupiter notebook in vscode extensions.



Python libraries:

**MATPLOTLIB :**

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

Installation on command prompt : pip install matplotlib.

**NUMPY:**

NumPy offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more.

Installation on command prompt : pip install numpy.

**SKLEARN:**

Sklearn or sickit learn is a free software machine learning library for the Python programming language.It features various classification, regression and clustering algorithms.

Installation on command prompt : pip install sklearn.

**SQLALCHEMY:**

SQLAlchemy is an open-source SQL toolkit and object-relational mapper (ORM) for the Python programming language released under the MIT License.

Installation on command prompt : pip install sqlalchemy.

**STATISTICS:**

Python has a built-in library called statistics. This module provides functions for calculating mathematical statistics of numeric (Real-valued) data.

Installation on command prompt : pip install statistics.

**SEABORN:**

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Installation on command prompt : pip install seaborn.

II-Introduction

II-1) what is Machine learning?

Machine learning is a branch of [artificial intelligence (AI)](https://www.ibm.com/cloud/learn/what-is-artificial-intelligence) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

II-2) why is Machine learning important?

Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies.

II-2) what are the different types of Machine learning?

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches:[supervised](https://www.techtarget.com/searchenterpriseai/definition/supervised-learning) learning, [unsupervised](https://www.techtarget.com/whatis/definition/unsupervised-learning) learning, semi-supervised learning and reinforcement learning. The type of algorithm data scientists choose to use depends on what type of data they want to predict.

**Supervised learning:** In this type of machine learning, [data scientists](https://www.techtarget.com/searchenterpriseai/definition/data-scientist) supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified. Supervised learning algorithms are good for the following tasks:

* **Binary classification:**Dividing data into two categories.
* **Multi-class classification:**Choosing between more than two types of answers.
* **Regression modeling:** Predicting continuous values.
* **Ensembling:** Combining the predictions of multiple machine learning models to produce an accurate prediction.

**Unsupervised learning:** This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined. Unsupervised learning algorithms are good for the following tasks :

* **Clustering:** Splitting the dataset into groups based on similarity.
* **Anomaly detection:** Identifying unusual data points in a data set.
* **Association mining:** Identifying sets of items in a data set that frequently occur together.
* **Dimensionality reduction:**Reducing the number of variables in a data set.

**Semi-supervised learning:** This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled [training data](https://www.techtarget.com/searchenterpriseai/feature/Using-small-data-sets-for-machine-learning-models-sees-growth), but the model is free to explore the data on its own and develop its own understanding of the data set. Semi-supervised learning strikes a middle ground between the performance of supervised learning and the efficiency of unsupervised learning. Some areas where semi-supervised learning is used include:

* **Machine translation:** Teaching algorithms to translate language based on less than a full dictionary of words.
* **Fraud detection:** Identifying cases of fraud when you only have a few positive examples.
* **Labelling data:** Algorithms trained on small data sets can learn to [apply data labels](https://www.techtarget.com/whatis/definition/data-labeling) to larger sets automatically.

**Reinforcement learning:**Data scientists typically use [reinforcement learning](https://www.techtarget.com/searchenterpriseai/definition/reinforcement-learning) to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way. Reinforcement learning is often used in areas such as:

* **Robotics:** Robots can learn to perform tasks the physical world using this technique.
* **Video gameplay:** Reinforcement learning has been used to teach bots to play a number of video games.
* **Resource management:** Given finite resources and a defined goal, reinforcement learning can help enterprises plan out how to allocate resources.

II-2) what are the advantages and disadvantages of Machine learning?

Machine learning has seen use cases ranging from predicting customer behavior to forming the operating system for self-driving cars.

When it comes to advantages, machine learning can help enterprises understand their customers at a deeper level. By collecting customer data and correlating it with behaviors over time, machine learning algorithms can learn associations and help teams tailor product development and marketing initiatives to customer demand.

Some companies use machine learning as a primary driver in their business models. Uber, for example, uses algorithms to match drivers with riders. Google uses machine learning to surface the ride advertisements in searches.

But machine learning comes with disadvantages. First and foremost, it can be expensive. Machine learning projects are typically driven by data scientists, who command high salaries. These projects also require software infrastructure that can be expensive.

There is also the problem of machine learning bias. Algorithms trained on data sets that exclude certain populations or contain errors can lead to inaccurate models of the world that, at best, fail and, at worst, are discriminatory. When an enterprise bases core business processes on biased models it can run into regulatory and reputational harm.

II-2) how to choose the right machine learning model

The process of choosing the right machine learning model to solve a problem can be time consuming if not approached strategically.

**Step 1:** Align the problem with potential data inputs that should be considered for the solution. This step requires help from data scientists and experts who have a deep understanding of the problem.

**Step 2:** Collect data, format it and label the data if necessary. This step is typically led by data scientists, with help from data wranglers.

**Step 3:** Chose which algorithm(s) to use and test to see how well they perform. This step is usually carried out by data scientists.

**Step 4:** Continue to fine tune outputs until they reach an acceptable level of accuracy. This step is usually carried out by data scientists with feedback from experts who have a deep understanding of the problem.

III-Case study: Loan Eligibility

III-1) Objectives of this case

* Understanding how to use oracle to load a dataset and connect it to python.
* Analyzing and plotting data to understand the relationship between loan status and other independent variables.
* Building different machine learning models using python

Predicting loan status using different machine learning algorithms.

III-2) Datasets:

Choosing dataset:

Preparing data is a crucial step in machine learning, The larger your dataset, the harder it gets to make the right use .The dataset should be rich enough to let you play with it, and see some common phenomena.it should also have a reasonable mix of both continuous and categorical variables.

Data cleaning:

data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within dataset.

When combining multiple data source there are many opportunities for data to be duplicated or mislabeled.

steps of Data Cleaning :

==> Step1 : remove duplicate observations /remove irrelevant(sans importance) observations

==> Step2 : Fix structural errors (at the syntax level)

==> step3 : Filter unwanted outliers(unimportant data): If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

==> step4 : Handle missing data

==> step5 : Validate and QA

at the end you should be able to answer these questions as a part of basic validation like :

--> Does te data make sense ?

--> Does the data follow the appropriate rules for its field?

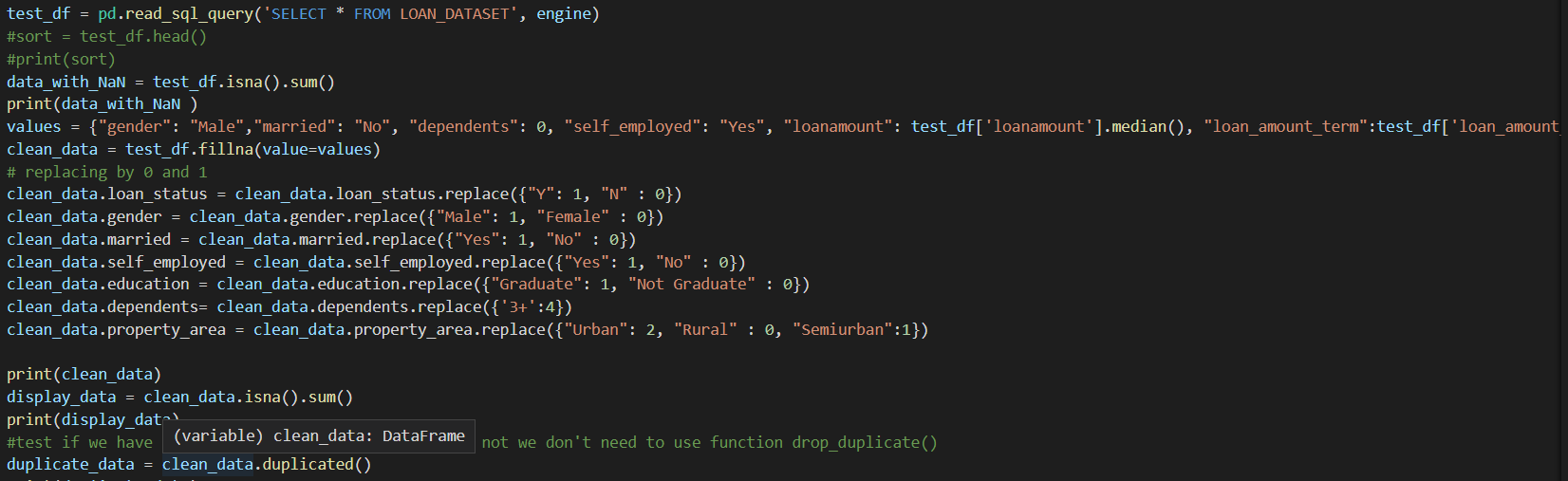
--> Does it prove or disprove your working theory, or bring any insight to light?

--> Can you find trends in the data to help you form your next theory?

--> If not, is that because of a data quality issue?

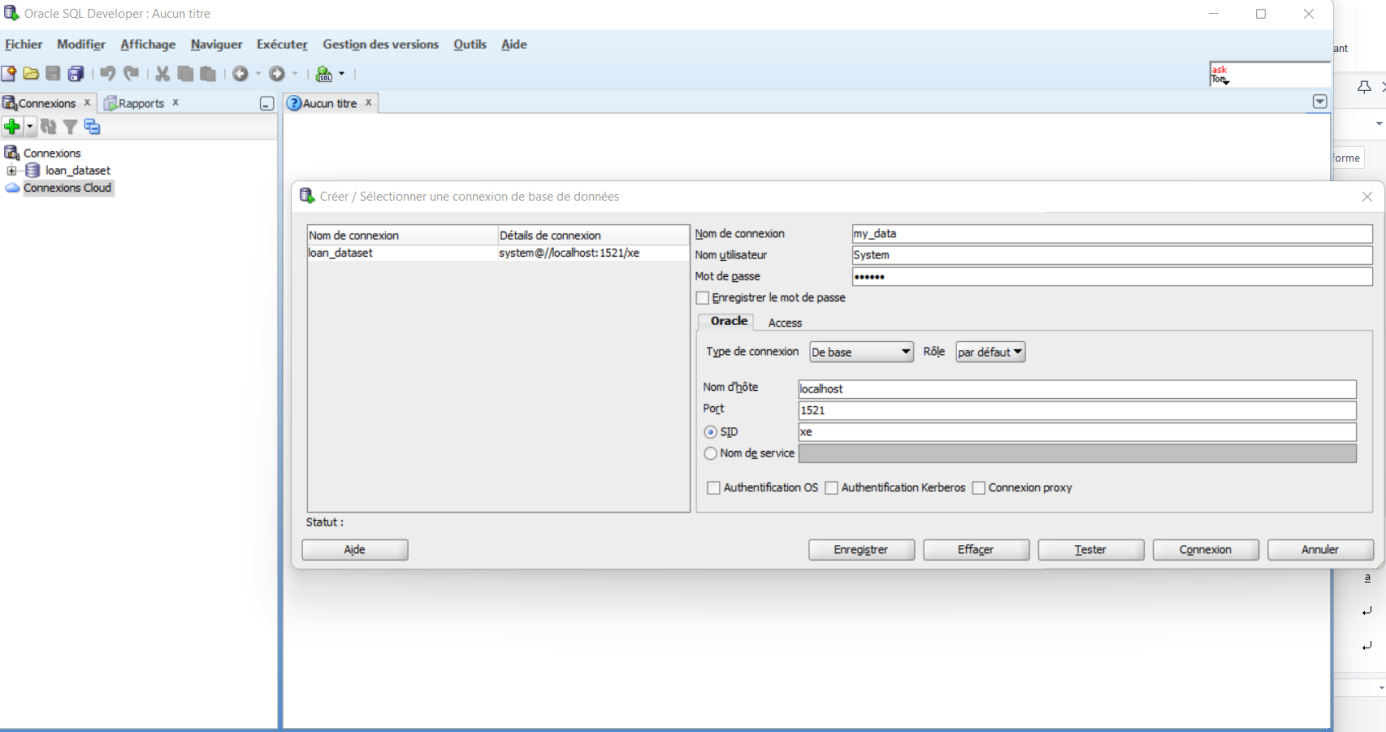
--- cleaning data in python:

1. Use function isna().sum to count NaN values in Dataframe.
2. Replacing continuous NaN values by median() : DataFrame[‘column name’].median(), but we replace categorical NaN values by 0 or 1.

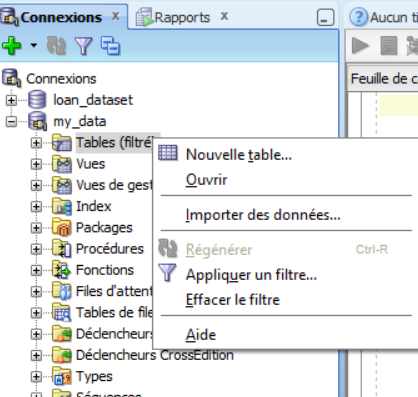


Importing data into SQL developer :

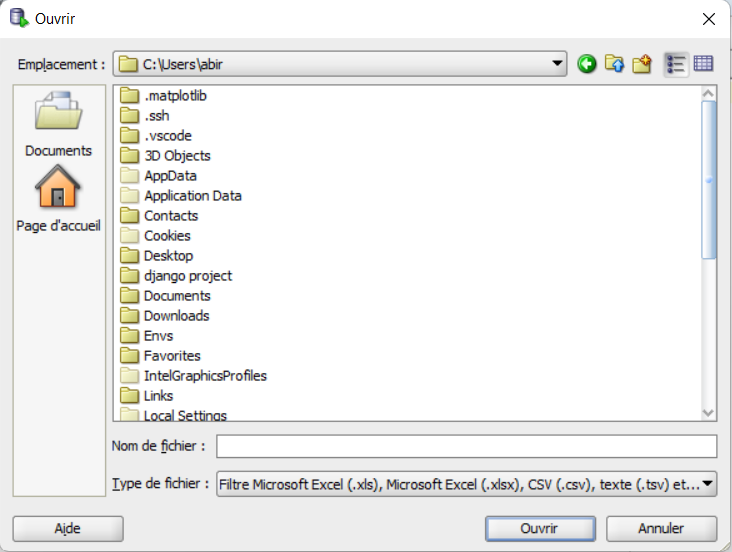
1-First open your SQL developer and create a new connection and fill in the fields with the right informations:



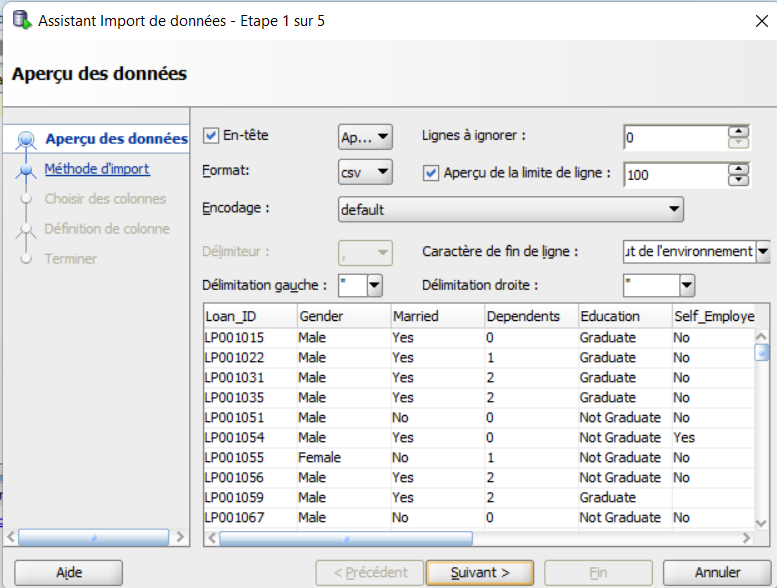
1. Your database connexion is created ,now you need to import the data ,right-click on tables and choose import data:



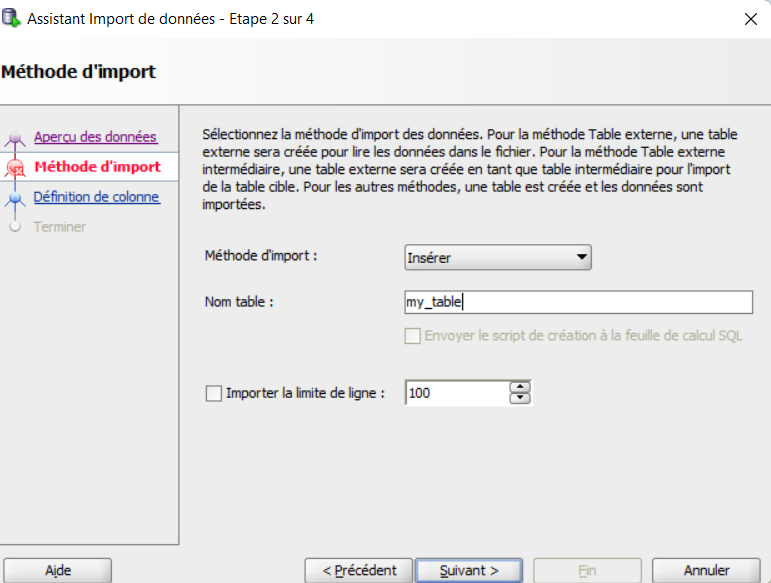
3-select your data-file to upload data (we used a csv file in this project):



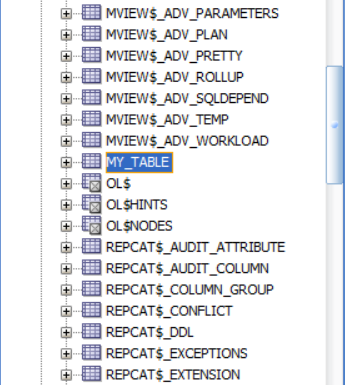
4-Your data will be uploaded like the following image, you can modify columns types and delimitations:



5-Name your Table:



6-The table is now uploaded successfully and you can manipulate your data as you want:

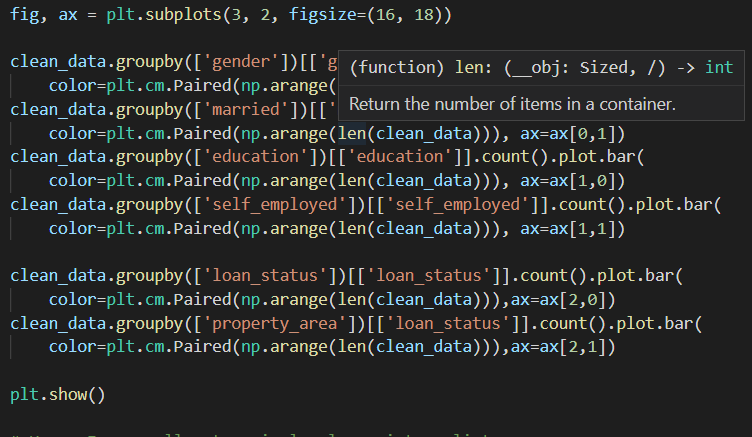
6-

III-3) Plotting and analyzing data:

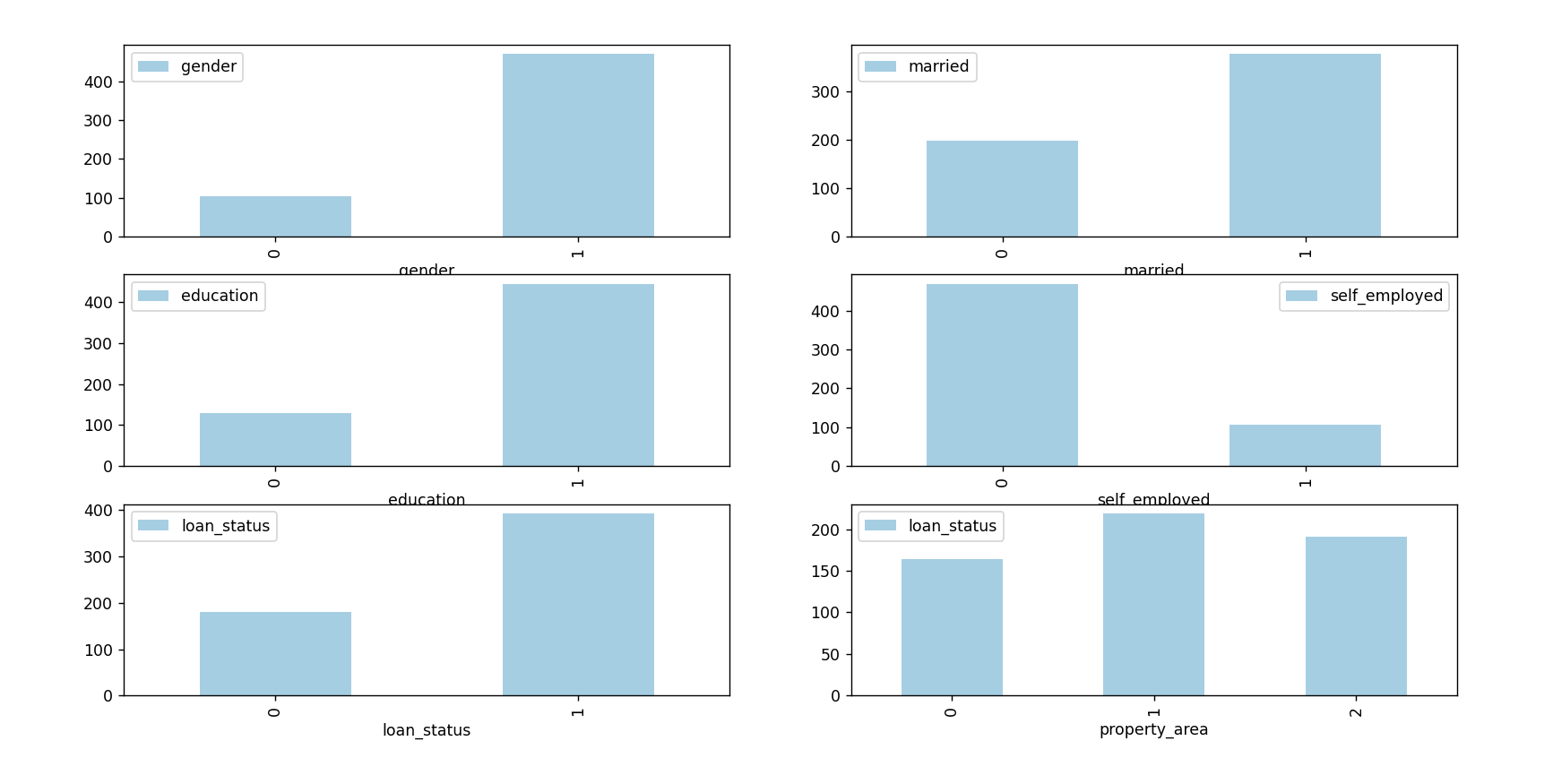
Plotting data is an important step of any machine learning project, the ability to visualize data leads to a better understanding of the case you’re dealing with .

To plot data using python we have many libraries such as matplotlib ,seaborn and pandas

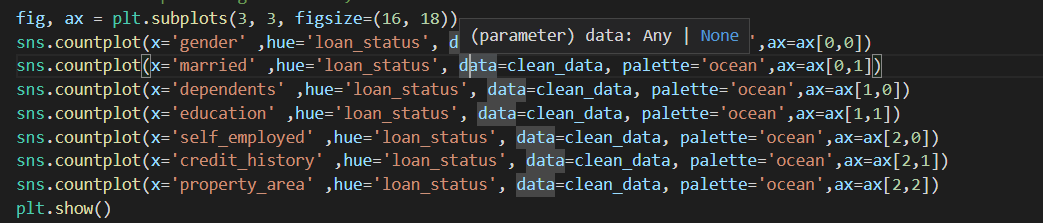
Countplot:



Output:

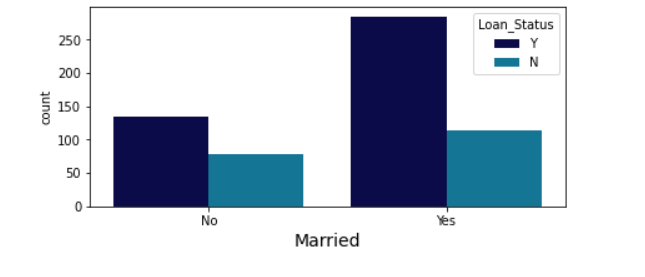
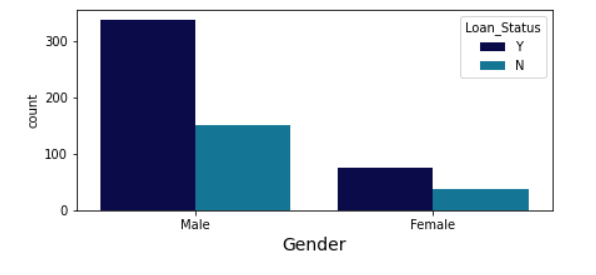
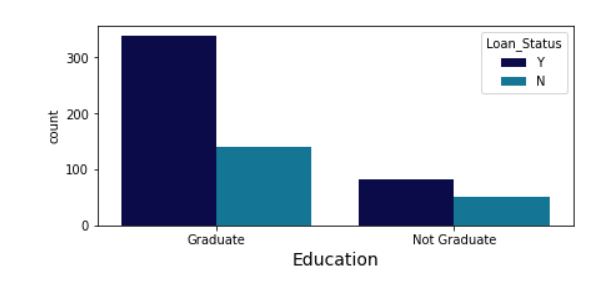
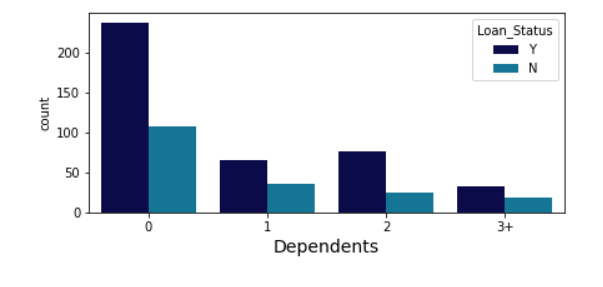


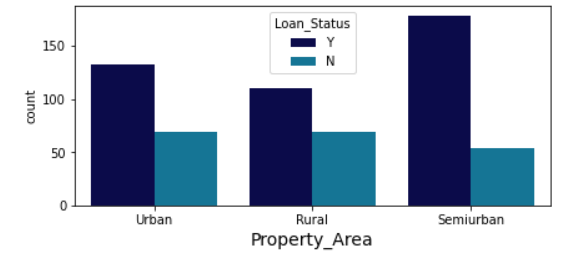
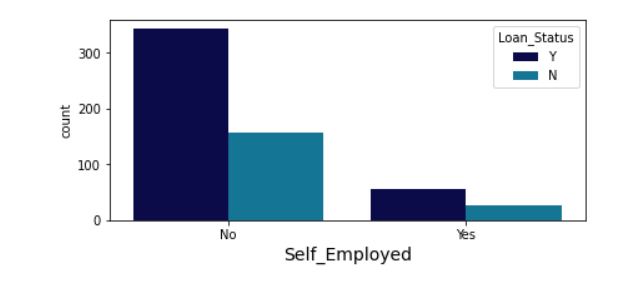
A countplot is **kind of like a histogram or a bar graph for some categorical area**. It simply shows the number of occurrences of an item based on a certain type of category.



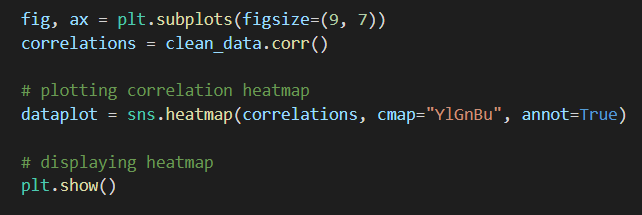
Output:

The following count plots show the relationship between loan status(target variable) and other variables:

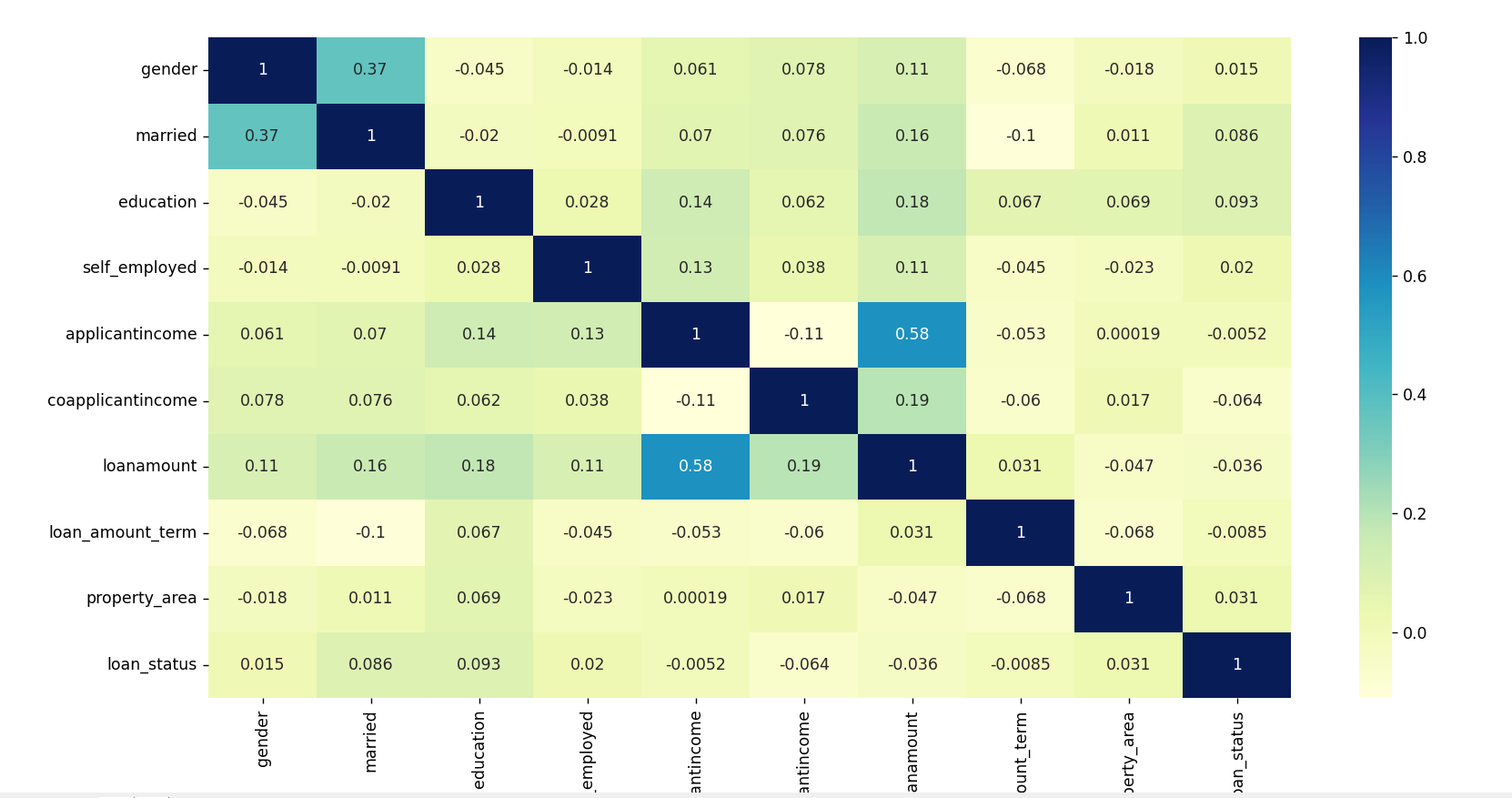
 



Heatmap: A heatmap is **a graphical representation of data where each value of a matrix is represented as a color**.



Output:



III-4) Using machine learning algorithm

1. Classification:

Before diving into any of the ml algorithms in this project we need to explain a major concept called classification , so what is classification ?

 Classification is an area of supervised machine learning that tries to predict which class or category some entity belongs to, based on its features.

For example, you might analyze the employees of some company and try to establish a dependence on the features or variables, such as the level of education, number of years in a current position, age, salary, odds for being promoted, and so on. The set of data related to a single employee is one observation. The features or [variables](https://realpython.com/python-variables/) can take one of two forms:

Independent variables, also called inputs or predictors, don’t depend on other features of interest (or at least you assume so for the purpose of the analysis).

Dependent variables, also called outputs or responses, depend on the independent variables.

There are two main types of classification problems:

Binary or binomial classification: exactly two classes to choose between (usually 0 and 1, true and false, or positive and negative)

Multiclass or multinomial classification: three or more classes of the outputs to choose from.

**In this project we will focus on Binary classification since we want to predict the loan status ,a binary dependent variable (0 for loan not accepted and 1 for accepted) using independent variables (gender,married,loan amount etc...).**

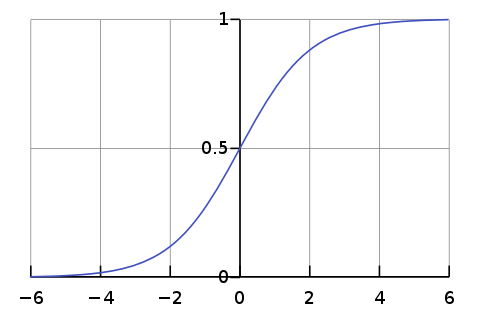
2-Logistic regression:

Logistic regression: is a fundamental classification technique. It belongs to the group of [linear classifiers](https://en.wikipedia.org/wiki/Linear_classifier) and is somewhat similar to polynomial and [linear regression](https://realpython.com/linear-regression-in-python/). Logistic regression is fast and relatively uncomplicated, and it’s convenient for you to interpret the results. Although it’s essentially a method for binary classification, it can also be applied to multiclass problems.

Sigmoid function:

Sigmoid function plays the role of an activation function in logistic regression.

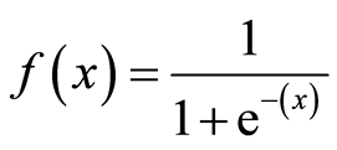
It takes the weighted sum of the input features and outputs the probability value of the outcome



The sigmoid function is presented by a curve .

For any value of x the value of 1 will remain between 0 and 1.

sigmoid formula :



When you’re implementing the logistic regression of some dependent variable y on the set of independent variables x=(x1,x2....xn) where n is the number of predictors(or inputs), you start with the known values of the predictors xi,and the corresponding actual response (or output) yi for each observation i=1....n

Your goal is to find the **logistic regression function** p(x) such that the **predicted responses p(xi)**

are as close as possible to the **actual response yi** for each observation i=1,2,3..n

Logistic regression is a linear classifier , so we’ll be using a linear function f(x)=b0+b1x1+b2x2.... also called the logit.the variable b0 is called the **intercept**,b1...bn are called the estimators of the r**egression coefficients** or predicted weights.

The logistic regression function p(x) is the sigmoid function of f(x):p(x)=1/(1+exp(-f(x)).

As such it’s often close to either 0 or 1.the fucntion p(x) is interpreted as the predicted probability that the output for a given x is equal to 1.Therefore ,1-p(x) is the probability that the output is 0.

Creating a logistic regression model in python:

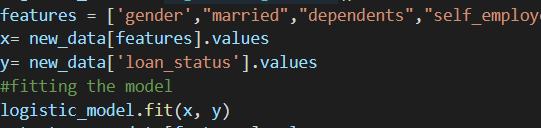
1. Creating model

:creating model

-Fitting training data :

Fitting your model to the training data is essentially the training part of the modeling process. It finds the coefficients for the equation specified via the algorithm being used

X are the independent variables or features used to predict loan status(gender,married,loanamount,selfemployed....)



Calculating intercept b0:

intercept

Output:

intercept output

Calculating coefficients b1,b2.....:

model coefficients

Output:

coeff output

-Predicting values of loan status:

predictions of loan status

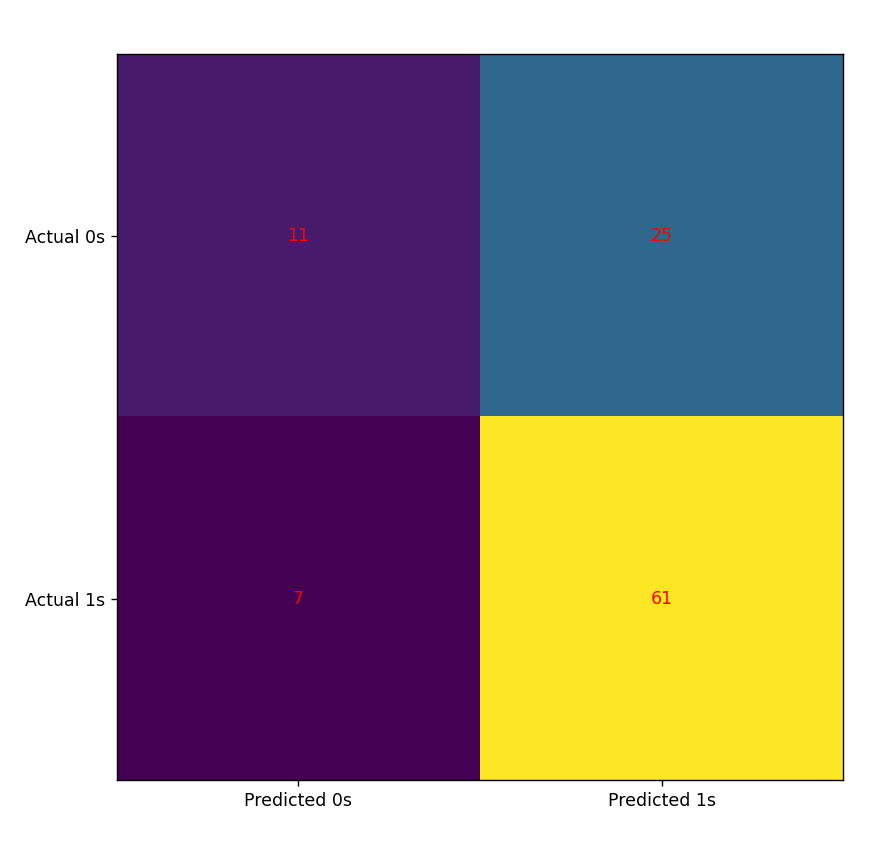
Output:

predictions output

-confusion matrix to evaluate the accuracy of a classification.:

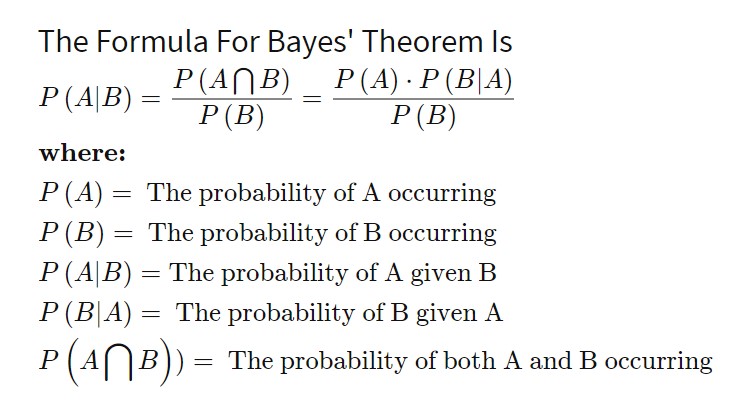
confusion matrix

Output:



1. naive bayes:

 naive Bayes classifiers are a family of simple "[probabilistic classifiers](https://en.wikipedia.org/wiki/Probabilistic_classification)" based on applying [Bayes' theorem](https://en.wikipedia.org/wiki/Bayes'_theorem) with strong (naive) [independence](https://en.wikipedia.org/wiki/Statistical_independence) assumptions between the features



There are two types of naive bayes:

Gaussian naive bayes:

Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution.

Bernoulli naive bayes:

Bernoulli Naive Bayes is a part of the family of Naive Bayes. It only takes binary values.

The decision rule for Bernoulli naive Bayes is based on:

bernoulli

According to the decision rule formula,x needs to be binary. Think about the formula in the case where xi=1 and the case where xi=0. So i is the event where xi=1 or the event where xi=0.

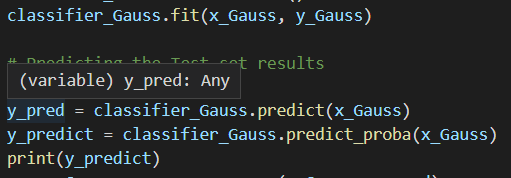
1-Creating model for Gaussian naïve bayes:



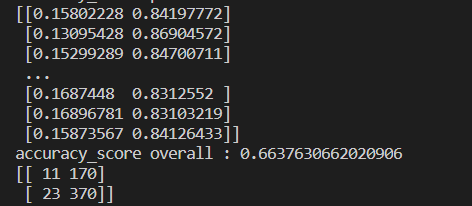
-Fitting training data:

Fitting your model to the training data is essentially the training part of the modeling process.

X are the continuous variables or features used to predict loan status( applicantincome, coapplicantincome, loanamount, loan\_amount\_term....)



Output:



2-creating model for Bernoulli naïve bayes:



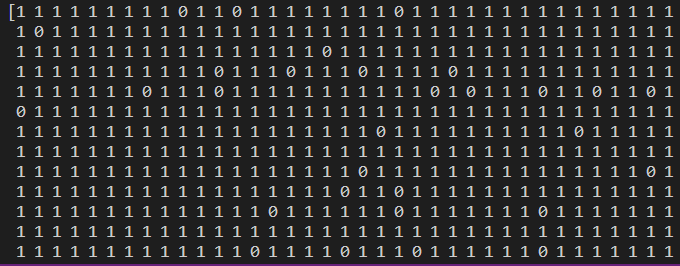
-Fitting training data:

Fitting your model to the training data is essentially the training part of the modeling process.

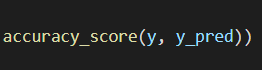
X are the continuous variables or features used to predict loan status( gender, married , education, self\_employed....)



Output:



Accuracy score:



Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

Accuracy=Number of correct predictions / Total number of predictions

Accuracy of the model:

The closer the accuracy to 1 the better the model.

Accuracy of loan prediction using Bernoulli:

accuracy output

3-Decision tree

* Decision Tree is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
* In a Decision tree, there are two nodes, which are the **Decision Node** and**Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed on the basis of features of the given dataset.
* ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***
* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
* In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**
* A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into sub-trees.

The diagram bellow explains the general structure of a decision tree: 

## **Why use Decision Trees?**

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

* Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
* The logic behind the decision tree can be easily understood because it shows a tree-like structure.

## **Decision Tree Terminologies**

 **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

 **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

 **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.

 **Branch/Sub Tree:** A tree formed by splitting the tree.

 **Pruning:** Pruning is the process of removing the unwanted branches from the tree.

 **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

## **Steps of decision tree algorithm**

* **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
* **Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**
* **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
* **Step-4:** Generate the decision tree node, which contains the best attribute.
* **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

1-creating model for decision tree:

decisiontree model

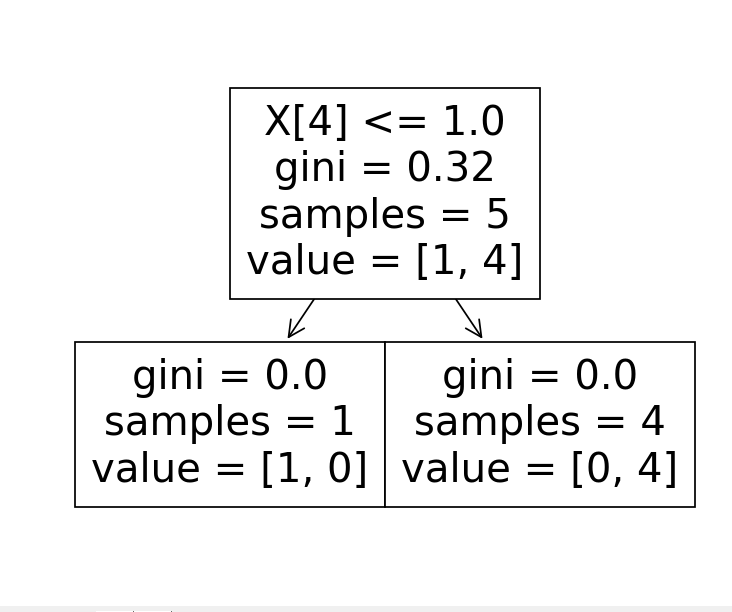
Fitting training data:

fitting decision tree

Plotting the tree:

plot dt

Output:



Gini Index: It is calculated by subtracting the sum of squared probabilities of each class from one.

VI)Summary:

* accuracy\_score of logistic regression model is: 0.764808362369338 wich means it is a good score.
* accuracy\_score of gaussian model is : 0.6637630662020906 wich is a medium score.
* bernoulli's accuracy score : 0.764808362369338
* Accuracy of decision tree: 0.7926829268292683

For our case logistic regression and decision tree had the most accurate results , with an accuracy score of 79%. Witch means those two algorithms are the best for this case.