**“Machine Learning Based Approaches for Cancer Prediction”**

A

*Thesis Synopsis submitted in partial fulfillment of the Award of the degree of*

MASTER OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

*By*

**CHIRANJIB PARIDA**

Roll No – **412VMIT20003**

Batch-(2020-22)

*Under the guidance of*

**Dr. NARAYAN NAYAK**

Associate Professor,

Electronics & Instrumentation Engineering

Silicon Institute of Technology, Bhubaneswar, Odisha 751024, India.

**CERTIFICATE**

This is to certify that the Thesis Synopsis entitled **“**Machine Learning Based Approaches for Cancer Prediction**”** submitted by **CHIRANJIB PARIDA**, Roll No. - **412VMIT20003**, in partial fulfillment of the requirements for the Master of Technology in INFORMATION TECHNOLOGY is a bonafide piece of work carried out by her under my supervision and guidance.

**To the best of my knowledge, this work has not been submitted to any other University or Institution for the award of any degree. In my opinion, the thesis fulfils the requirement of the regulations relating to the nature and standard of work for master of technology.**

**Dr. Narayan Nayak**

**Associate Professor,**

**Electronics & Instrumentation Engineering**

**Silicon Institute of Technology, Bhubaneswar, Odisha 751024, India.**

**CERTIFICATE OF APPROVAL**

This is to certify that we have examined the thesis synopsis entitled **“**Machine Learning Based Approaches for Cancer Prediction**”**submitted by **CHIRANJIB PARIDA** having **Roll No. 412VMIT20003** & we hereby accord our approval of the thesis work carried out and presented in a manner required for its acceptance for the partial fulfillment of the Degree of Master of Technology in Information Technology. It is understood that, this approval do not necessarily endorse or accept every statement made, opinion expressed or conclusions drawn as recorded in this thesis. It only signifies the acceptance of the thesis for the purpose for which it has been submitted.

***(External Examiner)***

**DECLARATION**

I hereby declare that the subject matter of this Thesis Synopsis entitled **“**Machine Learning Based Approaches for Cancer Prediction**”** is the record of work done by me. That to the best of my knowledge & belief the content of this thesis has not been submitted in any institution(s) for the sward of any previous higher degree. That the work done by me is free from plagiarism. This is being submitted to Utkal University for which it is required.

**CHIRANJIB PARIDA**

**Roll No-412VMIT20003**

**ABSTRACT**

The rapid expansion of data and databases has necessitated the development of new tools and procedures that can convert raw data into useable information rapidly and effectively. The advancement of information technology has prompted simultaneous growth in database storage and management capabilities in recent years. The need for retrieving the implicit information included in recorded data to improve decision-making in business, healthcare services, research, and other areas grows greater. As a result, obtaining meaningful information from data kept in vast repositories (also known as "knowledge discovery") is seen as a fundamental need in many fields. The subject field is known as “Machine Learning “has been a major issue in databases and Artificial Intelligence (AI). Cancer is one of the most serious problems facing society today, and identifying cancer at an early stage remains a difficult task for specialists. In order to create innovative ways for detecting and preventing cancer, it is critical to first identify and understand genetic and environmental variables. It is for this reason that a revolutionary multi-layered strategy that in corporate clustering and decision tree techniques is used in the development of a cancer risk prediction system. The suggested system accurately predicts lung, breast, oral, cervical, stomach, and blood cancers, and it is both user-friendly and cost-saving in terms of operation and maintenance. To identify possible cancer patients, this study uses data mining methods such as classification, clustering, and prediction to analyze large amounts of data. This cancer prediction method, which is based on data mining techniques, has been suggested by us. This approach predicts the likelihood of developing breast cancer at an early stage. The projected outcomes of this method are compared to the patient's past medical information, which serves as validation. The primary goal of this strategy is to offer consumers an early warning while still being cost effective for the user. Finally, a prediction system is established to examine risk levels, which aids in the development of a prognosis for the patient. This study aids in the discovery of a person's propensity to cancer prior to the administration of clinical and laboratory testing, which are both expensive and time-consuming.

Keywords: Artificial Intelligence (AI),Machine Learning, Cancer, Clustering, Classification.

**Contents**

**Title Page**

ACKNOWLEDGEMENTS II

CERTIFICATE III

DECLARATION IV

ABSTRACT V

ABBREVIATIONS

**CHAPTER 1: INTRODUCTION…………………………………………….[1 – 3]**

**CHAPTER 2: RELATED WORK ……………………………………………. [3 – 4]**

**CHAPTER 3: DATASET INFORMATION...........………………………. [4 – 8]**

**3.1 Data Preprocessing……………………………………………. [4]**

**3.2 Exploratory Data Analysis…………………………………… [5 - 7]**

**3.3 Feature Extraction……………………………………………… [8]**

**CHAPTER 4: OUR PROPOSED SYSTEM…………………………………. [8 – 10]**

**CHAPTER 5: PERFORMANCE EVALUATION & RESULTS…………... [11 – 15]**

**5.1 Confusion Matrix ……………………………………………………. [12]**

**5.2 Classification Report…………………………………………………[14]**

**CHAPTER 6: CONCLUSION and future scope……………………….........[16 – 17]**

**6.1 Future scope……………………………………………………………. [16]**

**CHAPTER 7: REFERENCES……………………………………………………….. [17]**

**LIST OF FIGURES**

**[2]**

**[4]**

**[5]**

**[5]**

**[6]**

**[6]**

**[6]**

**[9]**

**[11]**

**[22]**

* 1. **figure 1.1 ………………………………………………………………….**
  2. **figure 3.1………………………………………………………..**
  3. **figure 3.2………………………………………………………..**
  4. **figure 3.3………………………………………………………..**
  5. **figure 3.4………………………..................................................**
  6. **figure 3.5………………………………………………………..**
  7. **figure 3.6………………………………………………………..**
  8. **figure 4.1……………………………..…………………………**
  9. **figure 5.1………………………………………………………..**
  10. **figure 5.2…………………..…….. ………………… …………**

**CERTIFICATE OF PLAGIARISM**

CHIRANJIB PARIDA\_ 412VMIT20003\_SeminarReport\_l\_v1

ORIGINALITY REPORT

26%

SIMILARITY INDEX

19%

INTERNET SOURCES

32%

PUBLICATIONS

16%

STUDENTPAPERs

PRIMARY SOURCES

1 Mamatha Sai Yarabarla, Lakshmi Kavya Ravi, Dr. A. Sivasangari. " Breast Cancer Prediction via Machine Learning", Iranian Journal of Science and Technology, Transactions of Electrical Engineering

4%

Internet Source

3%

2 Syed Saba Raoof,Syed Aley Fathima, M A. Jabbar, " Lung Cancer Prediction using Machine Learning: A Comprehensive Approach", IEEE Transactions on Neural Networks and Learning Systems, 2021

Publication

3

medium.com

Internet Source %

3

1

diposit.ub.edu

Internet Source %

4

2

link.springer.com

Internet Source %

5

6

7

8

9

10

11

12

13

14

[www.coursehero.com](http://www.coursehero.com/)

Internet Source %

1

biomedpharmajournal.org

1

Internet Source %

Siddhant Thakur Thakur, Devarshi Patel Patel. "Patient Health Monitoring and Inferencing Arrhythmia Using ECG\_Data", Research Square Platform LLC, 2021

1%

Publication

datascience.sharerecipe.net

1

Internet Source %

Lina Chato, Shahram Latiﬁ. "Machine Learning and Deep Learning Techniques to Predict Overall Survival of Brain Tumor Patients using MRI Images", 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), 2017

1%

Publication

almazrestaurant.com

1

Internet Source %

paperswithcode.com

1

Internet Source %

[www.thinkautomation.com](http://www.thinkautomation.com/)

1

Internet Source %

mullikine.github.io

1

Internet Source %

15

16

17

18

19

20

21

22

Submitted to University of Portsmouth

Student Paper %

1

<1

askinglot.com

Internet Source %

<1%

Submitted to Liverpool John Moores University

Student Paper

<1

mnamd.com

Internet Source %

<1

bildenpoliziotto.com

Internet Source %

<1

[www.simplilearn.com](http://www.simplilearn.com/)

Internet Source %

<1%

R. Lakshmi Devi, Vijaya Samundeeswari V. "Detection and Automated Classiﬁcation of Brain Tumor Types in MRI Images using Convolutional Neural Network with Grid Search Optimization", 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2021

Publication

Ying Bi, Bing Xue, Mengjie Zhang. "An automated ensemble learning framework using genetic programming for image classiﬁcation", Proceedings of the Genetic and Evolutionary Computation Conference, 2019

23

24

25

26

27

28

29

Publication

huggingface.co

Internet Source

Kanishk Shah, Khushali Deulkar. "Lightweight Apparel Classiﬁcation with Residual and Inverted Residual Block based Architectures", 2021 IEEE Cloud Summit (Cloud Summit), 2021

Publication

doctorpenguin.com

Internet Source

ajse.aiub.edu

Internet Source

"Engineering Applications of Neural Networks", Springer Science and Business Media LLC, 2019

Publication

Karthika Veeramani, Suresh Jaganathan. "Intensiﬁed regularized discriminant analysis technique", International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), 2014

Publication

Submitted to Jaypee University of Information Technology

Student Paper

<1%

<1%

<1%

<1%

<1%

<1%

<1%

**Chapter – 1**

**Introduction**

Cancer is a disease in which certain cells in the body grow out of control and spread to other parts of the body. Cancer can occur almost anywhere in the human body, which is made up of trillions of cells. Normally, human cells grow and multiply (by a process called cell division) to form new cells as the body needs them. When the cells get old or damaged, they die, and new cells take their place. Sometimes this order of things breaks down, and abnormal or damaged cells grow and proliferate where they should not. These cells can form tumors, which are tissues. Abscesses may or may not be cancerous (benign). Cancer cells differ from normal cells in many ways. For example, cancer cells: they grow without signs to tell them to grow. Normal cells only grow when they receive such symptoms. Ignore the symptoms that most commonly tell cells to stop dividing or dying (a process known as programmed cell death, or apoptosis). Invade nearby areas and spread to other parts of the body. Normal cells stop growing when they come in contact with other cells, and most normal cells do not move in the body.

* 1. ***Objective:***  The objective of this paper is to create a Machine learning model that will predict whether the patient having cancer since all are cancer patients but this model will predict the level of cancer i.e. it is in initial or low stage, middle stage or high stage based on all the input features like 'Age', 'Gender', 'Air Pollution', 'Alcohol use', 'Dust Allergy', 'Occupational Hazards', 'Genetic Risk', 'chronic Lung Disease', 'Balanced Diet', 'Obesity', 'Smoking', 'Passive Smoker', 'Chest Pain', 'Coughing of Blood', 'Fatigue', 'Weight Loss', 'Shortness of Breath', 'Wheezing', 'Swallowing Difficulty', 'Clubbing of Finger Nails', 'Frequent Cold', 'Dry Cough' and 'Snoring'.

**Chapter -2**

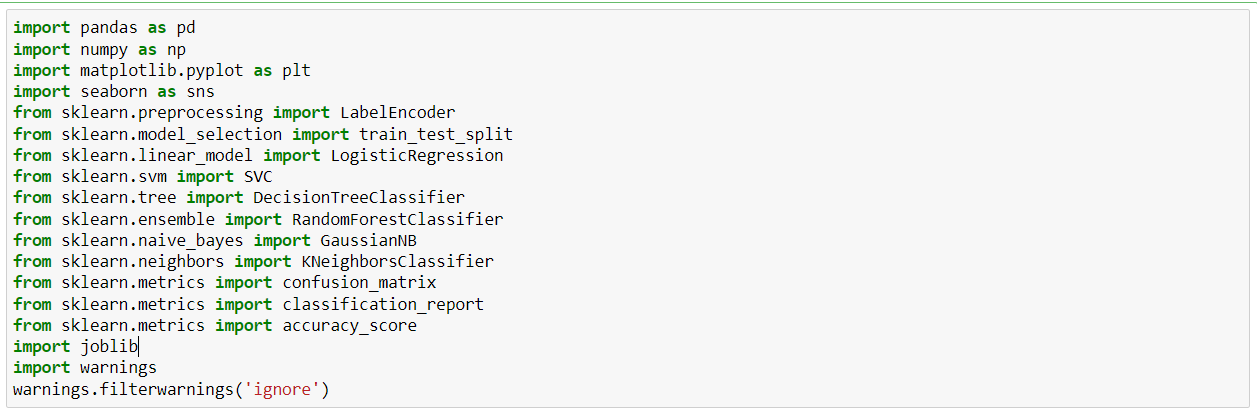
**Related Work**

Scientists are trying to minimize the effects of the deadly disease, and there are many questions about the survival of the fittest. Moataz M. Abdelwahab, Shimaa A. Abdelrahman predicted that cancer is among the most common genetic diseases that kill thousands of people each year. It grows on the human body due to genetic damage caused by a number of factors such as smoking, alcohol uses, Air Pollutions which activates on cases and binds the tumor suppressor to normal lung cells and produces genetic mutations that lead to genes. This creates an important need for robust, rapid, and sensitive diagnostic testing. Therefore, it has become increasingly important to develop a new approach based on genetic modification to predict cancer to win the battle against cancer.

**Chapter – 3**

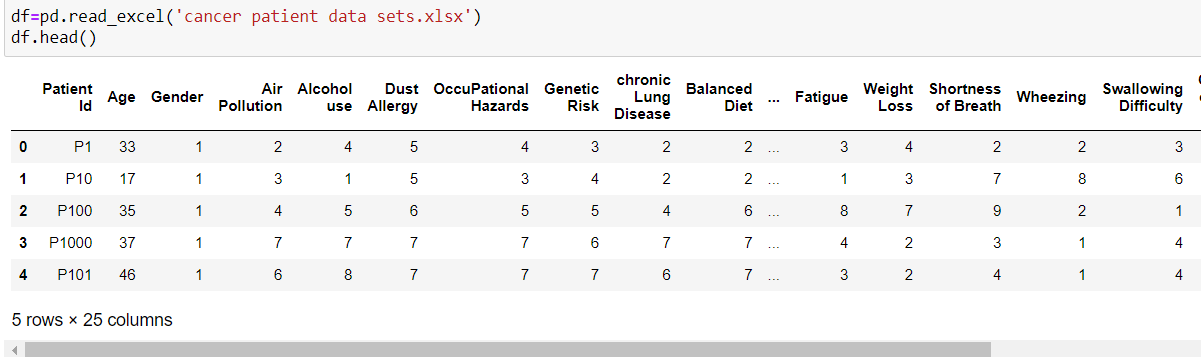
**Data Preprocessing**

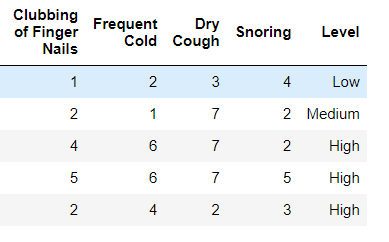
1. **Dataset information:** The Dataset is collected from Kaggle named as [*cancer\_patient\_data\_sets.xlsx*](https://data.world/cancerdatahp/lung-cancer-data/file/cancer%20patient%20data%20sets.xlsx)*.* It is also available in *data.world* site. This dataset is publicly available. The dataset contains 1000 (a thousand) rows and 25 (twenty five) columns. The features in the dataset are: *Patient Id, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, chronic, Lung, Disease, Balanced, Diet, Obesity, Smoking, Passive, Smoker, Chest, Pain, Coughing of Blood, Fatigue, Weight Loss, Shortness of Breath, Wheezing, Swallowing Difficulty, Clubbing of Finger Nails, Frequent Cold, Dry Cough, Snoring and Level*. Each row of the dataset contains all the above 25 information of a single patient.
   1. **Data Preprocessing:** Data preprocessing the first step of machine learning after the data collection. In Data preprocessing Data must be balanced, clean i.e. to remove duplicate values or fill the null values or remove the null values. IN this session all the records are in int64 format i.e. integer value and the only column ‘Patient Id’ and ‘Level’ are the object type (String value). So for model creation, Patient Id is not important feature, which means the patient having cancer or not it would not depend on the patients Id, so Patient Id would be removed. This dataset does not have any null value or duplicate value in the records of 1000 patients. Here ‘Level’ is the outcome or we can say it is the target of the dataset which is in object type or string format. As model could not understand the string values so it is needed to be converted in to integer or float for that LabelEncoder() is used, this function comes from preprocessing of sklearn library. After using this technique ‘Level’ columns values will be converted from [‘high’, ’medium’, ’low’] to [0, 1, 2]. When the data preprocessing comes for that at first we need to import several libraries. There are some important libraries we need to first import for the data preprocessing. Pandas, NumPy, matplotlib, seaborn, and Scikit-learn are the most important libraries that we need to import first before the data preprocessing.



**[Fig – 3.1.]**

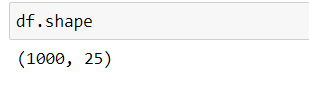
The libraries which are imported are given on above figure. Initially the head of the represented in Figure 3.2.

****

****

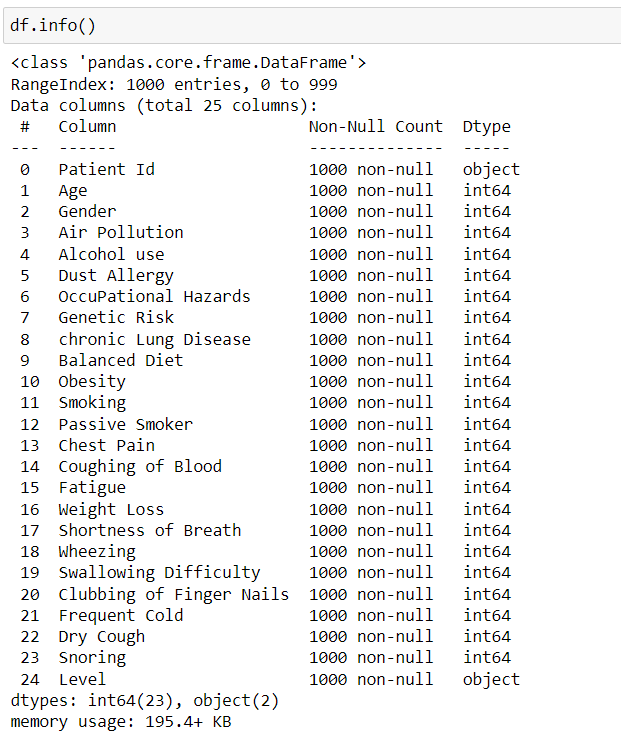
**[Fig - 3.2]**

1. The shape of the Data Frame is: (1000, 25) which indicates the dataset having 1000 number of rows and 25 number of columns it is represented in Figure 3.3.



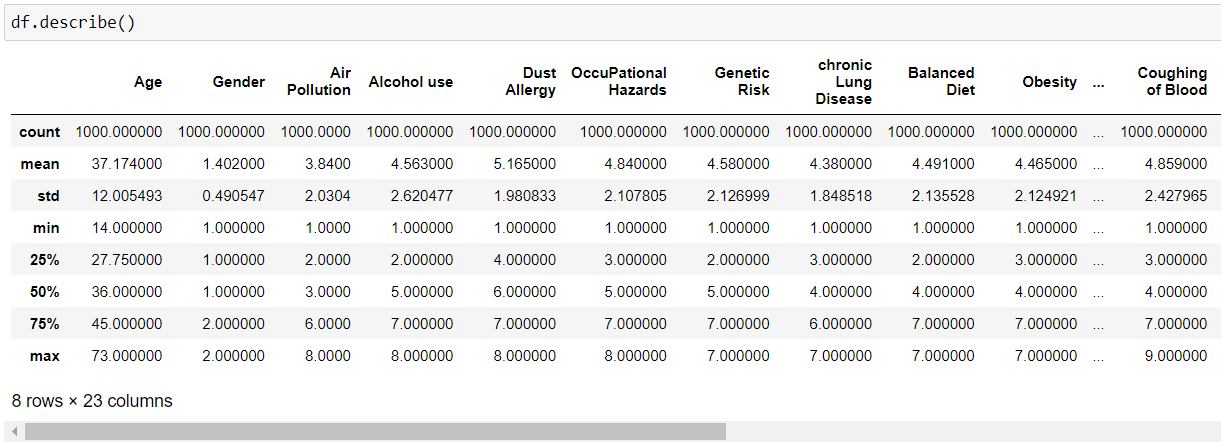
**[Fig – 3.3.]**

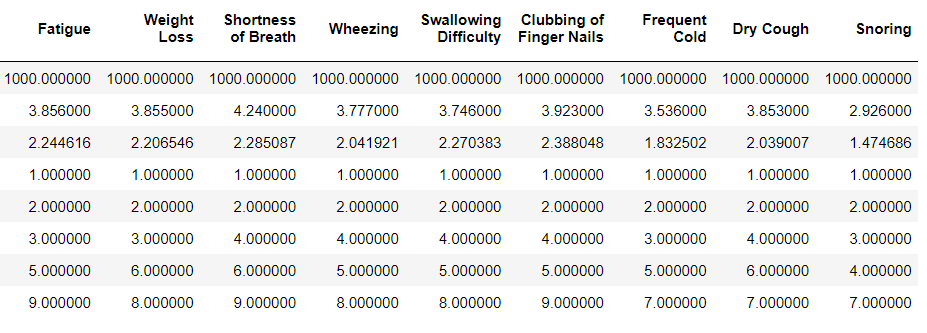
1. Some General Dataset information is given in the below figure 3.4. It describes the class, range index, data columns with non null content and Types. It describes the **patient Id** and the **Level** are **object** type of data where all other features are in **int 64** format. There are total 1000 rows and 25 columns in the dataset.



**[Fig – 3.4]**

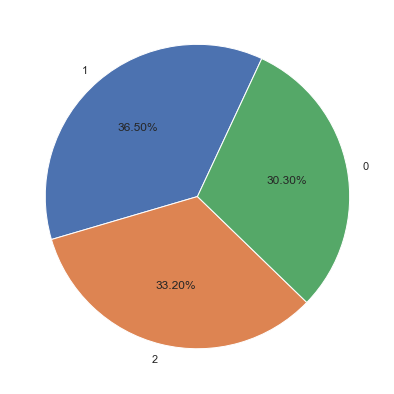
1. Some statistical information about the dataset is given in below figure 3.5. It represents the count means number of rows of each features, mean of each features, std means standard deviation of each features. Min represents the minimum value, 25% represents the 25% percent of the features it is called as Q1 and 50% represents the median of the features, 75% is Q3 represents the 75% value of each features and max represents the maximum value or 100% of the features.





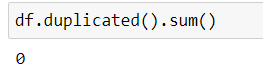
**[Fig – 3.5]**

* 1. **. Exploratory Data Analysis (EDA):** The below figure represents how different features are responsible for cancer disease. The below figure represents how much percentage of patients are from which stage among 100 patients, The Figure indicates [fig – 3.1] 36.50% of patients are in low stage of cancer and 33.20% patients are in middle stage of cancer and the rest 30.30% of patients are highly affected in cancer.

****

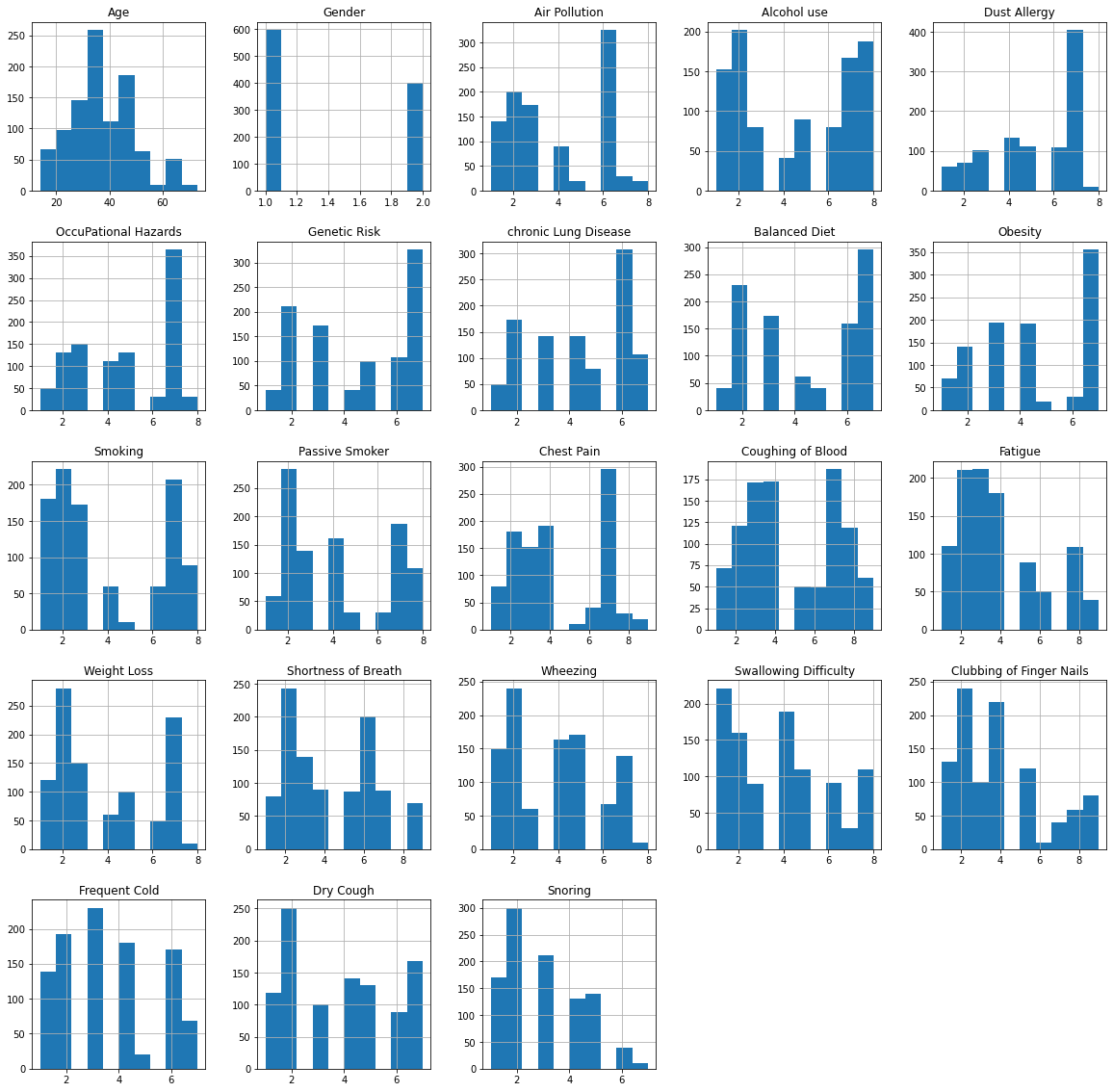
**[Fig -3.6.]**

* Checking for duplicated values in the dataset id present or not using duplicated () method.



**[Fig – 3.2.1.]**

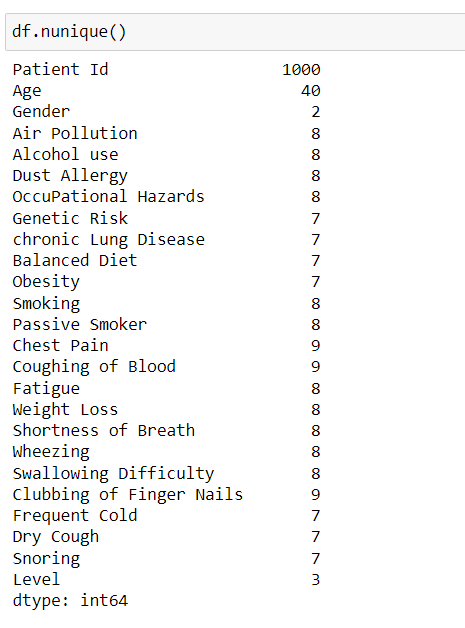
There is no duplicated values inside the dataset as you can see in above figure 3.2.1.



**[Fig – 3.2.2.]**

The Histogram of All the features in the dataset describes the amount of all the features. For cancer disease For example in [fig – 3.2.2] the second histogram indicates that among 1000 patients 600 cancer patients are Male and 400 patients are Female. Where the Air Pollution is in 6 the number of cancer patients are maximum i.e somewhere near 350.Also it can be noticed if the dust allergy is in somewhere 7 the number of cancer patients are maximum i.e greater than 400.

* The unique values present in each feature is given in below figure 3.2.3.



**[Fig – 3.2.3.]**

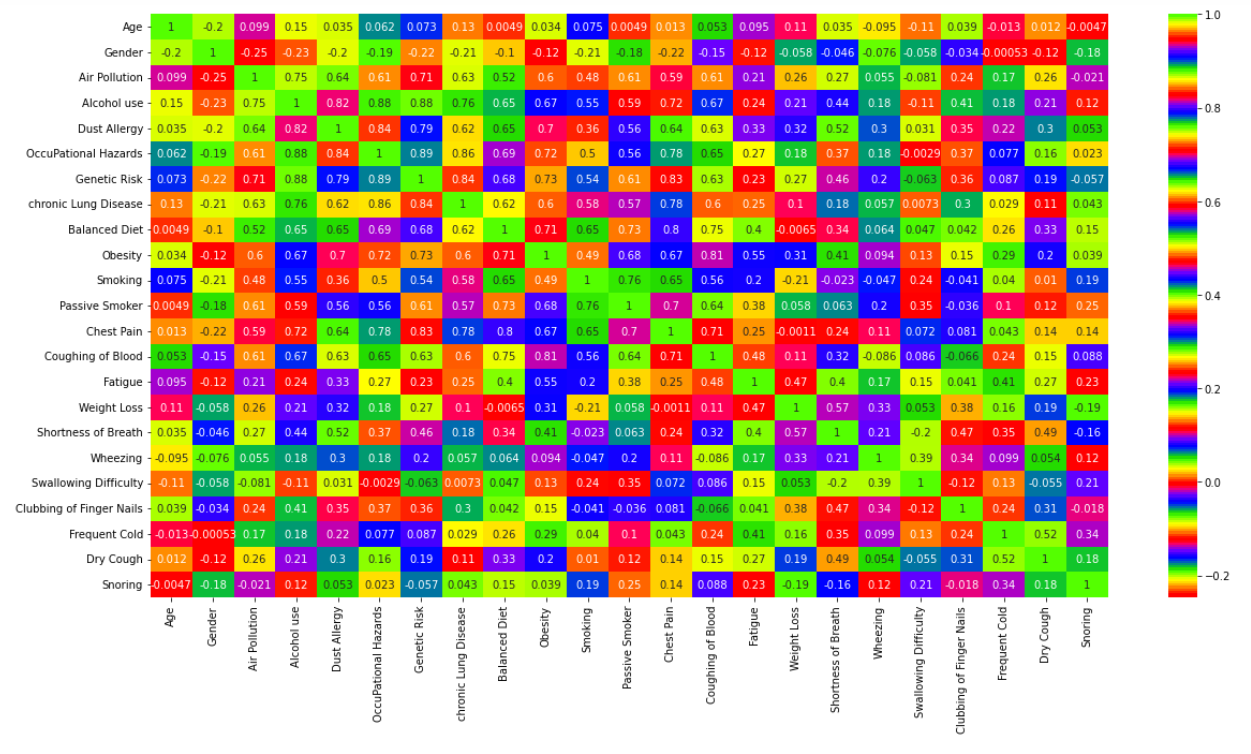
Patient Id contains one thousand features i.e., 1000 entries of patient’s data Age contains forty different values, Gender having two Unique values that seems to be male and Female. Similarly, ‘Air Pollution’ contains eight unique values. ‘Alcohol use’ contains eight unique values. ‘Dust Allergy’ contains eight unique values, ‘Occupational hazards’ having 8 different values, ‘Genetic Risk’ Contains seven unique values. ‘Chronic Lung Disease’ contains seven unique values, ‘Balanced Diet’ having seven different values. ‘Obesity’ contains seven different values. ‘Smoking’ having eight unique values ‘passive Smoker’ having eight unique values. ‘Chest pain’ having nine unique values. ‘Coughing of Blood’ also contains nine unique values. ‘Fatigue’ having eight unique values, ‘Weight Loss’ having eight different values, ‘Shortness of Breath’ having eight different unique values. ‘Wheezing’ contains eight different values. ’Swallowing Difficulty’ having 8 different values. ‘Clubbing of Finger Nails’ have nine unique values, ‘Frequent Cold’ having seven different values, ‘Dry Cough’ have seven different values. ‘Snoring’ also contains seven unique values, and finally the ‘Level’ contains three unique values.

* As we see the ‘Patient ID’ feature is not important for build or train a machine learning model because that only contains the patients number so we must delete that column.



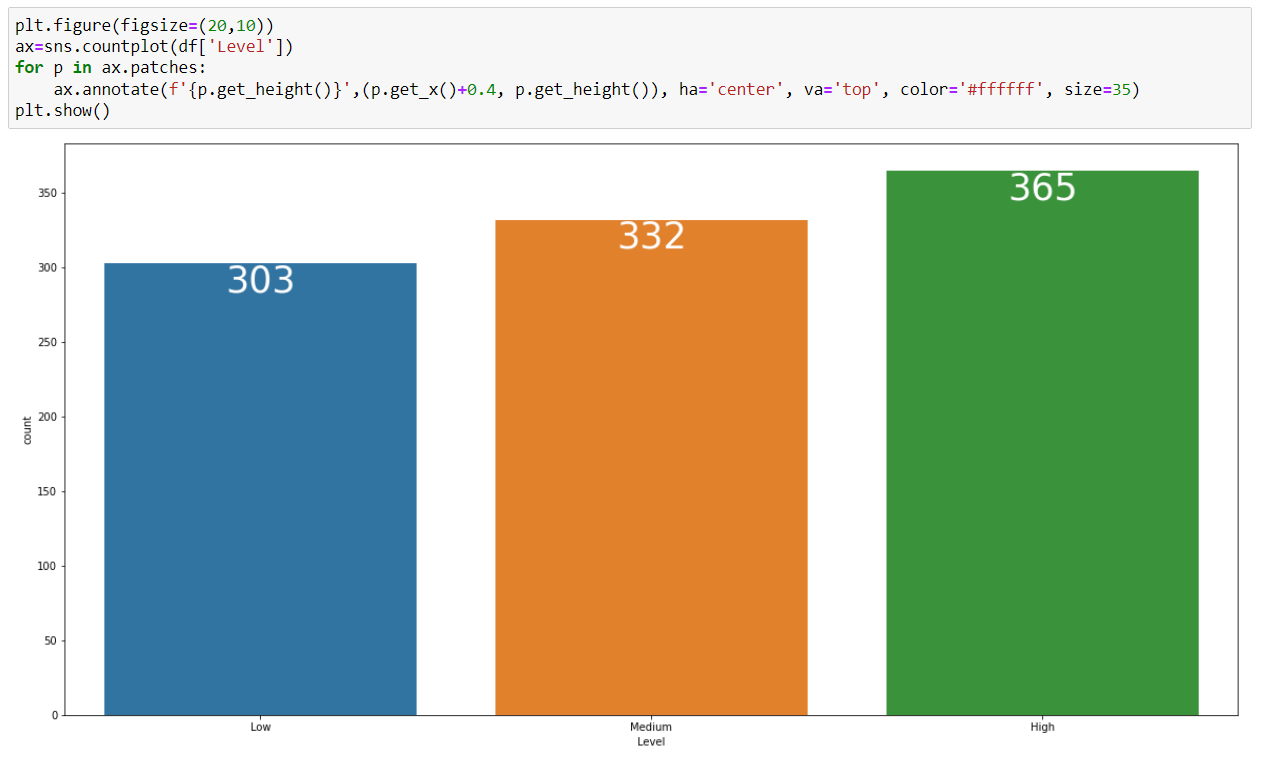
**[Fig – 3.2.4.]**

Here is the correlation heat-map at [fig – 3.2.5]. Correlation are generally 3 types (i) positively correlation (ii) Negative correlation and (iii) No correlation. If one feature of the data set is increasing then another feature which will increase, we can say that these two features are positively correlated to each other. If by increasing one feature another feature is decreasing then, these features are negatively correlated to each other. If by increasing or decreasing one feature it does not take any effect on any other feature it will be called these features having no correlation to each other. In above figure for example: [Air Pollution, Dust Allergy, Genetic Risk, chronic Lung Disease, Passive Smoker, Coughing of Blood] are positively correlated to Alcohol use since there correlation values are greater than 0.5, Where all other features those having value lower than 0.5 are negatively correlated to the Alcohol use. Swallowing Difficulty, Snoring and Gender are giving –ve value in correlation to the ‘Air Pollution’. I.e. Swallowing Difficulty, Snoring and Gender are negatively correlated to the ‘Air Pollution’. Weight Loss, Shortness of Breath, Wheezing, Clubbing of Finger Nails and Gender are negatively correlated to the 'Smoking'. While Passive Smoker, Chest Pain, Coughing of Blood are positively correlated to the Smoking.

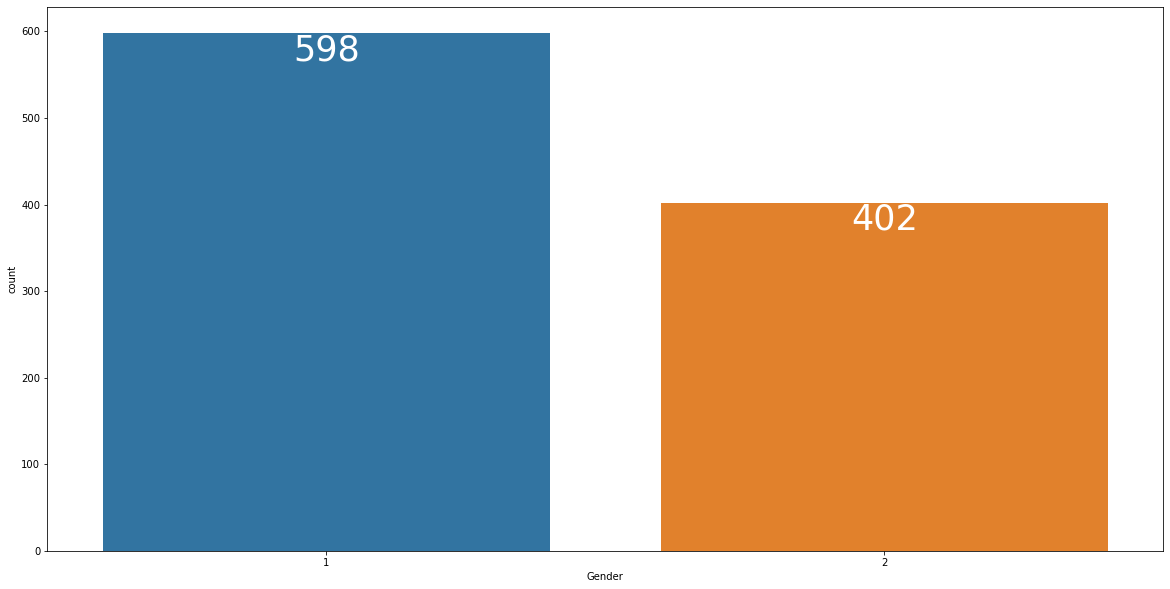


**[Fig – 3.2.5]**

The below figure [Fig – 3.2.6] shows how many patients are in which level/stage of cancer among the 1000 patients’ maximum patients are in high-level i.e., greater than 350. minimum number of patients are in middle stage of cancer.

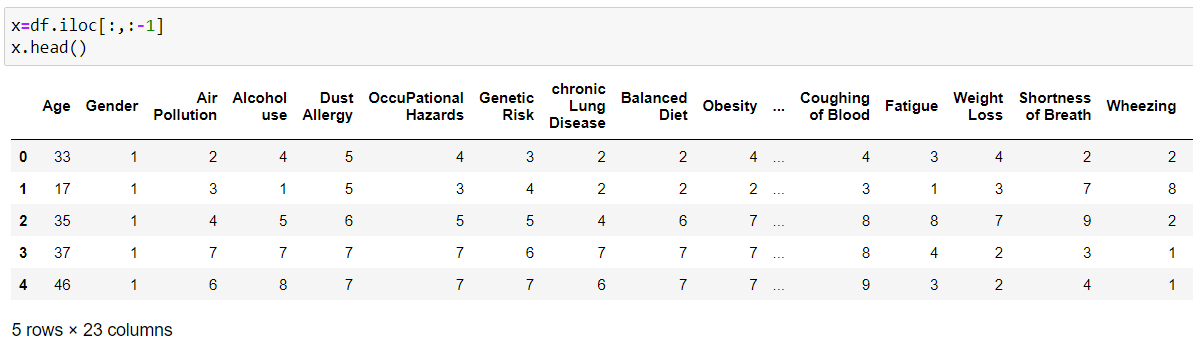


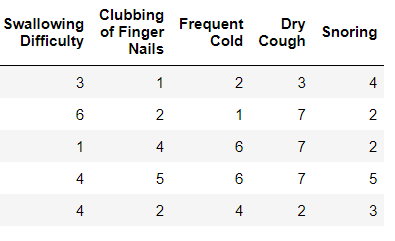
**[Fig – 3.2.6.]**

* How many Patients belongs to which Gender is given in below figure 3.2.7.

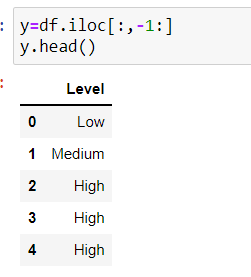
As you can see among 1000 patients 598 are male and 402 patients are Female. In dataset male represented as 1 and Female represented as 0.

* 1. **Feature Extraction:**In this section dataset we have to extract important features by analyzing that features machine learning model can predict the condition of patients i.e. the patient is in which stage of cancer. Here in this paper the ‘Level’ column is the outcome or the target i.e. dependent feature of the dataset. And rest all 23 columns are be input to the model so these are the independent feature of the dataset. In this paper X is the independent variable that contains all the independent features and the y is the dependent variable which contains the dependent feature/outcome/target i.e. the ‘Level’ column. The dataset must be divided into 2 parts that is training part and the testing part for which there is a function inside the model\_selection module of the sklearn library called train\_test\_split() which returns 4 variable named as X\_train (All the training dependent features), X\_test (Alll the testing dependent features), y\_train (Outcomes of the training data i.eX\_train) and y\_test (Outcomes of the testing data i.e (X\_train). here test\_size is assigned as 0.25 i.e 25% of data will be treated as the testing data (250 records) and rest 75% of data will be treated as training data (750 records).In this paper we have taken Random State as 42 for training the model with better accuracy.



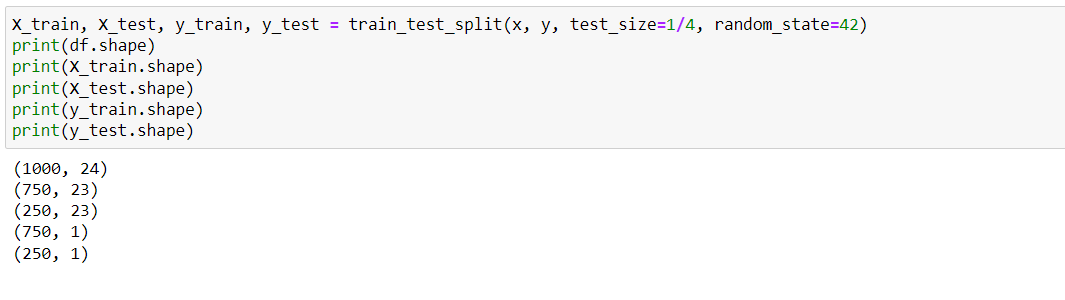


**[Fig – 3.3.1.]**



**[Fig – 3.3.2.]**

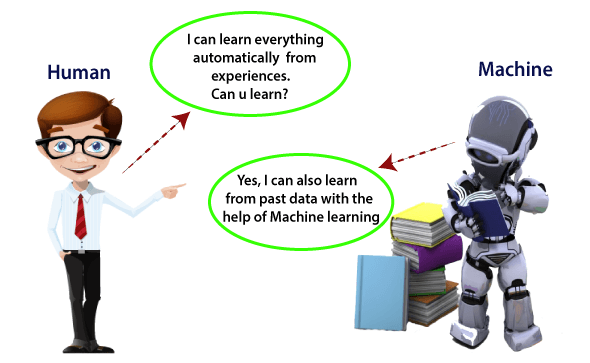
* Figure 3.3.1 represents all the independent features and figure 3.3.2 represents one dependent feature i.e., “Level”. After separating independent and dependent features train\_test\_split () method is used to splitting the dataset into training and testing part. Below figure represents the information about the volume of data present in the X\_train, X\_test, y\_train, and y\_test.



**[Fig – 3.3.3.]**

**Chapter – 4**

**Our Proposed Model**



**[Fig - 4.1.1.]**

* Machine Learning is said as a subset of artificial intelligence that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. We can define it in a summarized way as: Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.
* With the help of sample historical data, which is known as training data, machine learning algorithms build a mathematical model that helps in making predictions or decisions without being explicitly programmed. Machine learning brings computer science and statistics together for creating predictive models. Machine learning constructs or uses the algorithms that learn from historical data. The more we will provide the information, the higher will be the performance.
* A machine has the ability to learn if it can improve its performance by gaining more data.
* A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.
* Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:



**[Fig - 4.1.2.]**

**Classification of Machine Learning**

# 1) Supervised Learning

* Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.
* The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.
* The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is spam filtering.
* Supervised learning can be grouped further in two categories of algorithms:

1. Classification
2. Regression

2) Unsupervised Learning

* Unsupervised learning is a learning method in which a machine learns without any supervision.
* The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.
* In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data.

Machine learning Life cycle

* Machine learning has given the computer systems the abilities to automatically learn without being explicitly programmed. But how does a machine learning system work? So, it can be described using the life cycle of machine learning. Machine learning life cycle is a cyclic process to build an efficient machine learning project. The main purpose of the life cycle is to find a solution to the problem or project.

**Machine learning life cycle involves seven major steps, which are given below:**

* Gathering Data
* Data preparation
* Data Wrangling
* Analyze Data
* Train the model
* Test the model
* Deployment



* The most important thing in the complete process is to understand the problem and to know the purpose of the problem. Therefore, before starting the life cycle, we need to understand the problem because the good result depends on the better understanding of the problem.
* In the complete life cycle process, to solve a problem, we create a machine learning system called "model", and this model is created by providing "training". But to train a model, we need data, hence, life cycle starts by collecting data.

# 1. Gathering Data:

* Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.
* In this step, we need to identify the different data sources, as data can be collected from various sources such as files, database, internet, or mobile devices. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.

This step includes the below tasks:

* **Identify various data sources**
* **Collect data**
* **Integrate the data obtained from different sources**

By performing the above task, we get a coherent set of data, also called as a **dataset**. It will be used in further steps.

# 2. Data preparation

* After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.
* In this step, first, we put all data together, and then randomize the ordering of data.
* This step can be further divided into two processes:

1. **Data exploration: It is used to understand the nature of data that we have to work with. We need to understand the characteristics, format, and quality of data. A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends, and outliers.**
2. **Data pre-processing: Now the next step is preprocessing of data for its analysis.**

# 3. Data Wrangling

* Data wrangling is the process of cleaning and converting raw data into a useable format. It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step. It is one of the most important steps of the complete process. Cleaning of data is required to address the quality issues.
* It is not necessary that data we have collected is always of our use as some of the data may not be useful. In real-world applications, collected data may have various issues, including:

1. **Missing Values**
2. **Duplicate data**
3. **Invalid data**
4. **Noise**

So, we use various filtering techniques to clean the data.

* It is mandatory to detect and remove the above issues because it can negatively affect the quality of the outcome.

# 4. Data Analysis

* Now the cleaned and prepared data is passed on to the analysis step. This step involves:

1. **Selection of analytical techniques**
2. **Building models**
3. **Review the result**

* The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome. It starts with the determination of the type of the problems, where we select the machine learning techniques such as Classification, Regression, Cluster analysis, Association, etc. then build the model using prepared data, and evaluate the model.
* Hence, in this step, we take the data and use machine learning algorithms to build the model.

# 5. Train Model

* Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.
* We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and, features.

# 6. Test Model

* Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it.
* Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

# 7. Deployment

* The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system.
* If the above-prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system. But before deploying the project, we will check whether it is improving its performance using available data or not. The deployment phase is similar to making the final report for a project.

# **Supervised Machine Learning**

* Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.
* In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.
* Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).
* In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc.

# **How Supervised Learning Works?**

* In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.
* The working of Supervised learning can be easily understood by the below example and diagram:
* Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle, and Polygon. Now the first step is that we need to train the model for each shape.
* If the given shape has four sides, and all the sides are equal, then it will be labelled as a **Square**.
* If the given shape has three sides, then it will be labelled as a **triangle**.
* If the given shape has six equal sides then it will be labelled as **hexagon**.
* Now, after training, we test our model using the test set, and the task of the model is to identify the shape.
* The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the bases of a number of sides, and predicts the output.

# **Steps Involved in Supervised Learning:**

* First Determine the type of training dataset
* Collect/Gather the labelled training data.
* Split the training dataset into training dataset, test dataset, and validation dataset.
* Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
* Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
* Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.
* Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

**2. Classification**

* Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

Spam Filtering,

* Random Forest
* Decision Trees
* Logistic Regression
* Support vector Machines

# **Advantages of Supervised learning:**

* With the help of supervised learning, the model can predict the output on the basis of prior experiences.
* In supervised learning, we can have an exact idea about the classes of objects.
* Supervised learning model helps us to solve various real-world problems such as fraud detection, spam filtering, etc.

**Unsupervised Learning**

* As the name suggests, unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things. It can be defined as:
* Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.
* Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.
* Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset. The task of the unsupervised learning algorithm is to identify the image features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities between images.

**Below are some main reasons which describe the importance of Unsupervised Learning:**

* Unsupervised learning is helpful for finding useful insights from the data.
* Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
* Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.
* In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

# **Working of Unsupervised Learning**

* Here, we have taken an unlabeled input data, which means it is not categorized and corresponding outputs are also not given. Now, this unlabeled input data is fed to the machine learning model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.
* Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference between the objects.
* **Clustering**: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.

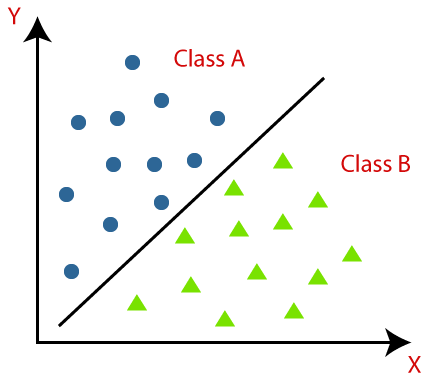
Unsupervised Learning algorithms:

* **K-means clustering**
* **KNN (k-nearest neighbors)**

# Advantages of Unsupervised Learning

* Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.
* Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.
* As we know, the Supervised Machine Learning algorithm can be broadly classified into Regression and Classification Algorithms. In Regression algorithms, we have predicted the output for continuous values, but to predict the categorical values, we need Classification algorithms.
* The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as: Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories.
* Unlike regression, the output variable of Classification is a category, not a value, such as "Green or Blue", "fruit or animal", etc. Since the Classification algorithm is a supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output.
* In classification algorithm, a discrete output function(y) is mapped to input variable(x).
* y=f(x), where y = categorical output
* The best example of an ML classification algorithm is Email Spam Detector.
* The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data.
* Classification algorithms can be better understood using the below diagram. In the below diagram, there are two classes, class A and Class B. These classes have features that are similar to each other and dissimilar to other classes.
* The algorithm which implements the classification on a dataset is known as a classifier. There are two types of Classifications:
* Binary Classifier: If the classification problem has only two possible outcomes, then it is called as Binary Classifier.

Examples: YES or NO, MALE or FEMALE, SPAM or NOT SPAM, CAT or DOG, etc.



* Multi-class Classifier: If a classification problem has more than two outcomes, then it is called as Multi-class Classifier.

Example: Classifications of types of crops, Classification of types of music.

# 2. **Confusion Matrix:**

* The confusion matrix provides us a matrix/table as output and describes the performance of the model.
* It is also known as the error matrix.
* The matrix consists of predictions result in a summarized form, which has a total number of correct predictions and incorrect predictions.

**Logistic Regression**

* Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
* Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
* Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
* In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
* The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
* Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
* Logistic regression uses the concept of predictive modeling as regression; therefore, it is called logistic regression, but is used to classify samples; therefore, it falls under the classification algorithm.
* Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



# **Logistic Function (Sigmoid Function):**

* The sigmoid function is a mathematical function used to map the predicted values to probabilities.
* It maps any real value into another value within a range of 0 and 1.
* The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the sigmoid function or the logistic function.
* In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

# **Assumptions for Logistic Regression**:

* The dependent variable must be categorical in nature.
* The independent variable should not have multi-collinearity.
* Logistic Regression Equation:
* The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

**Logistic Regression in Machine Learning**

* In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):
* Logistic Regression in Machine Learning
* But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:
* Logistic Regression in Machine Learning
* The above equation is the final equation for Logistic Regression.

# Type of Logistic Regression:

On the basis of the categories, Logistic Regression can be classified into three types:

* Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
* Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
* Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

# **Support Vector Machine:**

* Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
* The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
* SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using decision boundary or hyperplane.

SVM can be of two types:

* Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
* Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

# **How does SVM works?**

The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue. Consider the below image:



So as it is 2-d space so by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes. Consider the below image:



Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a hyperplane. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as margin. And the goal of SVM is to maximize this margin. The hyperplane with maximum margin is called the optimal hyperplane.

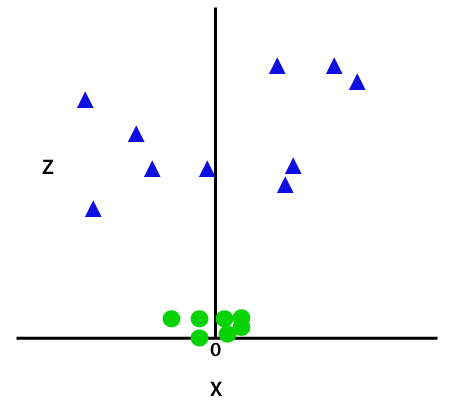


# **Non-Linear SVM:**

If data is linearly arranged, then we can separate it by using a straight line, but for non-linear data, we cannot draw a single straight line. Consider the below image:



So to separate these data points, we need to add one more dimension. For linear data, we have used two dimensions x and y, so for non-linear data, we will add a third dimension z. It can be calculated as: z=x2 +y2



By adding the third dimension, the sample space will become as below image:

So now, SVM will divide the datasets into classes in the following way. Consider the below image:



Since we are in 3-d Space, hence it is looking like a plane parallel to the x-axis. If we convert it in 2d space with z=1, then it will become as:



Hence we get a circumference of radius 1 in case of non-linear data.

* Non-linear retraction is a type of polynomial retreat. It is a way of modeling a non-linear relationship between dependent and independent variables. It is used in an area where the data show a curvy trend, and linear regression would not produce much more accurate results compared to non-linear regression. Here our proposed system is SVM (support vector machine) and for non-linear decomposition rbf - kernel is used. A support vector machine (SVM) is a type of in-depth learning algorithm that enables supervised learning to separate or reverse data groups. In AI and machine learning, supervised learning systems provide both input and desired output data, with a separation label. The non-linear decomposition with SVC (support vector classifier) model shows the 95% accuracy.
* Indirect separation: SVM can be extended to solve indirect separation tasks where a set of samples can be categorized. Using kernel functions, samples are drawn in the space of the maximum size, where linearization is possible.
* Linear Kernel is used when the data is linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a large number of Features in a particular Data Set.
* If we can separate the data in a straight line we use Non - Linear SVM. In this case, we have Kernel functions. They convert non-line spaces into line spaces. Converts the data into another size so that the data can be segmented.

# **Decision Tree Classification Algorithm:**

* 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 subtrees.
* Below diagram explains the general structure of a decision tree:

Note: A decision tree can contain categorical data (YES/NO) as well as numeric data.

## **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.

**How does the Decision Tree algorithm Work?**

* In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.
* For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below 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.

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as Attribute selection measure or ASM. By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:

* Information Gain
* Gini Index
* Information Gain: Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
* It calculates how much information a feature provides us about a class.
* According to the value of information gain, we split the node and build the decision tree.
* A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:
* Information Gain= Entropy(S)- [(Weighted Avg) \*Entropy(each feature)
* Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

Where,

S= Total number of samples

P(yes)= probability of yes

P(no)= probability of no

* Gini Index: Gini index is a measure of impurity or purity used while creating a decision tree in the CART (Classification and Regression Tree) algorithm.
* An attribute with the low Gini index should be preferred as compared to the high Gini index.
* It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.
* Gini index can be calculated using the below formula:
* Gini Index= 1- ∑jPj2

Pruning: Getting an Optimal Decision tree

* Pruning is a process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.
* A too-large tree increases the risk of overfitting, and a small tree may not capture all the important features of the dataset. Therefore, a technique that decreases the size of the learning tree without reducing accuracy is known as Pruning. There are mainly two types of tree pruning technology used:

Cost Complexity Pruning

Reduced Error Pruning.

# Advantages of the Decision Tree

* It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
* It can be very useful for solving decision-related problems.
* It helps to think about all the possible outcomes for a problem.
* There is less requirement of data cleaning compared to other algorithms.

# Disadvantages of the Decision Tree

* The decision tree contains lots of layers, which makes it complex.
* It may have an over fitting issue, which can be resolved using the Random Forest algorithm.
* For more class labels, the computational complexity of the decision tree may increase.

# **Random Forest Algorithm**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:



# **Assumptions for Random Forest**

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.

The predictions from each tree must have very low correlations.

**Why use Random Forest?**

Below are some points that explain why we should use the Random Forest algorithm:

It takes less training time as compared to other algorithms.

It predicts output with high accuracy, even for the large dataset it runs efficiently.

It can also maintain accuracy when a large proportion of data is missing.

How does Random Forest algorithm work?

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

# **Naïve Bayes Classifier Algorithm**

* Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
* It is mainly used in text classification that includes a high-dimensional training dataset.
* Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
* It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
* Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.
* Why is it called Naïve Bayes?
* The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

# **Bayes' Theorem:**

Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula for Bayes' theorem is given as: Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.

It can be used for Binary as well as Multi-class Classifications.

It performs well in Multi-class predictions as compared to the other Algorithms.

It is the most popular choice for text classification problems.

Applications of Naïve Bayes Classifier:

* It is used for Credit Scoring.
* It is used in medical data classification.
* It can be used in real-time predictions because Naïve Bayes Classifier is an eager learner.
* It is used in Text classification such as Spam filtering and Sentiment analysis.

**Types of Naïve Bayes Model:**

1. Gaussian: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.
2. Multinomial: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc. The classifier uses the frequency of words for the predictors.
3. Bernoulli: The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

# **K-NN Algorithm**

* There are two categories, Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:



* The K-NN working can be explained on the basis of the below algorithm:
* Step-1: Select the number K of the neighbors
* Step-2: Calculate the Euclidean distance of K number of neighbors
* Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
* Step-4: Among these k neighbors, count the number of the data points in each category.
* Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
* Step-6: Our model is ready.
* Suppose we have a new data point and we need to put it in the required category. Consider the below image:



* Firstly, we will choose the number of neighbors, so we will choose the k=5.
* Next, we will calculate the Euclidean distance between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



Euclidian distance between A1 and B2 =

By calculating Euclidian distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B.



* Below are some points to remember while selecting the value of K in the K-NN algorithm:
* There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
* A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
* Large values for K are good, but it may find some difficulties.

In this paper several machine learning algorithms are used these are Logistic Regression, Support Vector machine, Random Forest, Decision Tree, Naïve Bayes and KNN (K- Nearest Neighbor) also have been used. Depression predicts the outflow of phase-dependent variability. Therefore the result should be phase or separate value. Either Yes or No, 0 or 1, True or False, etc. but instead of giving a direct value such as 0 and 1, it provides possible values ​​between 0 and 1.Based on categories, Logistic Regression can be divided into three categories:

* Binomial: In binomial Logistic regression, there can be only two types of dependent elements, such as 0 or 1, Pass or Fail, etc.
* Multinomial: In most cases, there may be 3 or more random types of dependent variants, such as "cat", "dogs", or "sheep".
* Ordinal: In Ordinal Logistic retransmission, there may be 3 or more ordered types of possible dependencies, such as "Down", "Medium", or "Up".
* In Support Vector Machine, Support Vector Classifier SVC () is used to predict the outcomes. The SVC model is giving the accuracy of 96% with linear kernel, Where GaussianNB() of Naïve Bayes algorithm produces less accuracy than all other algorithm, which Is 89%.
* Decision Tree Classifier () of Decision Tree, Random Forest Classifier () of random Forest and K Neighbors Classifier () of KNN giving the accuracy of 100%. Here Decision Tree, random Forest and KNN model giving the accuracy score as 0.1 i.e., 100%.
* Here our proposed model is Random Forest after evaluating the performance of all the machine learning models.

**Data Preprocessing**

**Feature Extraction**

**Training Model**

**Evaluating Model**

**Implementing Algorithm**

**Data Collection**

**[Work Flow Diagram]**

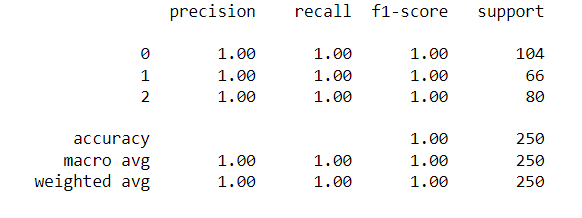
[Fig – 5.0]

**Chapter - 5**

**Performance Evaluation:**

**Classification Report**

* After model become trained classification report is shown below in figure [fig – 5.1]. Precision and recall are two extremely important model evaluation metrics. While precision refers to the percentage of your results which are relevant, recall refers to the percentage of total relevant results correctly classified by your algorithm.
* Precision is the measure of a classifier's correctness. In an imbalanced classification problem with two classes, precision is calculated as the number of true positives divided by the total number of true positives and false positives. The result is a value between 0.0 for no precision and 1.0 for full or perfect precision. Precision should ideally be 1 (high) for a good classifier. Precision becomes 1 only when the numerator and denominator are equal i.e. [TP = TP +FP], this also means FP is zero. As FP increases the value of denominator becomes greater than the numerator and precision value decreases, which we don't want.
* Recall is a measure of the classifier's completeness; the ability of a classifier to correctly find all positive instances. For each class, it is defined as the ratio of true positives to the sum of true positives and false negatives. i.e. [TP/TP + FN].
* Accuracy score is the sum of true positives and true negatives divided by the total number of samples. This is only accurate if the model is balanced. It will give inaccurate results if there is a class imbalance. Accuracy is (TP+TN/TP+FP+FN+TN)
* The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. It is primarily used to compare the performance of two classifiers. Suppose that classifier A has a higher recall, and classifier B has higher precision.

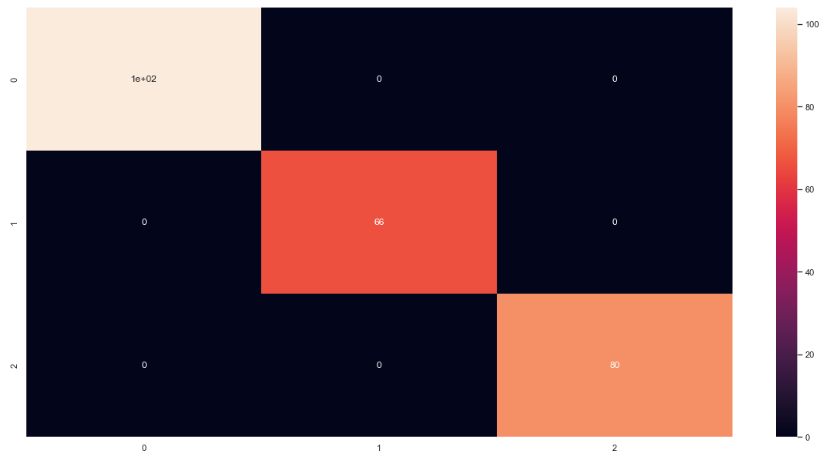


**[Fig – 5.1]**

* [fig – 5.1] represents the classification report of RandomForestClassifier model in which precision, recall, f1-score and support all shows the exactness of 100%, and also the accuracy is 100%.

**Confusion Matrix**

* The Confusion matrix of the model is shown below at [fig – 5.2]. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

****

**[Fig – 5.2]**

* For performance evaluation in this paper cross\_val\_score() is used which comes from model\_selection of Sklearn library.Which shows cross validation scores(0.996, 0.992), (0.964, 0.976), (1.0, 1.0), (1.0, 1.0), (0.996, 0.998), (0.888, 0.891) for Logistic Regression, Support vector machine, Decision Tree, random Forest, KNN, and Naive Bayes algorithms respectively.

**Chapter – 6**

**Literature Review**

* In this paper we use all the supervised machine learning algorithms Logistic Regression, Support Vector Machine, Decision Tree, random Forest, K-Nearest Neighbor and Naïve Bayes but by using Decision Trees and random Forest model predicts with a better accuracy and the accuracy is 0.1 i.e. 100%. Here in this paper we have taken Random Forest Classifier model for creation of model and for better accuracy because some advantages of Random Forest over Decision Tree is described below.
* Random forest is a kind of ensemble classifier which is using a decision tree algorithm in a randomized fashion and in a randomized way, which means it is consisting of different decision trees of different sizes and shapes, it is a machine learning technique that solves the regression and classification problems,
* The biggest advantage of Random Forest is that it relies on collecting various decision trees to arrive at any solution. This is an ensemble algorithm that considers the results of more than one algorithm of the same or different kind of classification.

**Chapter - 7**

**Conclusion & Future Scope**

* After analyzing the dataset and building the model and evaluate the performance of the model we found that The Random Forest Classifier is better for the prediction of Cancer and the by taking all the input parameters of the patients. In this Paper the data set is from Kaggle and the Cancer patient’s data contains 1000 records of individual patients.
  1. **. Feature Scope:**
* Every year, Epidemiologists diagnose 14 million new cancer patients worldwide. That's what millions of people will face years of uncertainty. Pathologists have been diagnosing and diagnosing cancer for decades. Most pathologists have a 96-98% success rate for cancer screening. They are very good at that part. If there is any website that contains the cancer patient’s data worldwide then it would be helpful to detect and classify the cancer patients are in which stage (low, medium and high) and also separate the dataset into different sets according to different types of cancer.

**Chapter – 7**

1. ***References:***

* F Bray, J Ferlay, I Soerjomataram, RL Siegel, LA Torre and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", *CA Cancer J Clin.*, 2018.
* View at: Publisher Site | Google Scholar
* Rebecca L. Siegel, Kimberly D. Miller and Ahmedin Jemal DVM, *Cancer statistics 2019*, 2019.
* View at: Publisher Site | Google Scholar
* SH Jafari, Z Saadatpour, A Salmaninejad et al., "Breast cancer diagnosis: Imaging techniques and biochemical markers", *J Cell Physiol.*, vol. 233, pp. 5200-5213, 2018.
* View at: Publisher Site | Google Scholar
* Kumar Ajay, Sushil Rama and Tiwari Arvind, "Cancer Survival Analysis Using Machine Learning", *Proceedings of International Conference on Sustainable Computing in Science Technology and Management (SUSCOM)*, 2019.
* View at: Publisher Site | Google Scholar
* R Jafari-Marandi, S Davarzani, M S Gharibdousti and BK Smith, "An Optimum ANN-based Breast Cancer Diagnosis: Bridging Gaps between ANayN Learning and Decision-making Goals", *Applied Soft Computing Journal*, 2018.
* View at: Publisher Site | Google Scholar
* M. H. Tafish and A. M. El-Halees, "Breast Cancer Severity Degree Predication Using Data Mining Techniques in the Gaza Strip", *2018 International Conference on Promising Electronic Technologies (ICPET)*, pp. 124-128, 2018.
* Ahmed M. Abdel-Zaher and Ayman M. Eldeib, Breast cancer classification using deep belief networks, 2016
* View at: Google Scholar
* M. Ashraf, G. Geng, X. Wang, F. Ahmad, and F. Abid, “A globally regularized joint neural architecture for music classification,” IEEE Access, vol. 8, pp. 220980–220989, 2020.
* View at: Publisher Site | Google Scholar
* Sun Y, Reynolds HM, Parameswaran B et al (2019) Multiparametric MRI and radiomics in prostate cancer: a review. Australas Phys Eng Sci Med 42:3–25
* View at: Publisher Site | Google Scholar
* Stanzione A, Imbriaco M, Cocozza S et al (2016) Biparametric 3T magnetic resonance imaging for prostatic cancer detection in a biopsy-naïve patient population: a further improvement of PI-RADS v2?EurJRadiol85:2269–227