Chapter 6 IMPLEMENTATION



**6.1.2 Implementation support**

**6.1.2.1 Anaconda**

Anaconda is a [free and open source](https://en.wikipedia.org/wiki/Free_and_open_source) distribution of the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) and [R](https://en.wikipedia.org/wiki/R_(programming_language)) programming languages for [data science](https://en.wikipedia.org/wiki/Data_science) and learning related applications (large-scale data processing, [predictive](https://en.wikipedia.org/wiki/Predictive_analytics) [analytics,](https://en.wikipedia.org/wiki/Predictive_analytics) [scientific computing),](https://en.wikipedia.org/wiki/Scientific_computing) that aims to simplify management and deployment. Anaconda3 includes Python 3.6. Anaconda Navigator is a desktop [graphical user interface](https://en.wikipedia.org/wiki/Graphical_user_interface) [(GUI)](https://en.wikipedia.org/wiki/Graphical_user_interface) included in Anaconda distribution that allows users to launch applications and manage anaconda packages, environments and channels without using [command-line commands.](https://en.wikipedia.org/wiki/Command-line_interface) Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install them in an environment, run the packages and update them. It is available for [Windows,](https://en.wikipedia.org/wiki/Windows) [macOS](https://en.wikipedia.org/wiki/MacOS) and [Linux.](https://en.wikipedia.org/wiki/Linux)

The following are the system requirements

* **License:** Free use and redistribution under the terms of the [Anaconda End User](https://docs.anaconda.com/anaconda/eula/) [License Agreement.](https://docs.anaconda.com/anaconda/eula/)
* **Operating system:** Windows Vista or newer, 64-bit macOS 10.10+, or Linux,including Ubuntu, RedHat, CentOS 6+, and others. Windows XP supported on Anaconda versions 2.2 and earlier. See lists. Download it from our [archive.](https://repo.anaconda.com/archive/)
* **System architecture:** 64-bit x86, 32-bit x86 with Windows or Linux, Power8 orPower9. Minimum 3 GB disk space to download and install.

**6.1.2.2 Installation of Anaconda on Windows**

The most common way of downloading anaconda is to go to [www.python.org](http://www.python.org/) and choose the appropriate version and click on download. The following steps must be followed to install anaconda, after downloading it.

1. [Download the Anaconda installer.](https://www.anaconda.com/download/#windows)
2. Optional: [Verify data integrity with MD5 or SHA-256.](https://docs.anaconda.com/anaconda/install/hashes/) [More info on hashes](http://conda.pydata.org/docs/download.html#what-about-cryptographic-hash-verification)
3. Double click the installer to launch. Click
4. Click Next.
5. Read the licensing terms and click “I Agree”.
6. Select an install for “Just Me” unless you’re installing for all users (which requires Windows Administrator privileges) and click next.
7. Select a destination folder to install Anaconda and click the Next button. See [FAQ.](https://docs.anaconda.com/anaconda/faq/#distribution-faq-windows-folder)



|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 27 |

Chapter 6 IMPLEMENTATION



1. Choose whether to add Anaconda to your PATH environment variable. It is recommended to not add Anaconda to the PATH environment variable, since this can interfere with other software. Instead, use Anaconda software by opening Anaconda Navigator or the Anaconda Prompt from the Start Menu.
2. Choose whether to register Anaconda as your default Python 3.6. Unless you plan on installing and running multiple versions of Anaconda, or multiple versions of Python, accept the default and leave this box checked.
3. Click the Install button. If you want to watch the packages Anaconda is installing, click Show Details.
4. Click the Next button.
5. Optional: To [install VS Code,](https://docs.anaconda.com/anaconda/user-guide/tasks/integration/vscode/) click the Install Microsoft VS Code button. After the install completes click the Next button.
6. Or to install Anaconda without VS Code, click the Skip button.
7. After your install is complete, verify it by opening Anaconda Navigator, a program that is included with Anaconda: from your Windows Start menu, select the shortcut Anaconda Navigator. If Navigator opens, you have successfully installed Anaconda. If not, check that you completed each step above, then see our [Help page.](https://docs.anaconda.com/anaconda/help-support/)

**6.1.3 Implementation using Logistic regression**

**6.1.3.1 Supervised Learning**

Supervised learning is the Data mining task of inferring a function from **labelled training data.** The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal).

A supervised learning algorithm analyses the training data and produces an inferred function, which can be used for mapping new examples.

* **Labelled data**: Data consisting of a set of training examples*,*where each example isa *pair* consisting of an input and a desired output value (also called the supervisory signal, labels, etc.)
* **Classification**: The goal is to predict discrete values, e.g. {1, 0}, {True, False},{spam, not spam}.
* **Regression**: The goal is to predict continuous values, e.g. home prices.

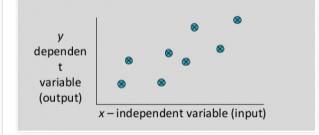


|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 28 |

Chapter 6 IMPLEMENTATION

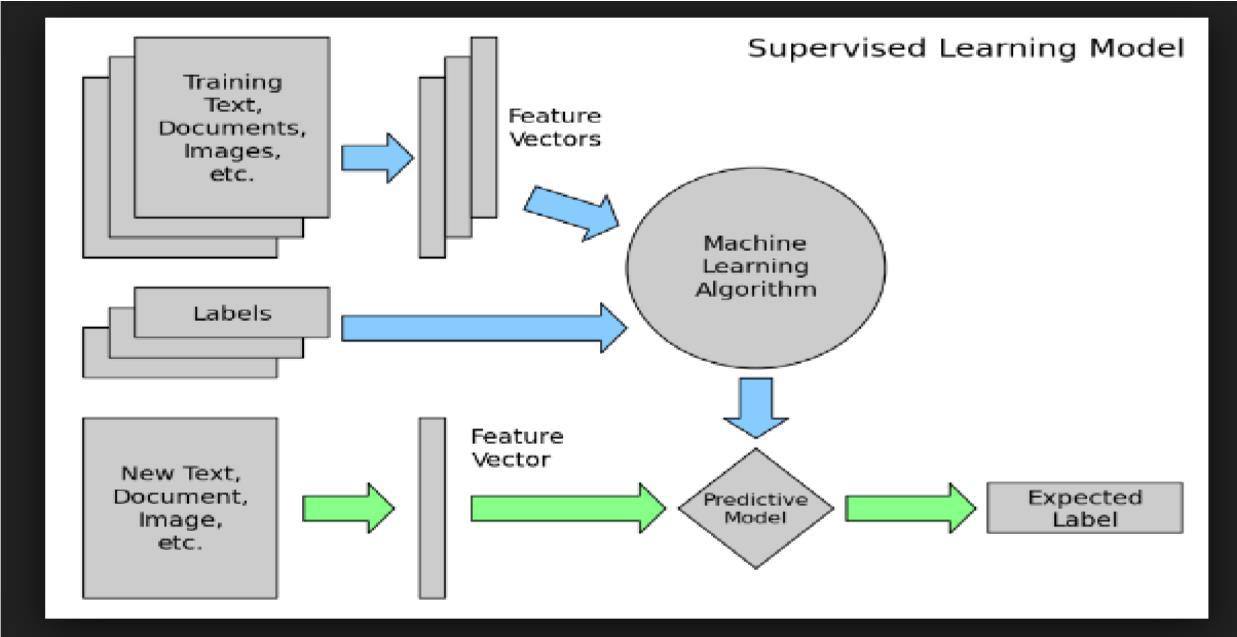


A regression problem is when the output variable is a real or continuous value, such as “salary” or “weight”. Many different models can be used; the simplest is the linear regression. It tries to fit data with the best hyper-plane which goes through the points.



**Figure 6.1 Dependency between Two Dataset**

**6.1.3.2 Supervised learning workflow**



**Figure 6.2 Supervised Learning Workflow**

The supervised learning workflow as shown in figure 6.2 tells us how the training and testing of the machine take place. There are certain rules/steps which we have to follow in the ordered to train our machine these steps are as follows:

1. One has to take a data set in order to train and test the machine, the data set should relevant to our problem, for which they are training the machine.
2. Then they have to find a relation between all the attribute of the data set and have to find the dependency of the data set with each other, this help to select most appropriate attributes to train the machine.



|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 29 |

Chapter 6 IMPLEMENTATION



1. Then they have to break our data set in two sets: a) Training data set (consist of 70 to 85 % of data set), b) Test data set (consist of about 15 to 30 % of the data set).
2. They select most appropriate algorithm to train the machine, in initial stage of machine there are some errors which can be resolved with the error function they have used in our algorithm in each repeated iterations of training.
3. After the machine is trained with the algorithm used and training data with the minimum error they have to calculate the accuracy of the machine by using the testing data.
4. The expected result is labelled and matches with original result to check the accuracy.

**6.1.3.3 Logistic Regression**

Logistic regression is used to find the probability of event\_ success and event failure. The dependent variable must be binary in nature (0 or 1). It is widely used for classification problems. Logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a log it function.

1. **Binomial:** target variable can have only 2 possible types: “0” or “1” which mayrepresent “win” vs. “loss”, “pass” vs. “fail”, “dead” vs. “alive”, etc.
2. **Multinomial:** target variable can have 3 or more possible types which are notordered (i.e. types have no quantitative significance) like “disease A” vs. “disease B” vs. “disease C”.
3. **Ordinal:** it deals with target variables with ordered categories. For example, a test

score can be categorized as: “very poor”, “poor”, “good”, “and very good”. Here, each category can be given a score like 0, 1, 2, and 3.

**6.2 Pseudocode**

Pseudocode is an informal [high-level](https://en.wikipedia.org/wiki/High-level_programming_language) description of the operating principle of a [computer](https://en.wikipedia.org/wiki/Computer_program) [program](https://en.wikipedia.org/wiki/Computer_program) or other algorithm. It uses the structural conventions of a normal [programming](https://en.wikipedia.org/wiki/Programming_language) [language,](https://en.wikipedia.org/wiki/Programming_language) but is intended for human reading rather than machine reading. Pseudocode typically omits details that are essential for machine understanding of the algorithm, such as [variable](https://en.wikipedia.org/wiki/Variable_declaration) [declarations,](https://en.wikipedia.org/wiki/Variable_declaration) system-specific code and some [subroutines.](https://en.wikipedia.org/wiki/Subroutine) Pseudocode resembles, but should not be confused with, programs which can be [compiled](https://en.wikipedia.org/wiki/Compiler) without errors.



|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 30 |

Chapter 6 IMPLEMENTATION



**6.2.1 Reading the CSV Data File and Cleaning the Data**



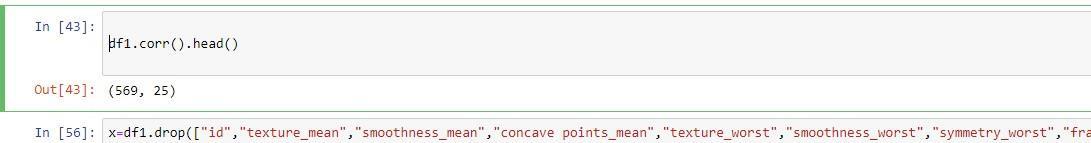
**Figure 6.3 Code for Reading the CSV Data File and Cleaning the Data**

As shown in the figure 6.3, the first step to train the machine is to read and clean the dataset which we are using to train and test the machine. We use read\_csv() function which is define under the pandas module, the read\_csv() function takes one argument which is the name of dataset file name.

In the next line we are using pandas.DataFrame.astype() function which we are using to cast a pandas object to specific Dtype in this case we are typecasting the attribute ‘DIAGNOSIS’ whose value is not ‘B’ as an integer value.

In the next step we are dropping the column value which contain ‘unnamed’ value using pandas.Series.str.Contains (pat,case=False). After this we will check for the attributes which contain the NULL values, we first use df.isnull ().sum () this give how many null value are present in the total data set.

**6.2.2 Finding the correlation between the data**



**Figure 6.4 Code for Finding the Correlation between the Data**



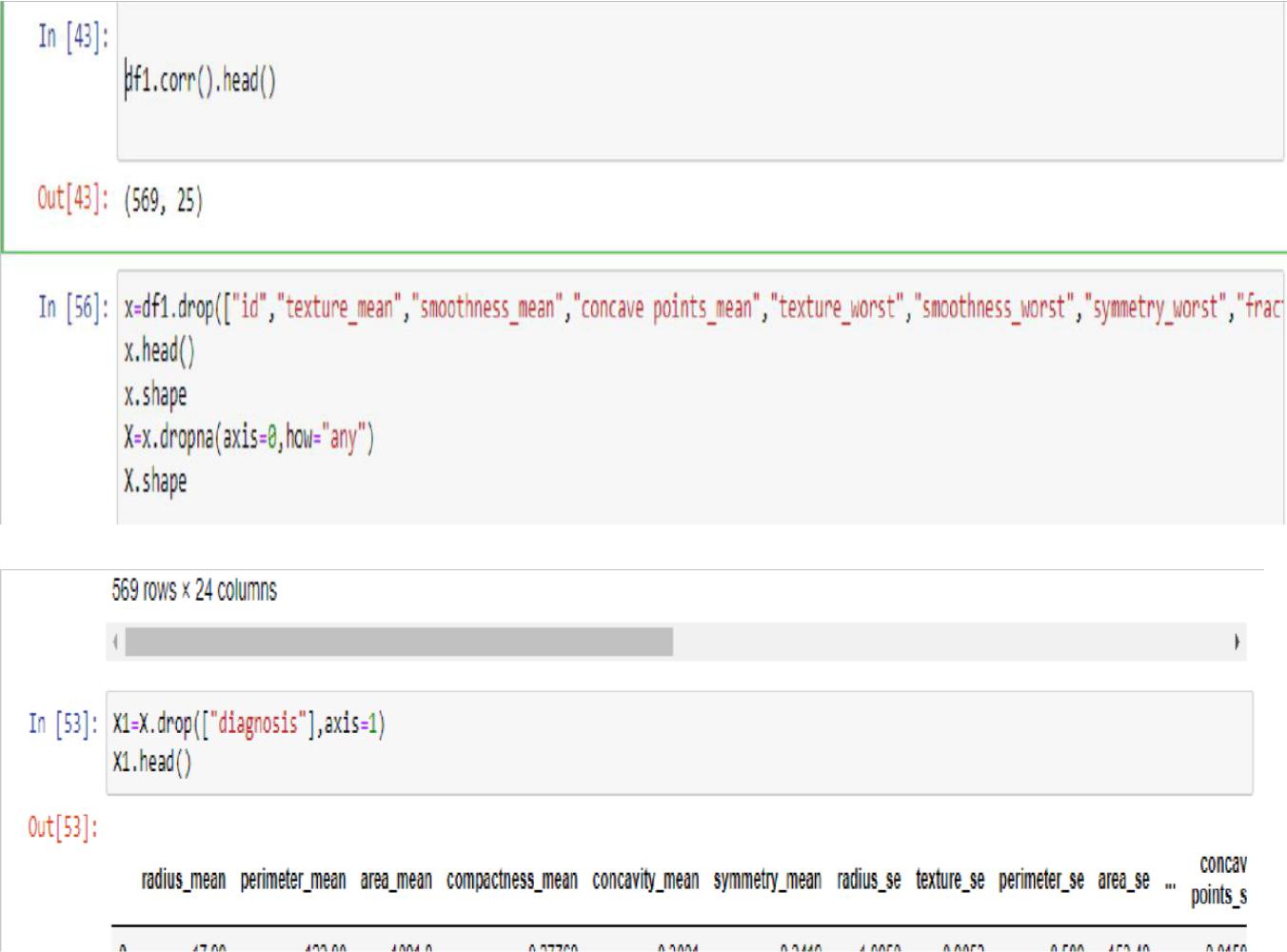
|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 31 |

Chapter 6 IMPLEMENTATION



After cleaning the data we have to find the correlation between the attribute for this we use dataframe.corr() module which is define in pandas. The correlation help us to judge which attribute data is most dependent as shown in the figure 6.4.

**6.2.3 Removing the least dependent attribute from dataset**



**Figure 6.5 Code to Remove the least dependent attribute from Dataset**

Now find the correlation between the data. All the attributes should be removed from the data set which affect the class label the least. As the result of detection does not depend on these data attribute. For this case we are dropping the attributes such as :

["id","texture\_mean","smoothness\_mean","concavepoints\_mean","texture\_worst","smoothne ss\_worst", "symmetry\_worst","fractal\_dimension\_mean"] and assign rest attribute as for X axis parameters.

We have to choose the parameters for x, we assign the x value to new variable x1 which do not contain the diagnosis value. As we do not want diagnosis as x parameter instead we assign diagnosis as y parameters to judge the output result with respect to y axis.

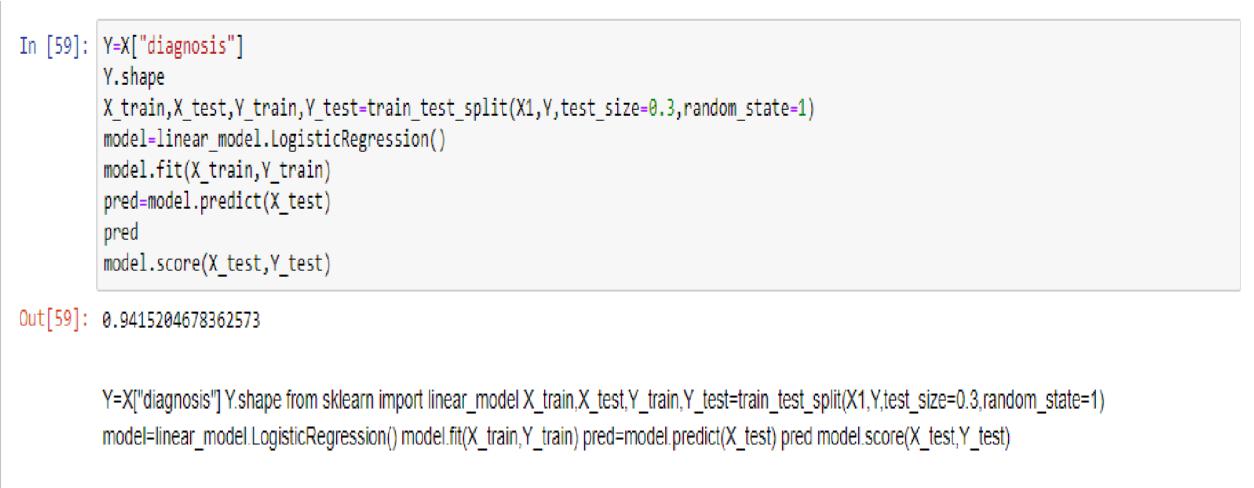


|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 32 |

Chapter 6 IMPLEMENTATION



**6.2.4 Training the Model and Testing the Model**



**Figure 6.6 Code to Train and Test the Model**

Last but not least the dataset has to be split into the training and testing dataset. It is distributed in the ratio of (70:30) % for training and testing part, then we train our machine using logistic regression which is supported in sklearn library.

In order to achieve this we split data set as x\_train and x\_test and Y\_train and Y\_test using train\_test\_split module , after this we set our model as Logistic Regression which is define in linear\_model, then we use model.fit(x\_train,Y\_train ) to train the model, after training gets over we test this against the testing data and predict the accuracy of our model using model.pred(), and then check the score of model.



|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 33 |