BREAST CANCER CLASSIFICATION

PROJECT REPORT

By

NAME OF THE CANDIDATE(S)

NAME	DEPARTMENT	YEAR	SEM	ROLL NO
Ishika Acharjee	**Computer Science and Engineering**	**3rd**	**5th**	**11900121181**
Monali Chaki	**Computer Science and Engineering**	**3rd**	**5th**	**11900121194**
Sajit Tamang	**Computer Science and Engineering**	**3rd**	**5th**	**11900121186**
Saurav Karmakar	**Computer Science and Engineering**	**3rd**	**5th**	**11900121199**
Sourav Tamang	**Computer Science and Engineering**	**3rd**	**5th**	**11900121200**

Department of Computer Science and Engineering

Siliguri Institute of Technology

September,2022

BONAFIDE CERTIFICATE

This is to certify that this project report entitled **"BREAST CANCER CLASSIFICATION" submitted** to **Siliguri Institute of Technology, Sukna** is a bonafide record of work done by "**Ishika Acharjee"** under my supervision from **"14/9/2022**" to " **15/9/2022"…**

X Ishika Acharjee

ISHIKA ACHARJEE
GROUP LEADER

X Arpan Samanta

ARPAN SAMANTA
PROJECT GUIDE

SIKHARTHY INFOTECH PVT. LTD

Place Date

15-09-2022

Declaration by Author(s)

This is to declare that this report has been written by me. No part of the report is plagiarized from other sources. All information included from other sources have been duly acknowledged. I aver that if any part of the report is found to be plagiarized, I shall take full responsibility for it.

X Ishika Acharjee

ISHIKA ACHARJEE
GROUP LEADER

Place Date

15-09-2022

ACKNOWLEDGEMENT

It is a great pleasure for me to acknowledge the assistance and participation of a large number of individuals to this attempt. My project report has been structured under the valued suggestion, support and guidance of **Mr.Arpan Samanta**. Under his guidance I have accomplished the challenging task in a very short time.

Finally, I express my sincere thankfulness to my family members for inspiring me all throughout and always encouraging me.

ISHIKA ACHARJEE

Department of CSE

TABLE OF CONTENTS

2.Technical Description

3.Coding & Output Section

INTRODUCTION:

Breast cancer is a widely occurring cancer in women worldwide and is related to high mortality. The objective of this project was to present several approaches to investigate the data on machine learning (ML) approach for early breast cancer detection.

Technologies in healthcare include maintenance and retrieval of electronic medical records of patients and devices involved. Cancer detection has always been a challenge in the diagnosis and treatment plan for haematological diseases. Currently, an overwhelming percentage of the population is affected by one or more diseases. Recent years have seen tremendous advances in medical science. Despite these advancements, there is still a huge lack of information among the public regarding health and disease. A large proportion of the population likely suffer from health issues, some of which may even be fatal. In addition to improving the accuracy of the rapid detection of fatal conditions, adopting safe, realistic techniques and using modern technology can reduce the need for caregivers and reduce overall health care costs. Several lives could be saved through innovations in intelligent decision-making strategies and technologies.

ML is a branch of artificial intelligence (AI) that is used to classify data based on models which have been developed and for predictive analytics, in particular breast cancer25,26. It provides tools by which large quantities of data can be automatically analyzed.

This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this i have used machine learning classification methods to fit a function that can predict the discrete class of new input.

PROJECT OVERVIEW

This app BREAST CANCER CLASSIFICATION can be used to predict malignant or benign cancer, it will support both stand alone and networking environment.

The app use the main modules involved in this system are:

BREAST CANCER CLASSIFICATION USING MACHINE LEARNING WITH PYTHON

Types of Tumor

1) 1. Benign Tumor (tumors which doesn't move to the other parts of the body, non-cancerous, slow growing)
2) 2. Malignant Tumor (tumors which can move from one part of the body to other part, cancerous, fast growing)
3) Dataset
4) Fine needle aspiration (type of biopsy procedure, a thin needle which is inserted into an area of abnormal-appearing tissue or body fluid ,it can help to make a diagnosis or rule out conditions such as cancer)
5) In Diagnosis (zero(0) represents the presence of malignant tumor and one(1) represents that the particular tumor is balanced)
6) Radius, Texture…(all properties of cells)
7) Predict the cancer type using the input data.

Scope for development of this project

Collaborative Effort- This app is developed through the facilitated under the supervision of dedicated students.

- Access/ Search information.
- Login to the system through the first page of the application
- View every paragraph, there are many options.
- Can contact us for any doubts or faults.
- Students can give feedback on college/staff/any other student.

Need for Proposed System

The main objective of the existing system is to provide a user friendly interface. The system, which is proposed, now computerizes all the details that are maintained manually. Once the details are fed into the computer there is no need for various persons to deal with separate sections. The security can also be given as per the requirement of the user.

➢Maintenance of this app is flexible.

➢Has varieties.

➢can easily access any point.

➢Accurate information are made.

➢Less manpower required.

General description User Characteristics: The target audience for CMS product is the college students/staff (Technical/Non technical) .The users for this system are Product Perspective: The product will be a standalone application and may be run on multiple systems within an Internet network. The product will require a keyboard, mouse and monitor or a mobile phone to interface with the users. The minimum hardware requirements for the product are specified in this document.

Overview of Functional Requirements:

The client requires the following features:

- The medical students, or non medical students can access this.
- A mechanism to uniquely identify each data.
- The user can view every point of the data.
- The system has no login.
- The system have help feature.

Feasibility Study

Feasibility study is a step towards identification of the system as a feasible product. First the studies often pre-suppose that when the feasibility document is being prepared, the analyst is able to evaluate solutions. Second, most studies tend to overlook the confusion inherent in system development – the constraints and the assumed attitudes.

If the feasibility study is to serve as a decision document, it must answer three questions:

- Is there a new and better way to do a job that will benefit the user?
- What are the costs and savings of the alternatives?
- What is recommended?

Feasibility Considerations:

There are three key considerations to the feasibility study:

1) Economic
2) Technical
3) Behavioural

Feasibility analysis of the system:

Economic Feasibility:

The project was considered to be economical feasible because-

- The cost involved in developing the system well under the budget of the organization.
- The cost-benefit analysis tells us that the value of the benefits offered by the proposed system is much higher than cost.

Technical Feasibility:

The proposed system is technically feasible because all the necessary hardware and software required for developing and installing the system is available with the organization.

Behavioural Feasibility:

The proposed system is also behaviourally feasible as it is very user friendly. Extensive training of the user is not required. The users can easily learn to use the system and can adapt themselves according to the system.

ANALYSIS:

Analysis is the detailed study of the various operations performed by a system and their relationships within and outside of the system. A key question is what must be done to solve the problems? One aspect of analysis are defining the boundaries of the system and determining whether or not a candidate system should consider other related system. During analysis, data are collected on the available files, decision points, and transactions handled by the present system. Some logical system models and tools that are used in analysis. DFD interviews, onsite observations, and questionnaires are examples. The interview is a commonly used tool in analysis. It requires special skills and sensitivity to the subjects being interviewed. Bias in data collection and interpretation can be a problem. Training, experience, common sense is required for collection of the information needed to do analysis.

Once analysis is completed, the analyst has firm understanding of what is to be done. The next step is to decide how the problem might be solved. Thus, in the system design, we move from the logical to the physical aspect of the life cycle.

System Specifications

Hardware Requirements: 1) A desktop with internet connection

2) CPU- Processor- AMD Ryzen 3 2200g

Ram- 8GB (3200 MHz)

Storage- 1TB HDD

Software Requirements: 1) Google colab

2) web browser Google Chrome

Project Phases:

Software Requirement Specification (SRS)

Introduction:

The introduction of the software requirement specification states the goals and objectives of the software, describing it in the context of the computer based system.

Information Description:

It provides detailed description of the problem that the software must solve. Information content, flow and structure and documented. Hardware, software and human interfaces are described for external system elements and internal functions.

Functional Description:

A processing narrative is provided for each function, design constraints are stated and justified, performance characteristics are stated, and one or more diagrams

are included to graphically represent the overall structure of the software and interplay among software and other system elements.

Behavioural Description:

This section of the SRS examines the operation of software as a consequences of external events and internally generated control characteristics.

Validation criteria:

It is probably the most important and ironically the most often neglected section of the Software Requirement Specification. The section is neglected because completing it demands a thorough understanding of the software requirements.

Specification of validation criteria acts as an implicit review of all other requirements, so it is essential that time and attention to this section.

Requirement Analysis

Requirement analysis is the first technical step in the process of any software development. A careful analysis can help the software designer and programmer to have a better insight of the product to be created. A careless analysis can result into incomplete or dysfunctional software. To avoid such situation it is very important to properly identify the required software's features and create an effective design for it. It is also important to analyse and find out whether the application being developed suits the current hardware and software platform

available or not. The application should be developed well within time and should meet the specified requirements. If the application is being developed for commercial purposes then a cost-benefit analysis becomes must to find out the real value of the software product.

Saurav Karmakar and Ishika Acharjee are handling this phase.

DESIGN

The most creative and challenging phase of the system life cycle is the system design. The term design describes a final system and the process by which it is developed. It refers to the technical specifications that will be applied in implementing the candidate system. It also includes the construction of programs and program testing. The key question here is: How should the problem be solved? The major steps in design are:

The first step is to determine how the output is to be produced and l what format. Sample of the output (are input) are also presented. Second, input data and master files (data base) have to be design to meet the requirements of the proposed outputs. The operational phases are handled through program construction and testing, including a list of the programs needed to meet the system and an estimate of the impact of the candidate system on the user and the organization are documented and evaluated by the management as a step towards implementation.

The final report prior to the implementation phase includes procedural flowcharts, record layouts, report layouts, and workable plans for implementing the candidate system. Information on personal, money, h/w, facilities and their estimated cost

must also be available. At this point, projected costs must be close to actual cost of implementation.

In some firms, separate group of programs do the programming where as other firms employ analyst-programmers that do the analysis and design as well as code programs. For this discussion, we assume that two separate persons carry out analysis and programming. There are certain functions, though, that the analyst must perform while programs are being written.

Sourav Tamang and Sajit Tamang and Monali Chaki are the in charge of this phase.

IMPLEMENTATION

The implementation phase is less creative than system design. It is primarily concentrated with user training, site preparation, and file conversion. When the candidate system is linked to terminals to remote sites, the telecommunication network and test of the network along with the system are also included under implementation.

During the final testing user acceptance is tested followed by the user training. Depending on the nature of the system, extensive user training may be required. Conversation usually takes place at about the same time the user is being trained or later.

In the extreme, the programmer is falsely viewed as some who ought to be isolated from other aspects of system development. Programming is itself design work, however. The initial parameters of the candidates of the system should be modified as a result of programming efforts. Programming provides a “reality test” for the assumption made by the analyst it is therefore a mistake to exclude programmers from the initial system design.

System testing checks the readiness and accuracy of the system to access update and retrieve data from new files. Once the program becomes available test data are read into the computer and processed against the file provide for testing in most conversation a parallel run is conducted where the new system runs simultaneously with the old system this method through costly provides added assurance against errors in the candidate system.

POST-IMPLEMENTATION AND MAINTANACE:

After the installation phase is completed and the user staff is adjusted to the changes created by the candidate system evaluation and maintenance begin. Like any system there is an aging process that requires periodic maintenance of hardware and software if the new information is inconsistent with the design specification then changes to be made. Hardware also requires maintenance to keep in tune with design specification.

Monali Chaki and Sajit Tamang are part of this phase.

Testing

Software testing is a critical element of software quality assurance and represents the ultimate review of application, design, coding. The aim of the testing process is to identity all defects existing in a software product. Testing provides a practical way of reducing defect in a system and increasing the user's confidence in a developed system.

TESTING OBJECTIVE:

- Testing is a process of executing a program with the intent of finding of error.

- A good test case one of that has a highly probability of finding an as yet undiscovered error.
- A successful test is one that uncovers as yet undiscovered error. The objective is to design test cases that systemically uncover different Classes of error or do so with a minimum amount of time and effort. This product has two parts:

a) Planning: This involves writing and reviewing unit, integration, functional, validation and acceptance test plans.
b) Execution: This involves executing these test plans, measuring, collecting data and verifying if its meets the quality criteria. Data collected is used to make appropriate changes in the plans related to developments and testing. The quality of the product or item can be achieved by ensuring that the product meets the requirements by planning and conducting the following tests at various stages.

The main type of software testing are:

COMPONENTS:

Starting of the button the first level is "**Component Testing**", sometimes called unit testing. It involves checking that each specified in the "**Component Design**" has been implemented in the component. In theory an independent tester should do this, but in practice the developer usually does it, as they are the only people who understand how a components work. The problem with a component is that it performs only a small part of functionality of a system, and it relies on co-operative with the other part of the system, which may not have been built yet. To cover come this, the developer either builds, or uses special software to trick the component into believing it is working in a fully functional system. **In our project we perform three levels of testing.**

1. Unit testing
2. Integration testing
3. System testing **Unit Testing:**

In unit testing we tested different modules of in isolation.

To test a program we provide set of inputs to program and observe whether it behaves as expected in unit testing we user white-box approach to develop the test cases.

Integration Testing:

During integration testing we integrated both modules of system using integration plan. In integration testing we tested the module interfaces. For integration testing we tested the module interfaces. We use incremental approach to integrate the system module.

a. System Testing:

System testing is actually series of test whose primary purpose is to fully exercise the project. We performed three type of system testing

- Validation testing
- Recovery testing
- Security testing

- **Validation Testing:** The purpose of validation testing is to know whether requirements of project are fulfilled or not. For this we perform two type testing.

  - Alpha testing
  - Beta testing

- **Alpha Testing:**

For alpha testing we simulated the required environment within the organization and used the system.

Beta Testing:

For beta testing we launch the system temporarily any test it form outside the organization.

Recovery Testing:

In recovery testing we forced the project to fall in variety of ways and verify that recovery is performed properly. Here the recovery testing has been done from all the aspects including automatic reinstallation, check pointing mechanism, data recovery etc. All the above points are evaluated for correctness.

Security Testing:

Security testing attempts to verify that protection mechanism built into a system evil in fact protect it from improper penetration. Here we are testing that any unauthorized user could not change the information of website. The project has been thoroughly processed through security testing and given enough time and resource for testing the penetration in to the project.

Saurav Tamang tested every units.

Technical Description

2.1 Machine Learning

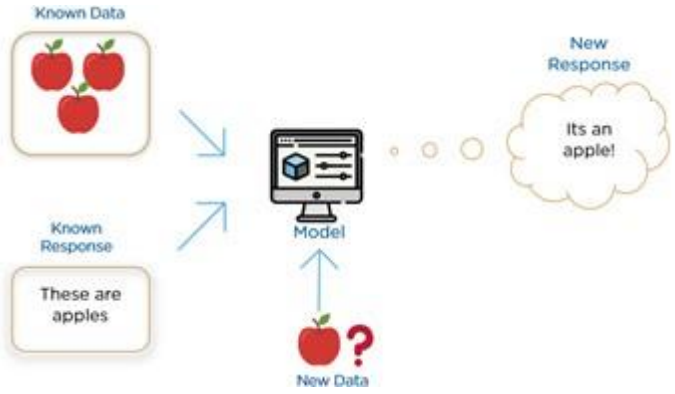
At a high-level, machine learning is simply the study of teaching a computer program or algorithm how to progressively improve upon a set task that it is given. On the research-side of things, machine learning can be viewed through the lens of theoretical and mathematical modeling of how this process works. However, more practically it is the study of how to build applications that exhibit this iterative improvement.

SUPERVISED LEARNING **UNSUPERVISED LEARNING** **REINFORCEMENT LEARNING**

Supervised Learning

Supervised learning is the most popular paradigm for machine learning. It is the easiest to understand and the simplest to implement. It is very similar to teaching a child with the use of flash cards.

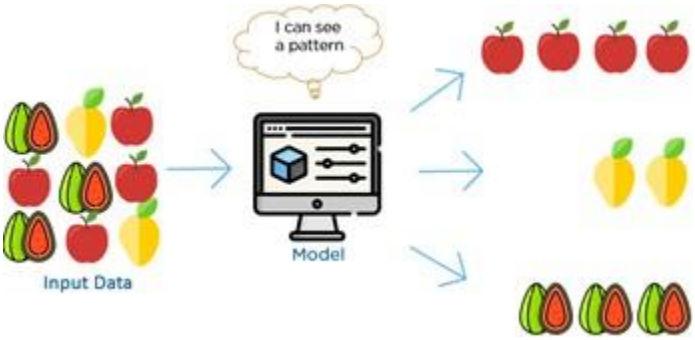
Given data in the form of examples with labels, we can feed a learning algorithm these example label pairs one by one, allowing the algorithm to predict the label for each example, and giving it feedback as to whether it predicted the right answer or not. Over time, the algorithm will learn to approximate the exact nature of the relationship between examples and their labels. When fully trained, the supervised learning algorithm will be able to observe a new, never-before-seen example and predict a good label for it.



Unsupervised Learning

Unsupervised learning is very much the opposite of supervised learning. It features no labels. Instead, our algorithm would be fed a lot of data and given the tools to understand the properties of the data. From there, it can learn to group, cluster, and/or organize the data in a way such that a human (or other intelligent algorithm) can come in and make sense of the newly organized data.

What makes unsupervised learning such an interesting area is that an overwhelming majority of data in this world is unlabeled. Having intelligent algorithms that can take our terabytes and terabytes of unlabeled data and make sense of it is a huge source of potential profit for many industries. That alone could help boost productivity in a number of fields.



Reinforcement Learning

Reinforcement learning is fairly different when compared to supervised and unsupervised learning. Where we can easily see the relationship between supervised and unsupervised (the presence or absence of labels), the relationship to reinforcement learning is a bit murkier.
Some people try to tie reinforcement learning closer to the two by describing it as a type of learning that relies on a time-dependent sequence of labels, however, my opinion is that that simply makes things more confusing.

I prefer to look at reinforcement learning as learning from mistakes. Place a reinforcement learning algorithm into any environment and it will make a lot of mistakes in the beginning. So long as we provide some sort of signal to the algorithm that associates good behaviors with a positive signal and bad behaviors with a negative one, we can reinforce our algorithm to prefer good behaviors over bad ones. Over time, our learning algorithm learns to make less mistakes than it used to.

2.2 Regression Analysis in Machine Learning

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as temperature, age, salary, price, etc.

Dependent Variable : The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called target variable.

Independent Variable : The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variables, also called as a predictor.

Why do we use Regression Analysis?

As mentioned above, Regression analysis helps in the prediction of a continuous variable. There are various scenarios in the real world where we need some future predictions such as weather conditions, sales prediction, marketing trends, etc., for such a case we need some technology which can make predictions more accurately. So for such a case we need Regression analysis which is a statistical method and used in machine learning and data science.

Below are some other reasons for using Regression analysis:

- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data.
- It helps to predict real/continuous values.
- By performing the regression, we can confidently determine the most important factor, the least important factor, and how each factor is affecting the other factors.

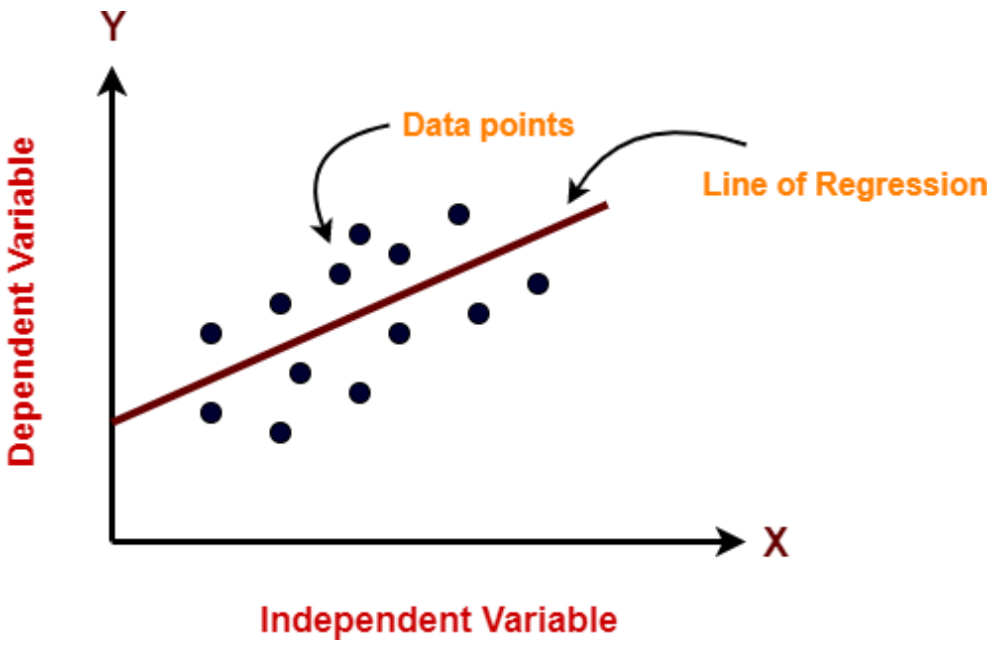
2.3 Linear Regression

- Linear regression is a statistical regression method which is used for predictive analysis.
- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
- If there is only one input variable (x), then such linear regression is called simple linear regression. And if there is more than one input variable, then such linear regression is called multiple linear regression.
- The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of the year of experience.

Below is the mathematical equation for Linear regression:

$Y= aX+b$

Here, Y = dependent variables (target variables), X= Independent variables (predictor variables), a and b are the linear coefficients.



2.4 Logistic Regression

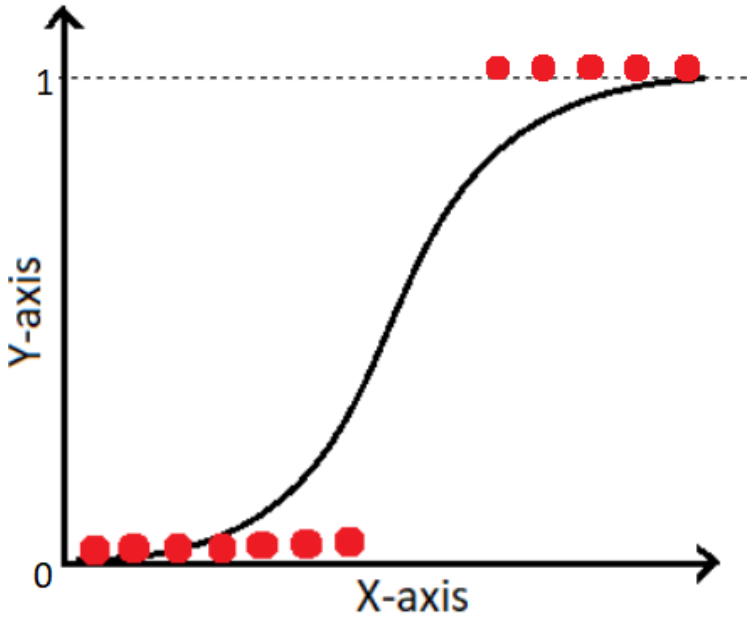
- Logistic regression is another supervised learning algorithm which is used to solve the classification problems. In classification problems, we have dependent variables in a binary or discrete format such as 0 or 1.
- Logistic regression algorithm works with the categorical variable such as 0 or 1, Yes or No, True or False, Spam or not spam, etc.
- It is a predictive analysis algorithm which works on the concept of probability.
- Logistic regression is a type of regression, but it is different from the linear regression algorithm in the term how they are used.
- Logistic regression uses sigmoid function or logistic function which is a complex cost function. This sigmoid function is used to model the data in logistic regression.

The function can be represented as:

f(x)=1/(1+e^-x)

f(x)= Output between the 0 and 1 value, x= input to the function and e= base of natural logarithm.

When we provide the input values (data) to the function, it gives the S-curve as follows:

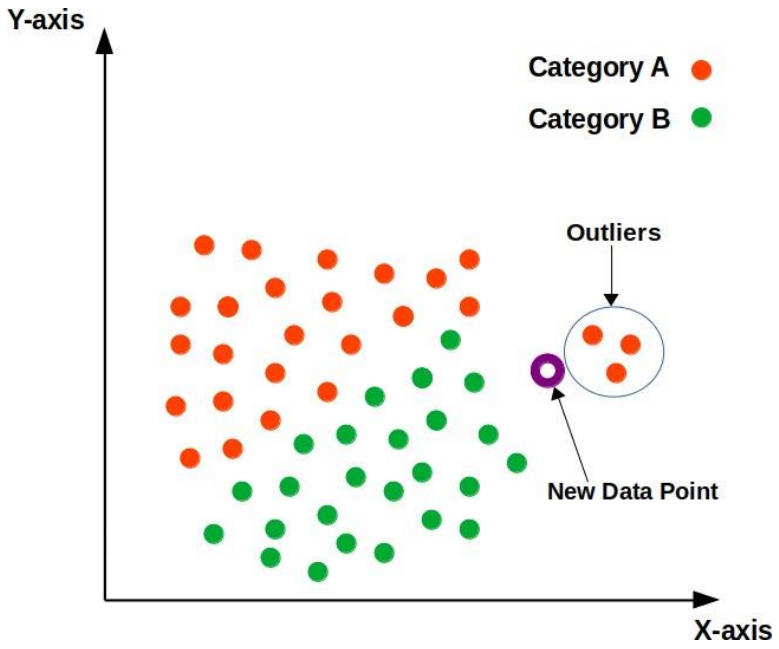


It uses the concept of threshold levels, values above the threshold level are rounded up to 1, and values below the threshold level are rounded up to 0.

2.5 K-Nearest Neighbor (KNN) Algorithm for Machine Learning

- K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

The KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.



2.6 K-Means Algorithm

K-means clustering algorithm computes the centroids and iterates until it finds the optimal centroid. It assumes that the number of clusters are already known. It is also called a flat clustering algorithm. The number of clusters identified from data by algorithm is represented by 'K' in K-means.

In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum. It is to be understood that less variation within the clusters will lead to more similar data points within the same cluster.

2.7 The Elbow Method

For the k-means clustering method, the most common approach for answering this question is the so-called elbow method. It involves running the algorithm multiple times over a loop, with an increasing number of cluster choices and then plotting a clustering score as a function of the number of clusters.

The elbow method finds the optimal value for k (clusters).