**Breast Cancer Prediction using Machine Learning Algorithms**

An Internship Report submitted in partial fulfilment of the requirements for

the award of the degree of

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE ENGINEERING**

Submitted by

**Pagadala Vamsi Krishna**

**Regd.no: 171FA04227**

Under the esteemed guidance of

**Mr. Mohan Rao sir**

(Speaker)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**Vignan’s**

**(Deemed to be University)**

**Vadlamudi**

**2016-2020**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**Vignan’s University**



**Certificate**

This is to certify that **Mittapalli Madhuri** pursuing B. tech in Gitam University under Computer Science and Engineering has completed a four-week **Internship** cum hands-on training program conducted by **Smartbridge** in collaboration with **IBM** at Jawaharlal Nehru Technological University in **“Machine Learning with python and IBM Watson Studio”** during the time period of 27th May, 2019 to 27th June, 2019.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**Vignan’s University**



**DECLARATION**

We hereby declare that the internship review entitled **“Breast Cancer Prediction using Machine Learning Algorithms”** is an original work done in the Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM (Deemed to be University) submitted in partial fulfilment of the requirements for the award of the degree of B. Tech in Computer Science and Engineering.

*Date:*

**171FA04227 Pagadala Vamsi Krishna**

*(Registration No) ( Name ) Signature*

ABSTRACT

The number and size of medical databases are increasing rapidly but most of these data are not analysed for finding the valuable and hidden knowledge. Machine Learning techniques can be used to discover hidden patterns and relationships. Models developed from these techniques are useful for medical practitioners to make right decisions. The present research studied the application of Machine Learning techniques to develop predictive models for breast cancer recurrence in patients who were followed-up. Breast cancer represents one of the diseases that make a high number of deaths every year. It is the most common type of all cancers and the main cause of women’s deaths worldwide. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions. In this project, a performance comparison between different machine learning algorithms such as, Support Vector Machine (SVM), Decision Tree, Naive Bayes (NB) and k Nearest Neighbours (k-NN) on the Breast\_cancer dataset is conducted.

The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, sensitivity and specificity. Experimental results show that SVM gives the highest accuracy (97.13%) with lowest error rate. All experiments are executed using Jupyter Notebook with Anaconda 3(64-bit) as the Base.

CONTENTS

1. Title of the project 6
   1. Introduction 6
   2. Objectives of Research 10
   3. Problem Statement 11
   4. Industry Profile 11
2. Review of Literature 12
3. Data Collection 14
4. Methodology 14
   1. Import Libraries 14
   2. Import Data 16
   3. Exploratory Data Analysis 16
   4. Cleaning Of data 19
   5. Splitting of Data 19
   6. Algorithm 20
   7. Evaluation 25
5. Findings and Suggestions 25
6. Conclusion 27
7. References 27
8. **Predicting Breast Cancer**
   1. **Introduction**

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs** that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. **The primary aim is to allow the computers learn automatically** without human intervention or assistance and adjust actions accordingly.

Some machine learning methods

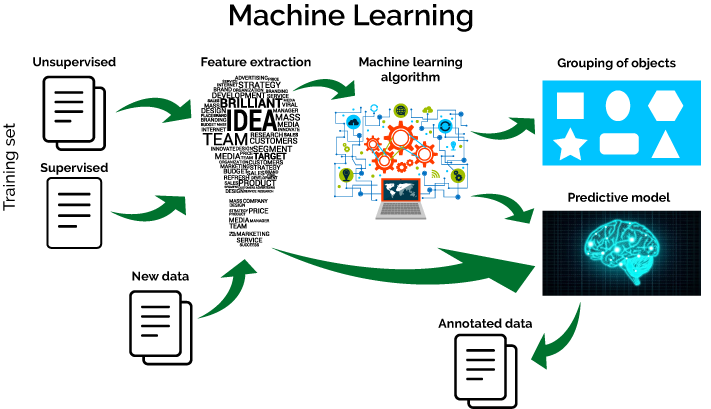
Machine learning algorithms are often categorized as supervised or unsupervised.

**Supervised machine learning algorithms**can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

In contrast, **unsupervised machine learning algorithms**are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn’t figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

**Semi-supervised machine learning algorithms** fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiringunlabeled data generally doesn’t require additional resources.

**Reinforcement machine learning algorithms**is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly. Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information.

Breast cancer is [cancer](https://en.wikipedia.org/wiki/Cancer) that develops from [breast](https://en.wikipedia.org/wiki/Breast) tissue. Signs of breast cancer may include a [lump](https://en.wikipedia.org/wiki/Breast_lump) in the breast, a change in breast shape, [dimpling](https://en.wikipedia.org/wiki/Dimpling) of the skin, fluid coming from the nipple, a newly inverted nipple, or a red or scaly patch of skin. In those with [distant spread of the disease](https://en.wikipedia.org/wiki/Metastatic_breast_cancer), there may be [bone pain](https://en.wikipedia.org/wiki/Bone_pain), swollen [lymph nodes](https://en.wikipedia.org/wiki/Lymph_node), [shortness of breath](https://en.wikipedia.org/wiki/Shortness_of_breath), or [yellow skin](https://en.wikipedia.org/wiki/Jaundice).

Risk factors for developing breast cancer include being female, [obesity](https://en.wikipedia.org/wiki/Obesity), lack of physical exercise, drinking [alcohol](https://en.wikipedia.org/wiki/Alcoholic_drink), [hormone replacement therapy](https://en.wikipedia.org/wiki/Hormone_replacement_therapy) during [menopause](https://en.wikipedia.org/wiki/Menopause), [ionizing radiation](https://en.wikipedia.org/wiki/Ionizing_radiation), early age at [first menstruation](https://en.wikipedia.org/wiki/Menarche), having children late or not at all, older age, prior history of breast cancer, and family history. About 5–10% of cases are due to genes [inherited](https://en.wikipedia.org/wiki/Heredity)  from a person's parents, including [BRCA1](https://en.wikipedia.org/wiki/BRCA1) and [BRCA2](https://en.wikipedia.org/wiki/BRCA2) among others. Breast cancer most commonly develops in cells from the lining of [milk ducts](https://en.wikipedia.org/wiki/Lactiferous_duct) and the [lobules](https://en.wikipedia.org/wiki/Lobules) that supply the ducts with milk. Cancers developing from the ducts are known as [ductal carcinomas](https://en.wikipedia.org/wiki/Mammary_ductal_carcinoma), while those developing from lobules are known as [lobular carcinomas](https://en.wikipedia.org/wiki/Lobular_carcinoma). In addition, there are more than 18 other sub-types of breast cancer. Some cancers, such as [ductal carcinoma in situ](https://en.wikipedia.org/wiki/Ductal_carcinoma_in_situ), develop from [pre-invasive lesions](https://en.wikipedia.org/wiki/Pre-invasive_lesions). The diagnosis of breast cancer is confirmed by taking a [biopsy](https://en.wikipedia.org/wiki/Breast_biopsy) of the concerning lump. Once the diagnosis is made, further tests are done to determine if the cancer has spread beyond the breast and which treatments are most likely to be effective.

The first noticeable [symptom](https://en.wikipedia.org/wiki/Symptom) of breast cancer is typically a [lump](https://en.wikipedia.org/wiki/Breast_lump) that feels different from the rest of the breast tissue. More than 80% of breast cancer cases are discovered when the woman feels a lump. The earliest breast cancers are detected by a [mammogram](https://en.wikipedia.org/wiki/Mammogram). Lumps found in lymph nodes located in the armpits can also indicate breast cancer.

Indications of breast cancer other than a lump may include thickening different from the other breast tissue, one breast becoming larger or lower, a nipple changing position or shape or becoming inverted, skin puckering or dimpling, a rash on or around a nipple, discharge from nipple/s, constant pain in part of the breast or armpit, and swelling beneath the armpit or around the collarbone. Pain ("[mastodynia](https://en.wikipedia.org/wiki/Mastodynia" \o "Mastodynia)") is an unreliable tool in determining the presence or absence of breast cancer, but may be indicative of other [breast health](https://en.wikipedia.org/wiki/Breast_health) issues.

Another symptom complex of breast cancer is [Paget's disease of the breast](https://en.wikipedia.org/wiki/Paget%27s_disease_of_the_breast). This syndrome presents as skin changes resembling eczema, such as redness, discoloration, or mild flaking of the nipple skin. As Paget's disease of the breast advances, symptoms may include tingling, itching, increased sensitivity, burning, and pain. There may also be discharge from the nipple. Approximately half of women diagnosed with Paget's disease of the breast also have a lump in the breast.

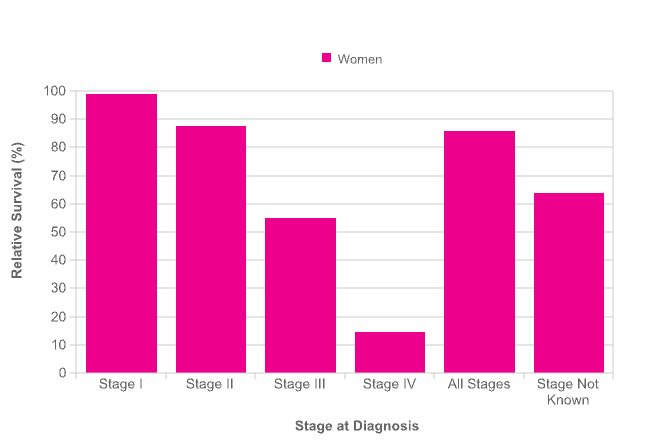
Obesity and drinking alcoholic beverages are among the most common modifiable risk factors. Although the correlation between these factors and breast cancer is anything but linear. Studies show that those who rapidly gain weight in adulthood are at higher risk than those who have been overweight since childhood. Likewise, excess fat in the midsection seems to induce a higher risk than excess weight carried in the lower body. This implies that the food one eats is of greater importance than ones [BMI](https://en.wikipedia.org/wiki/Body_mass_index).

The consumption of alchol seems to be linked to the risk for breast cancer. [Drinking alcoholic beverages increases the risk of breast cancer](https://en.wikipedia.org/wiki/Alcohol_and_breast_cancer), even at relatively low (one to three drinks per week) and moderate levels. The risk is highest among heavy drinkers. Dietary factors which may increase risk include a high-fat diet and obesity-related [high cholesterol](https://en.wikipedia.org/wiki/Hypercholesterolemia) levels. Dietary iodine deficiency may also play a role. Evidence for fiber is unclear. A 2015 review found that studies trying to link fiber intake with breast cancer produced mixed results. In 2016 a tentative association between low fiber intake during adolescence and breast cancer was observed.

[Smoking tobacco](https://en.wikipedia.org/wiki/Tobacco_smoking) appears to increase the risk of breast cancer, with the greater the amount smoked and the earlier in life that smoking began, the higher the risk. In those who are long-term smokers, the risk is increased 35% to 50%. A lack of physical activity has been linked to about 10% of cases. [Sitting](https://en.wikipedia.org/wiki/Sitting) regularly for prolonged periods is associated with higher mortality from breast cancer. The risk is not negated by regular exercise, though it is lowered.

Breast cancer (BC) is the most common cancer in women, affecting about 10% of all women at some stages of their life. In recent years, the incidence rate keeps increasing and data show that the survival rate is 88% after five years from diagnosis and 80% after 10 years from diagnosis. Early prediction of breast cancer is one of the most crucial works in the follow-up process. Machine Learning methods can help to reduce the number of false positive and false negative decisions. Consequently, new methods try to identify and exploit patterns and relationships among large number of variables, and predict the outcome of a disease using historical cases stored in datasets.

In this project, using Machine Learning techniques, we developed models to predict the recurrence of breast cancer by analysing data. The next sections of this project review related work, describe background of this study, evaluate classification models explain the methodology used to conduct the prediction, present experimental results, and the last part of the paper is the conclusion. To estimate validation of the models, accuracy, sensitivity, and specificity were used as criteria, and were compared.

* 1.  **Objectives of Research**

The prediction of breast cancer survivability has been a challenging research problem for many researchers. Since the early dates of the related research, much advancement has been recorded in several related fields. For instance, thanks to innovative biomedical technologies, better explanatory prognostic factors are being measured and recorded; thanks to low cost computer hardware and software technologies, high-volume better-quality data is being collected and stored automatically; and finally, thanks to better analytical methods, those voluminous data are being processed effectively and efficiently. Therefore, the main objective of this manuscript is to report on a research project where we took advantage of those available technological advancements to develop prediction models for breast cancer survivability.

We used two popular Machine Learning algorithms along with a most commonly used statistical methods to develop the prediction models using a large dataset. We also used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes.

* 1. **Problem Statement**

Breast cancer is the most common cancer among women in the United States. Every thirteen minutes a woman dies with the diagnosis of breast cancer. These facts have led researchers to continue studying how to treat and detect breast cancer in women, especially older women, who are of higher risk. Sonography (also known as ultrasound) has become a great addition to mammography and magnetic resonance imaging (MRI) as imaging techniques dedicated to providing breast cancer screening. There has been an increasing interest in a new imaging technique to detect breast cancer deposits in axillary lymph nodes and to provide a noninvasive means to evaluate the stage of the disease in patients. The segmentation of images can provide this outcome. Image segmentation refers to the procedure of partitioning a digital image into various sections. The sections are created to alter the depiction of an image into digital components which make it more straightforward to analyze. Image segmentation is utilized to find discrete objects and boundaries within background images. It would be highly useful to employ image segmentation in medical images because it could assist with determining tissue volumes, diagnosis, localized pathology, and studying anatomical structures5-7. While image segmentation is not a new tool, as far as an assessment technique, it has rarely been used with sonography images. The few clinical research studies that employed segmentation of sonographic images proved to aid in determining heel density, ovarian cysts, breast 1 cysts, fetal, liver, and cardiac pathology8-14. The next step is to determine that the evaluation of axillary lymph nodes with image segmentation has the potential to increase diagnostic accuracy for detecting breast cancer.

1.4 **Industry Profile**

The healthcare industry (also called the medical industry or health economy) is an aggregation and integration of sectors within the [economic system](https://en.wikipedia.org/wiki/Economic_system) that provides goods and services to treat patients with [curative](https://en.wikipedia.org/wiki/Curative_care), [preventive](https://en.wikipedia.org/wiki/Preventive_medicine), [rehabilitative](https://en.wikipedia.org/wiki/Physical_therapy), and [palliative care](https://en.wikipedia.org/wiki/Palliative_care). It includes the generation and commercialization of goods and services lending themselves to maintaining and re-establishing health. The modern healthcare industry is divided into many sectors and depends on the [interdisciplinary teams](https://en.wikipedia.org/wiki/Interdisciplinary_team) of trained professionals and [paraprofessionals](https://en.wikipedia.org/wiki/Paraprofessional) to meet health needs of individuals and populations.

The healthcare industry is one of the world's largest and fastest-growing [industries](https://en.wikipedia.org/wiki/Industries). Consuming over 10 percent of [gross domestic product](https://en.wikipedia.org/wiki/Gross_domestic_product) (GDP) of most [developed nations](https://en.wikipedia.org/wiki/Developed_nations), health care can form an enormous part of a country's economy.

For the purpose of finance and management, the healthcare industry is typically divided into several areas. As a basic framework for defining the sector, the United Nations [International Standard Industrial Classification](https://en.wikipedia.org/wiki/International_Standard_Industrial_Classification) (ISIC) categorizes the healthcare industry as generally consisting of: Hospital activities; Medical and [dental](https://en.wikipedia.org/wiki/Dentistry) practice activities; "Other [human health](https://en.wikipedia.org/wiki/Health) activities".

This third class involves activities of, or under the supervision of, nurses, midwives, physiotherapists, scientific or diagnostic laboratories, pathology clinics, residential health facilities, or other [allied health professions](https://en.wikipedia.org/wiki/Allied_health_profession), e.g. in the field of optometry, hydrotherapy, medical massage, yoga therapy, music therapy, occupational therapy, speech therapy, chiropody, homeopathy, chiropractic, acupuncture, etc.

The [Global Industry Classification Standard](https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard) and the [Industry Classification Benchmark](https://en.wikipedia.org/wiki/Industry_Classification_Benchmark) further distinguish the industry as two main groups: healthcare equipment and services; and [pharmaceuticals](https://en.wikipedia.org/wiki/Pharmaceuticals), [biotechnology](https://en.wikipedia.org/wiki/Biotechnology) and related [life sciences](https://en.wikipedia.org/wiki/Life_sciences).

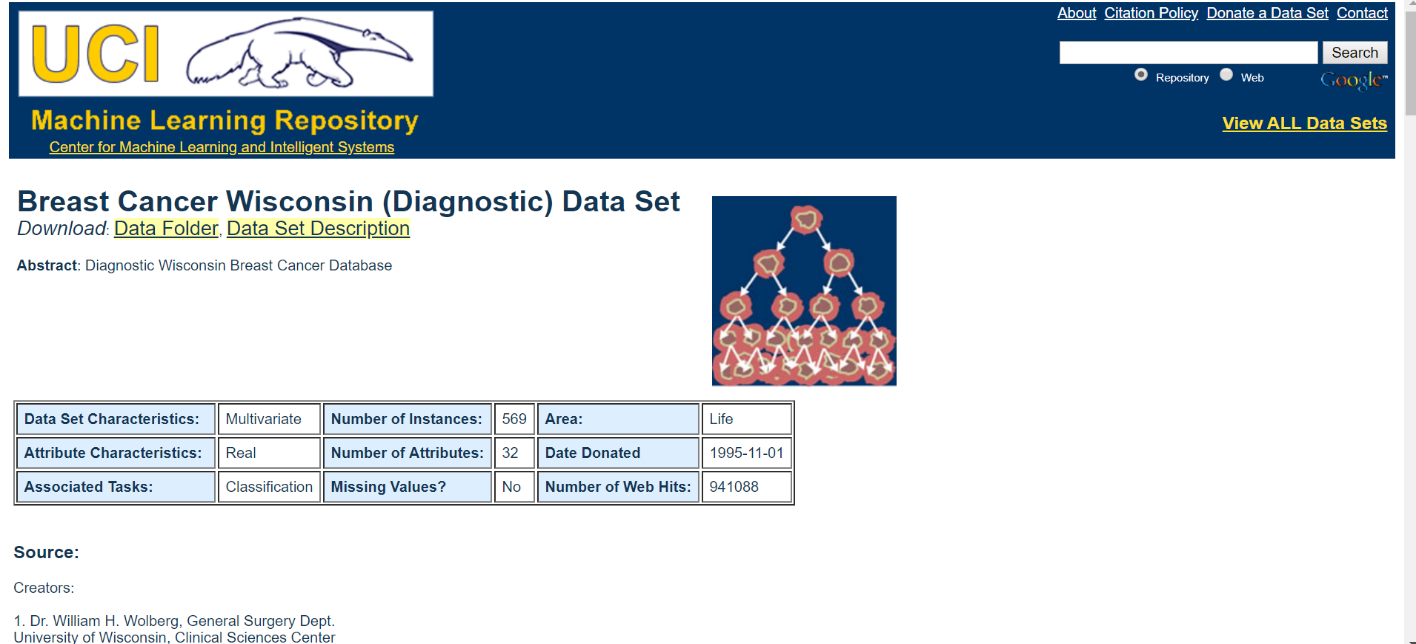
The healthcare equipment and services group consists of companies and entities that provide medical equipment, medical supplies, and healthcare services, such as hospitals, home healthcare providers, and [nursing homes](https://en.wikipedia.org/wiki/Nursing_home). The latter listed industry group includes companies that produce biotechnology, pharmaceuticals, and miscellaneous scientific services.

Other approaches to defining the scope of the healthcare industry tend to adopt a broader definition, also including other key actions related to health, such as education and training of health professionals, regulation and management of health services delivery, provision of [traditional](https://en.wikipedia.org/wiki/Traditional_medicine) and complementary medicines, and administration of [health insurance](https://en.wikipedia.org/wiki/Health_insurance).

1. **Review of Literature**

A literature review shows that there have been several studies on the survival prediction problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and recurrence using Machine Learning approaches such as decision trees. Delen et al. used artificial neural networks, decision trees and logistic regression to develop prediction models for breast cancer survival by analysing a large dataset, the SEER cancer incidence database. Lundin et al. used ANN and logistic regression models to predict 5, 10, and 15 -year breast cancer survival. They studied 951 breast cancer patients and used tumour size, axillary nodal status, histological type, mitotic count, nuclear pleomorphism, tubule formation, tumour necrosis, and age as input variables. Pend Harker patterns in breast cancer. In this study, they showed that data mining could be a valuable tool in identifying similarities (patterns) in breast cancer cases, which can be used for diagnosis, prognosis, and treatment purposes. These studies are some examples of researches that apply data mining to medical fields for prediction of diseases.

AlirezaOsarech, BitaShadgar used SVM classification technique on two different benchmark datasets for breast cancer which got 98.80% and 96.63% accuracies. MandeepRana, PoojaChandorkar, AlishibaDsouza worked on the diagnosis and the prediction of recurrence of breast cancer by applying KNN, SVM, Naïve Bayes and Logistic Regression techniques, programmed in MATLAB. The classification techniques are applied on two datasets taken from UCI depository. A dataset of them is used for identification of disease(WDBC) and the next one is used for recurrence prediction (WPBC)[3].VikasChaurasia, BB Tiwari and Saurabh Pal used three famous algorithms such as J48, Naive bayes, RBF, to build predictive models on breast cancer prediction and compared their accuracy. The results had shown that Naive Bayes predicted well among them with an accuracyof97.36%. Haifeng Wang and Sang Won Yoon compared Naive Bayes Classifier, Support Vector Machine (SVM), AdaBoost tree, Artificial Neural Networks (ANN), to find a powerful model for breast cancer prediction. They implemented PCA for dimensionality reduction. S.Kharya worked on breast cancer prediction and stated that artificial neural networks are widely used. The paper featured about the advantages and short comings of using machine learning methods like SVM, Naive Bayes, Neural network and Decision trees. NareshKhuriwal, Nidhi Mishra took data from Wisconsin Breast Cancer database and worked on breast cancer diagnosis. The results of their experiments proved that ANN and Logistic Algorithm worked better and provided a good solution. It achieved an accuracy of 98.50%.

1. **Data Collection**

**Data Source:**

<https://archive.ics.uci.edu/ml/datsets/Breast+cancer+Wisconsin+(Diagnostic)>

The data was duly collected from the above link and reverified for changed attributes by visiting hospitals and checking for latest changes. The data set was modified accordingly and then used.

1. **Methodology**

**4.1** Importing Libraries

To perform majority of tasks in python it is very much important to import libraries that are used in python same as in C, C++ and java. There are different kinds of libraries that their respective functions in python and some of them are,

* + 1. **NumPy**

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

* A powerful N-dimensional array object.
* Sophisticated (broadcasting) functions.
* Tools for integrating C/C++ and Fortran code.
* Useful linear algebra, Fourier transform, and random number capabilities.
  + 1. **Pandas**

Pandas is the most popular python library that is used for data analysis. It provides highly optimized performance with back-end source code is purely written in C or Python.

We can analyse data in pandas with:

* Series
* Data Frames
  + 1. **Matplotlib**

Matplotlib is a python library used to create 2D graphs and plots by using python scripts. It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc. It supports a very wide variety of graphs and plots namely - histogram, bar charts, power spectra, error charts etc. It is used along with NumPy to provide an environment that is an effective open source alternative for MATLAB. It can also be used with graphics toolkits like PyQt and wxPython.

* + 1. **Seaborn**

Seaborn is a library for making statistical graphics in Python. It is built on top of [matplotlib](https://matplotlib.org/) and closely integrated with [pandas](https://pandas.pydata.org/) data structures.

Here is some of the functionality that seaborn offers:

* A dataset-oriented API for examining [relationships](https://seaborn.pydata.org/examples/scatter_bubbles.html#scatter-bubbles) between [multiple variables](https://seaborn.pydata.org/examples/faceted_lineplot.html#faceted-lineplot).
* Specialized support for using categorical variables to show [observations](https://seaborn.pydata.org/examples/jitter_stripplot.html#jitter-stripplot) or [aggregate statistics](https://seaborn.pydata.org/examples/pointplot_anova.html#pointplot-anova).
* Options for visualizing [univariate](https://seaborn.pydata.org/examples/distplot_options.html#distplot-options) or [bivariate](https://seaborn.pydata.org/examples/joint_kde.html#joint-kde) distributions and for [comparing](https://seaborn.pydata.org/examples/horizontal_boxplot.html#horizontal-boxplot) them between subsets of data.
* Automatic estimation and plotting of [linear regression](https://seaborn.pydata.org/examples/anscombes_quartet.html#anscombes-quartet) models for different kinds [dependent](https://seaborn.pydata.org/examples/logistic_regression.html#logistic-regression) variables.
  + 1. **Sklearn**

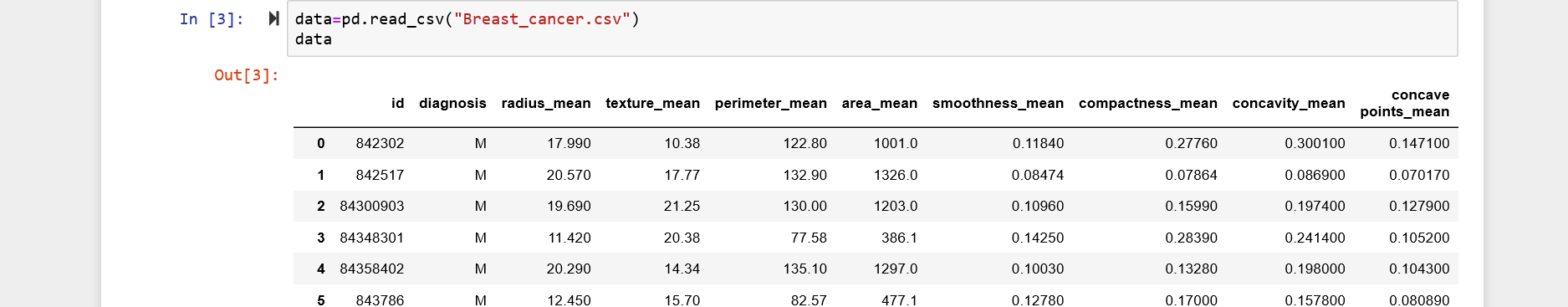
Scikit-learn (formerly scikits.learn) is a [free software](https://en.wikipedia.org/wiki/Free_software) [machine learning](https://en.wikipedia.org/wiki/Machine_learning) [library](https://en.wikipedia.org/wiki/Library_(computing)) for the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) programming language. It features various [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and [clustering](https://en.wikipedia.org/wiki/Cluster_analysis) algorithms including [support vector machines](https://en.wikipedia.org/wiki/Support_vector_machine), [random forests](https://en.wikipedia.org/wiki/Random_forests), [gradient boosting](https://en.wikipedia.org/wiki/Gradient_boosting), [k-means](https://en.wikipedia.org/wiki/K-means_clustering) and [DBSCAN](https://en.wikipedia.org/wiki/DBSCAN), and is designed to interoperate with the Python numerical and scientific libraries [NumPy](https://en.wikipedia.org/wiki/NumPy) and [SciPy](https://en.wikipedia.org/wiki/SciPy).

* 1. Import Data

To pull data from a CSV file, you must use the reader function to generate a reader object.

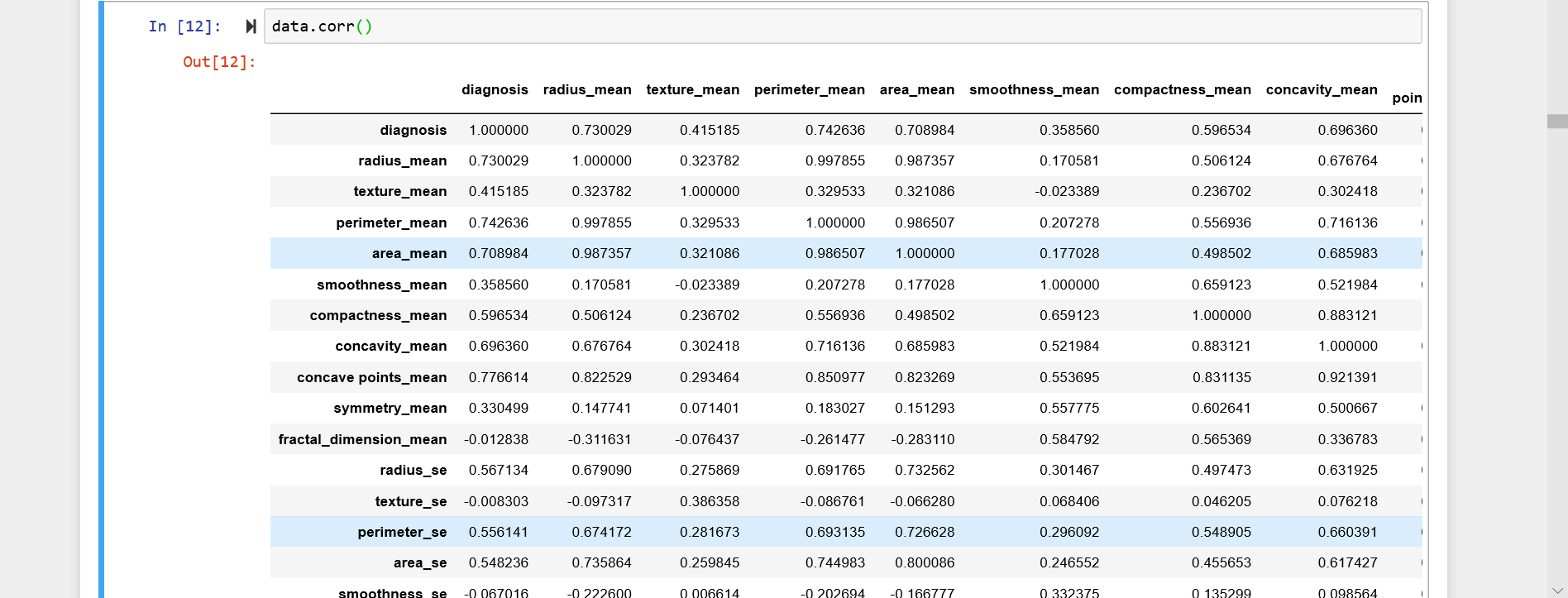
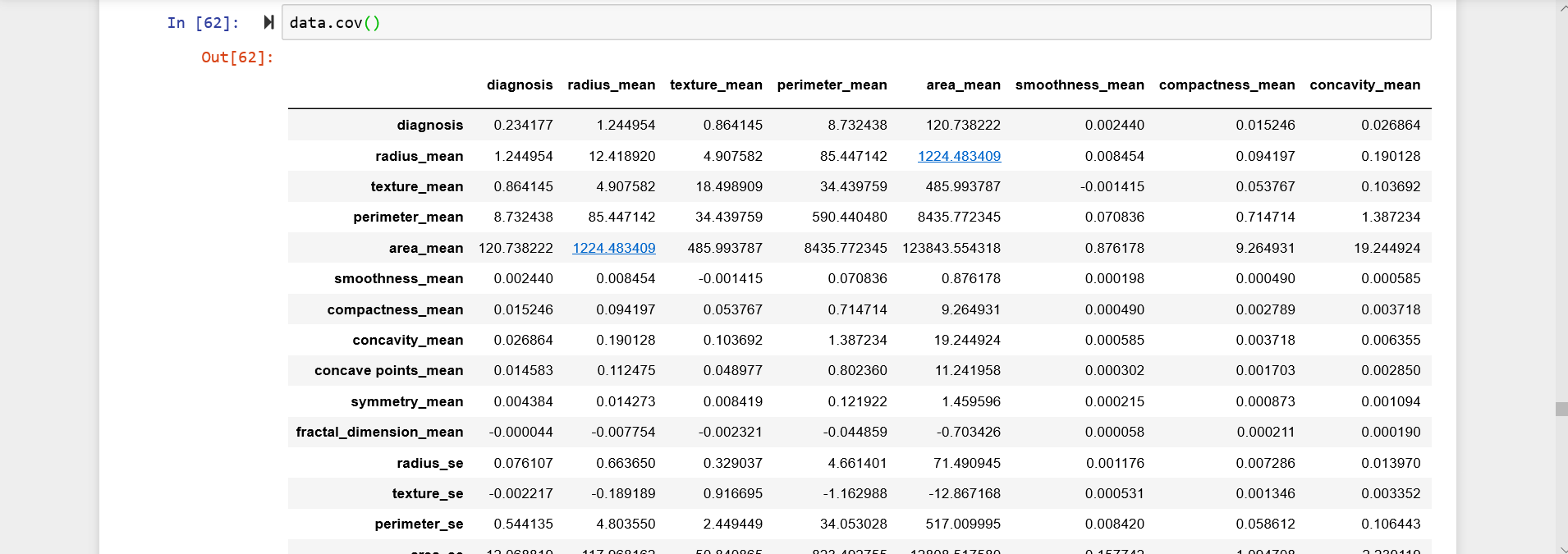
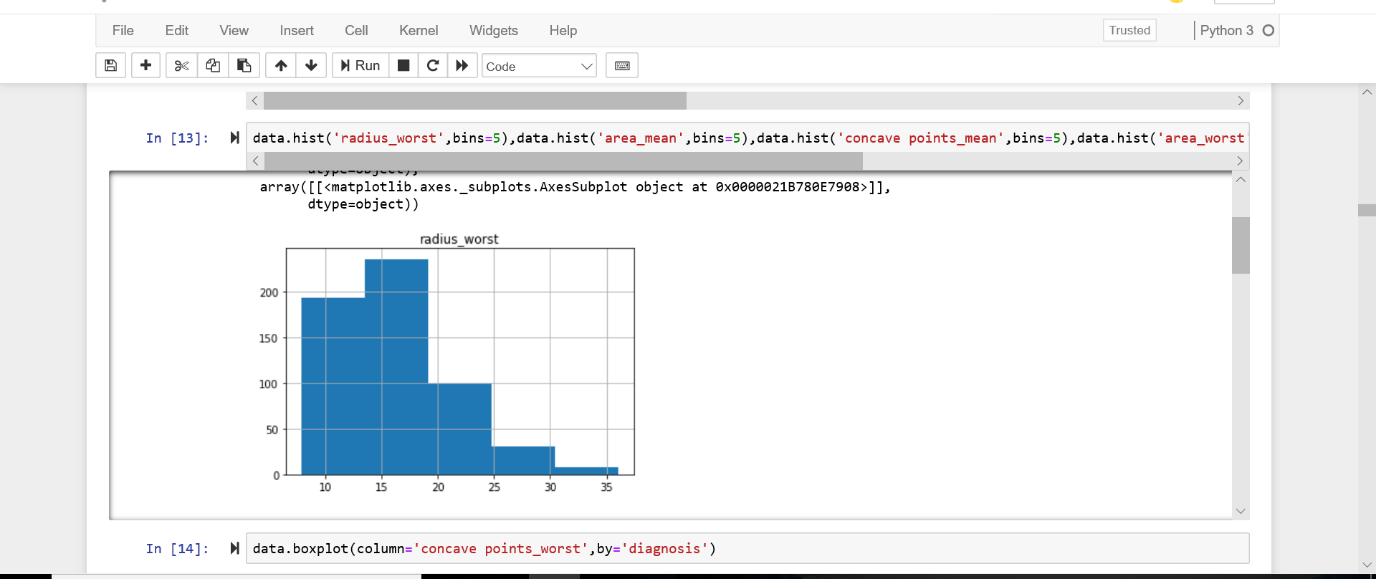
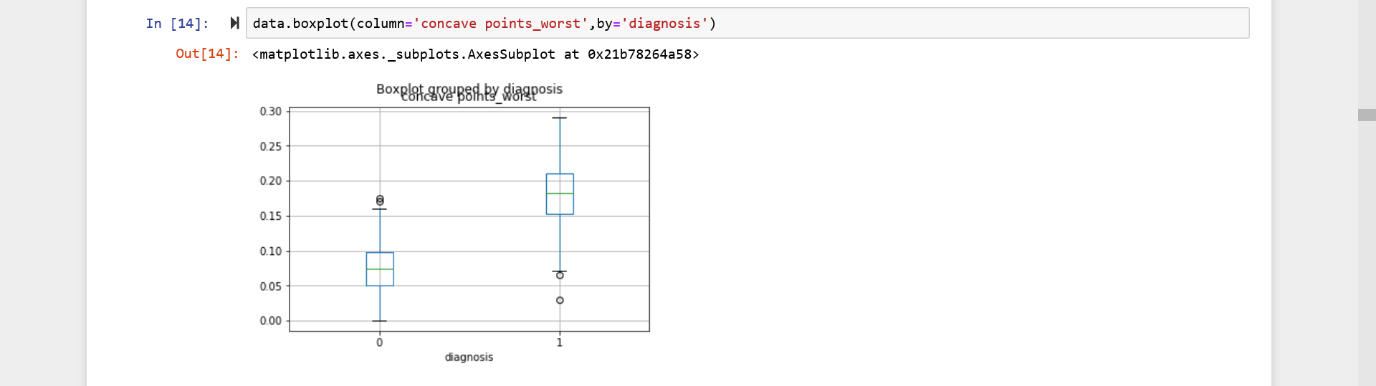
The reader function is designed to take each line of the file and make a list of all columns. Then, you just choose the column you want the variable data for.

It sounds a lot more complicated than it is. To prove it, let’s take a look at an example.

Next, we create the reader object, iterate the rows of the file, and then print them. Finally, we get Data.

* 1. Exploratory Data Analysis

In [statistics](https://en.wikipedia.org/wiki/Statistics) and in Machine Learning, exploratory data analysis (EDA) is an approach to analysing [data sets](https://en.wikipedia.org/wiki/Data_set) to summarize their main characteristics, often with visual methods. A [statistical model](https://en.wikipedia.org/wiki/Statistical_model) can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. Exploratory data analysis was promoted by [John Tukey](https://en.wikipedia.org/wiki/John_Tukey) to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from [initial data analysis (IDA)](https://en.wikipedia.org/wiki/Data_analysis#Initial_data_analysis), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA. There are many subparts in Exploratory Data Analysis and some of them are,

* + 1. Correlation
    2. Covariance
    3.  Histogram
    4.  Boxplot
    5.  Pie chart
  1. Cleaning of Data

The cleaning of data is very much important to increase all the parameters of the data. Cleaning up of data means the removal of null values from the data that makes it more robust to perform different types of operations.

* 1. Splitting of Data

When you’re working on a model and want to train it, you obviously have a dataset. But after training, we have to test the model on some test dataset. For this, you’ll a dataset which is different from the training set you used earlier. But it might not always be possible to have so much data during the development phase.

In such cases, the obviously solution is to split the dataset you have into two sets, one for training and the other for testing; and you do this before you start training your model.

But the question is, how do you split the data? You can’t possibly manually split the dataset into two. And you also have to make sure you split the data in a random manner. To help us with this task, the SciKit library provides a tool, called the Model Selection library. There’s a class in the library which is, aptly, named ‘[train\_test\_split](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html" \t "_blank).’ Using this we can easily split the dataset into the training and the testing datasets in various proportions.

There are a few parameters that we need to understand before we use the class:

**test\_size** — This parameter decides the size of the data that has to be split as the test dataset. This is given as a fraction. For example, if you pass 0.5 as the value, the dataset will be split 50% as the test dataset. If you’re specifying this parameter, you can ignore the next parameter.

**train\_size** — You have to specify this parameter only if you’re not specifying the test\_size. This is the same as test\_size, but instead you tell the class what percent of the dataset you want to split as the training set.

**random\_state** — Here you pass an integer, which will act as the seed for the random number generator during the split. Or, you can also pass an instance of the RandomState class, which will become the number generator. If you don’t pass anything, the RandomState instance used by np.random will be used instead.

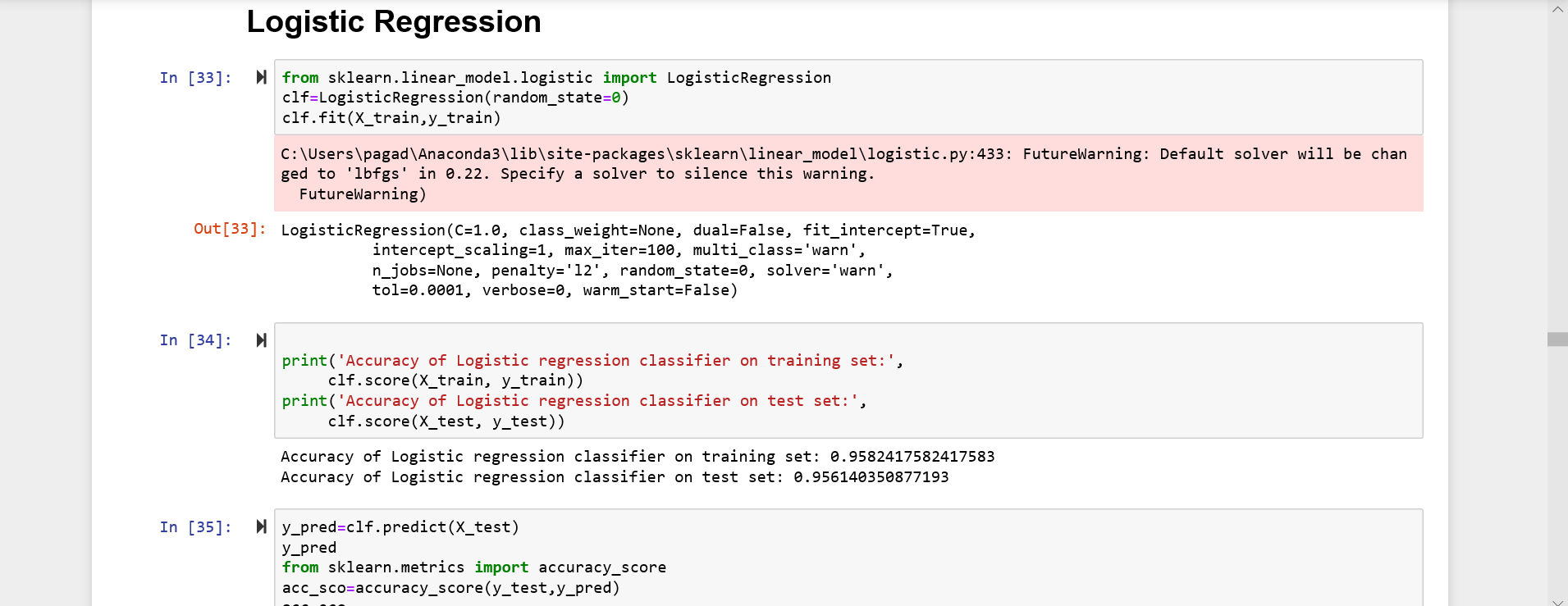
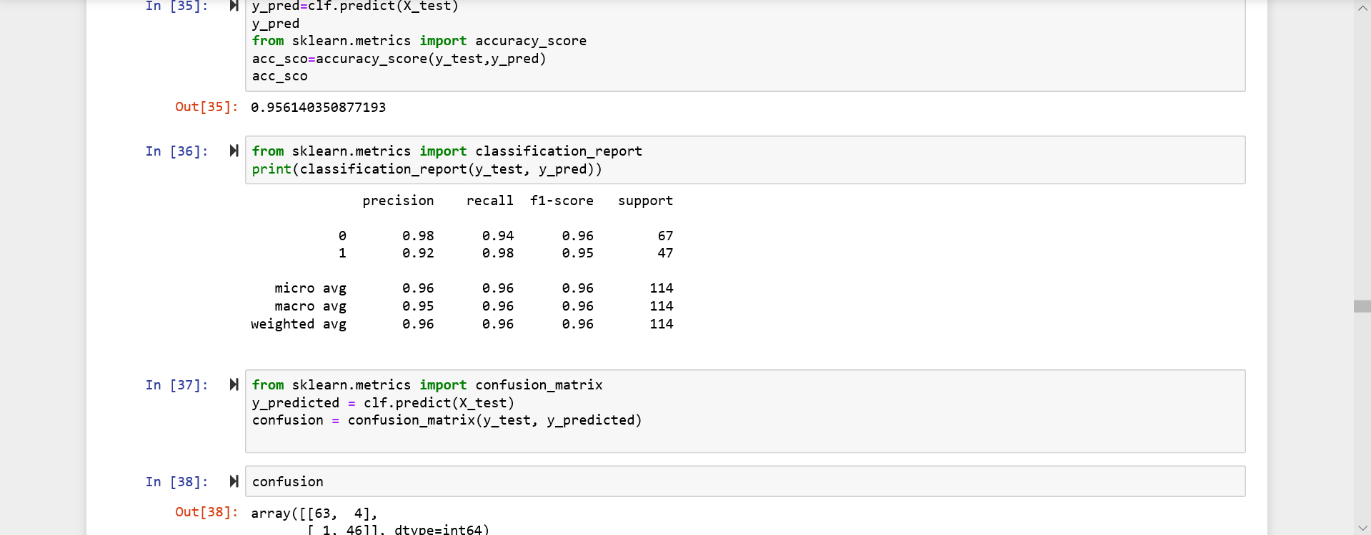
* 1. Algorithm

As the data is read into the python notebook using pandas and os, the data is cleaned i.e. the missing values are replaced and now the data is ready to be used in an algorithm for training and testing the data and predicting new values for the dependent variable. Some of the Algorithms are,

* + 1. Logistic Regression

**Definition:**Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.

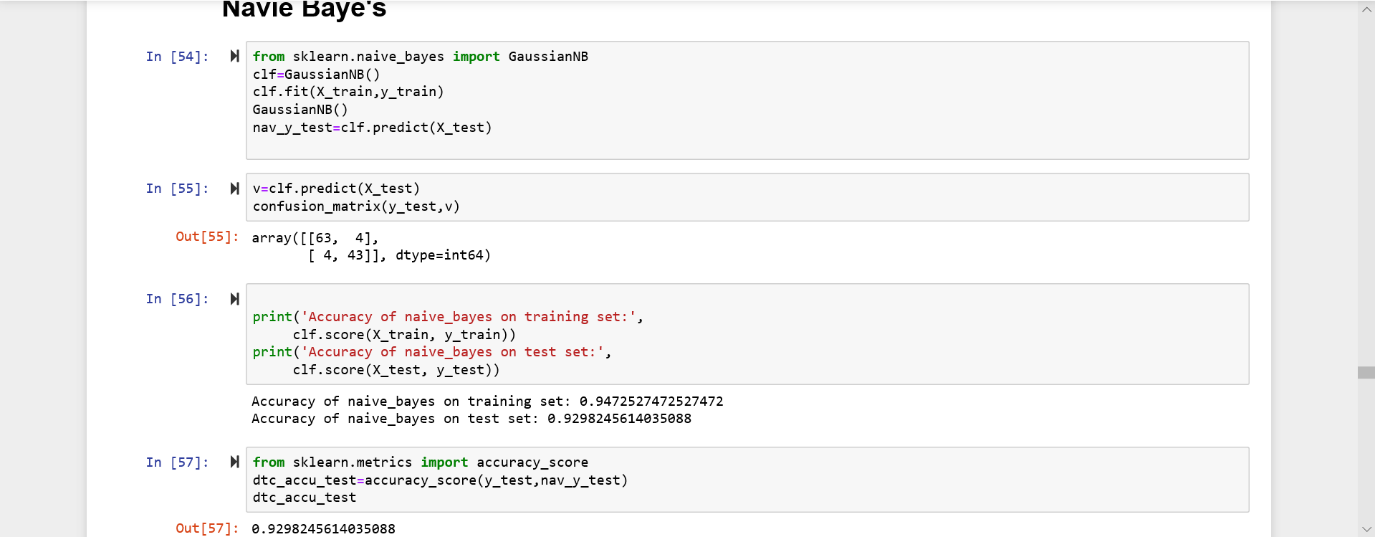
**Advantages:** Logistic regression is designed for this purpose (classification), and is most useful for understanding the influence of several independent variables on a single outcome variable.

**Disadvantages:** Works only when the predicted variable is binary, assumes all predictors are independent of each other, and assumes data is free of missing values.

* + 1. Naive Bayes

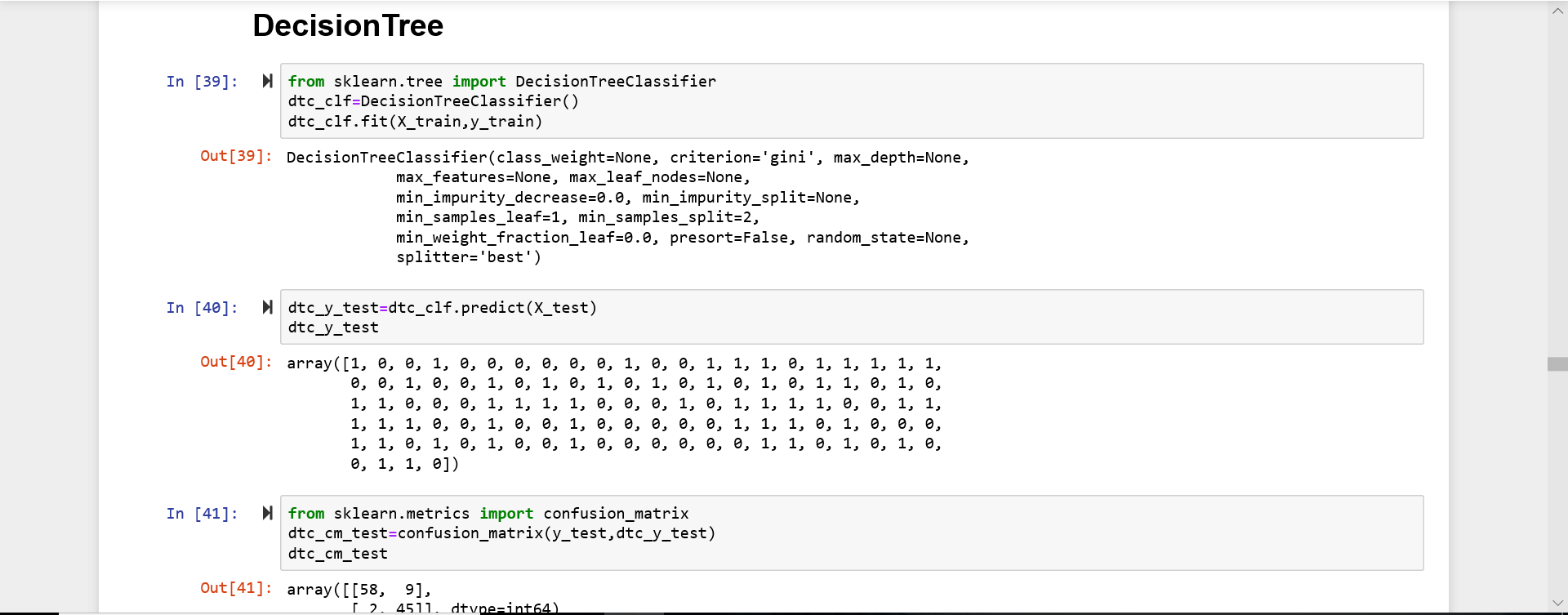
**Definition:**Naive Bayes algorithm based on Bayes’ theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as document classification and spam filtering.

**Advantages:**This algorithm requires a small amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely fast compared to more sophisticated methods.

**Disadvantages:**Naive Bayes is is known to be a bad estimator.

* + 1. Decision Tree

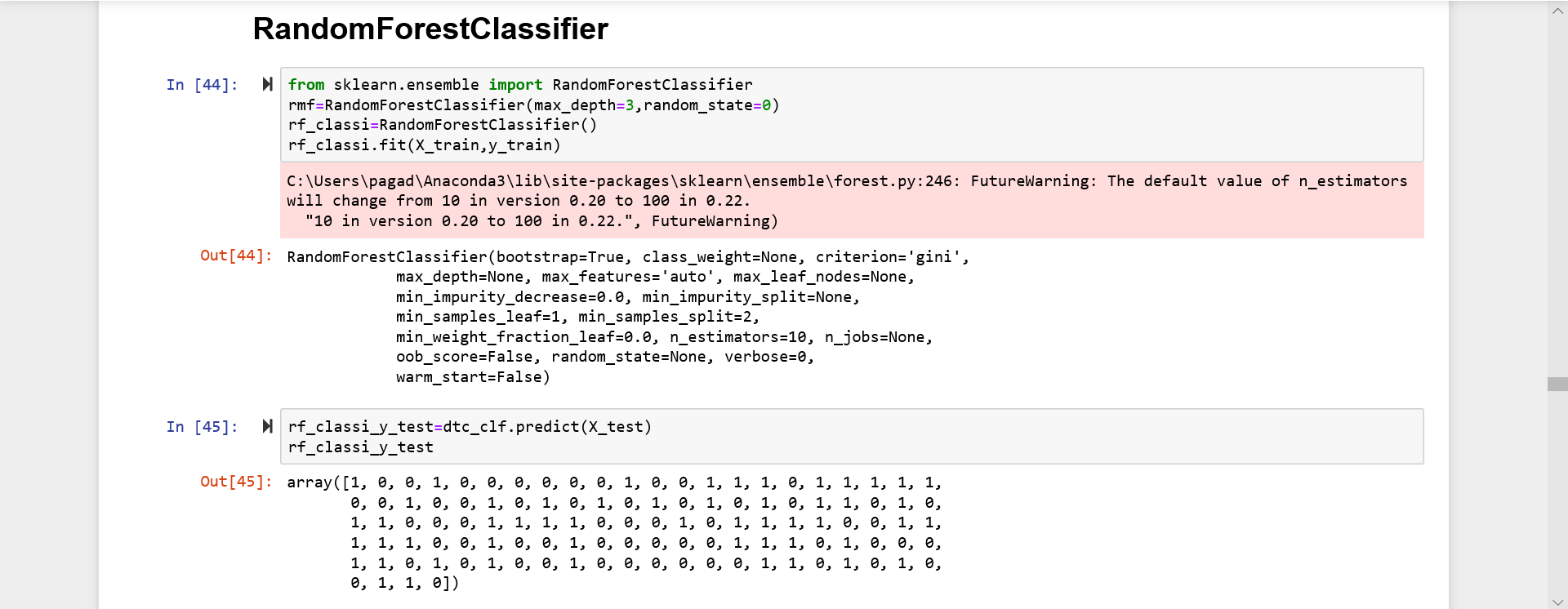
**Definition:** Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.

**Advantages:** Decision Tree is simple to understand and visualise, requires little data preparation, and can handle both numerical and categorical data. **Disadvantages:**Decision tree can create complex trees that do not generalise well, and decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

* + 1. Random Forest

**Definition:**Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

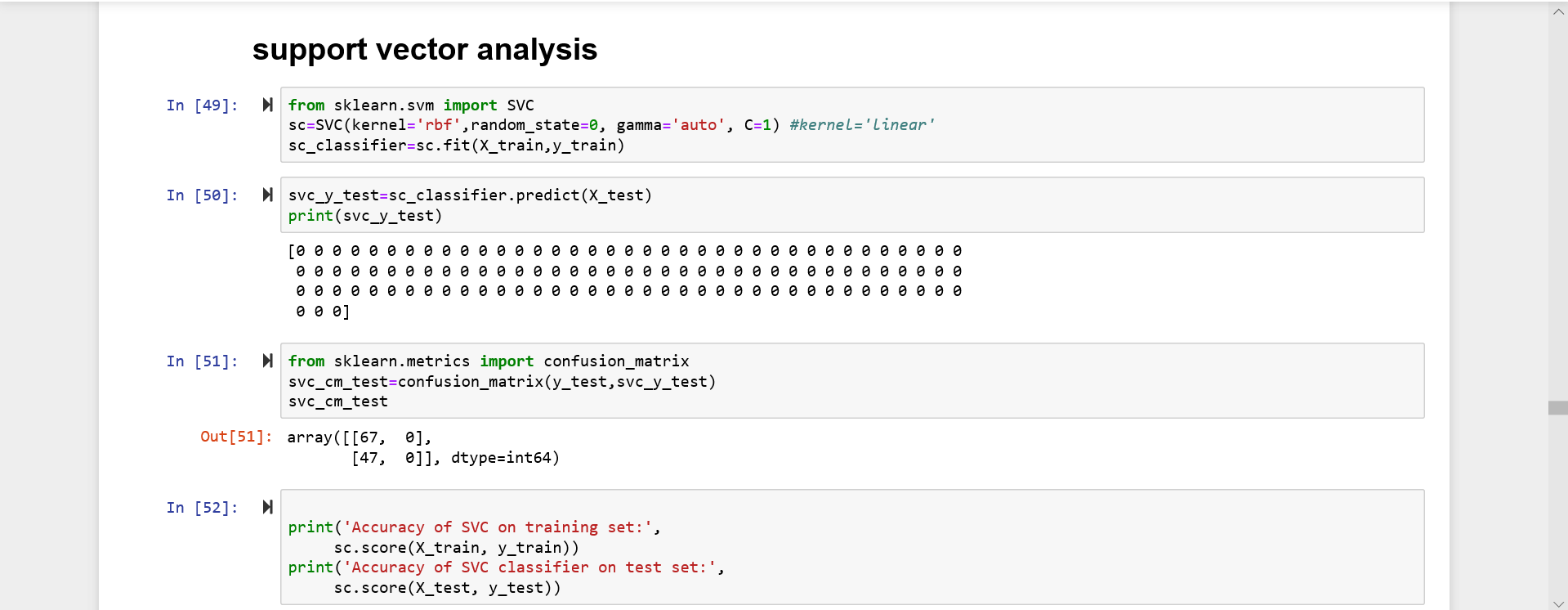
**Advantages:**Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases.

**Disadvantages:**Slow real time prediction, difficult to implement, and complex algorithm.

* + 1. Support Vector Classification

**Definition:**Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

**Advantages:**Effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient.

**Disadvantages:**The algorithm does not directly provide probability estimates; these are calculated using an expensive five-fold cross-validation.

* + 1. KNN

**Definition:** Neighbours based classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbours of each point.

**Advantages:** This algorithm is simple to implement, robust to noisy training data, and effective if training data is large.

**Disadvantages:** Need to determine the value of K and the computation cost is high as it needs to computer the distance of each instance to all the training samples.

* 1. Evaluation

****As the algorithm is successfully applied and we found the prediction values successfully. Now the question arises that to which extent the predicted values are correct. To Measure that we find the accuracy of that particular algorithm using accuracy score from sklearn and confirm with the accuracy and decide that it may happen or not. Basically, an accuracy of 80% and above are considered to be the best model. In the same way the sensitivity and specificity are also measured to determine some other specific factors of the data and the predicted values. Accuracy is done as follows,

1. **Findings and Suggestions**

The data that we processed is about Breast Cancer prediction collected from a trusted source. The data was imported and then cleaned up. The data was then studied using Exploratory data analysis and the relation between every factor was observed. All the columns of the data except Diagnosis which was the target were taken as the predictors as this is the health care industry, definitely each and every factor can affect the health of the human so there is no was of neglecting/dropping a column whereas we can unconditionally drop the id which has no relation. Then the data is split into training and testing data with a ratio of 80% to the training data and 20% to the test data, which is used to train and predict values.

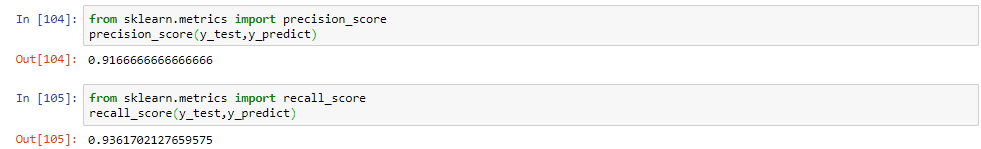
Different classification Algorithms were applied on the data and the Diagnosis values were predicted. To know exactly which model produced the most accurate values we found the testing and training accuracies of the algorithms. The following table on the next page consists the accuracies in percentage. After analysing the algorithms specifically, it is found that KNN algorithm produces the most accurate values as the percentage values of the training and testing data are very closer and hence the KNN algorithm is then used to predict values for the split of data of 90:10, 70:30 , 60:40, 50:50.

|  |  |  |
| --- | --- | --- |
| S. No | Algorithms | Accuracy1 |
| 1 | Logistic Regression | Train - 94% |
|  |  | Test – 92% |
| 2 | Decision Tree | Train – 100% |
|  |  | Test – 88% |
| 3 | Random Forest | Train – 99% |
|  |  | Test – 93% |
| 4 | Support Vector Classification | Train – 100% |
|  |  | Test – 58% |
| 5 | Naive Bayes | Train – 94% |
|  |  | Test – 92% |
| 6 | K- Nearest Neighbours | Train – 94% |
|  |  | Test – 93% |

After the algorithm was applied to different ratios of training and testing data the following are the accuracies respectively,

|  |  |  |
| --- | --- | --- |
| S. No | Ratio of Split | Accuracy |
| 1 | 90% - 10% | Train – 95% |
|  |  | Test – 89% |
| 2 | 70% - 30% | Train – 93% |
|  |  | Test – 94% |
| 3 | 60% - 40% | Train – 93% |
|  |  | Test – 95% |
| 4 | 50% - 50% | Train – 94% |
|  |  | Test – 93.8% |

The next Code shows the specificity and the sensitivity of the data using kNN algorithm,



1. **Conclusion**

To analyse medical data, various machine learning methods are available. An important challenge in machine learning areas is to build accurate and computationally efficient classifiers for Medical applications. In this study, we employed four main algorithms such as SVM, NB, k-NN, logistic, Decision tree and Random Forest on the Breast\_Cancer dataset. We tried to compare efficiency and effectiveness of those algorithms in terms of accuracy, sensitivity and specificity to find the best classification accuracy. Of kNN reaches and accuracy of 93.8% and outperforms, therefore, all other algorithms. In conclusion, kNN has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of accuracy and low error rate.

1. **References**

7.1 <https://archive.ics.uci.edu/ml/datsets/Breast+cancer+Wisconsin+(Diagnostic)>

7.2 [www.wikipedia.com](http://www.wikipedia.com)

7.3 [www.geeksforgeeks.com](http://www.geeksforgeeks.com)

7.4 [www.tutorialspoint.com](http://www.tutorialspoint.com)

7.5 [www.stackoverflow.com](http://www.stackoverflow.com)

7.6 [www.python.org](http://www.python.org)