*A Project Report*

*on*

**Classification of Acute Myeloid and Lymphocytic Leukemia using Machine Learning Algorithms based on Gene Expression Profiling**

*Submitted in partial fulfillment of the requirements*

*for the award of the degree of*

#### BACHELOR OF TECHNOLOGY

*in*

## Computer Science & Engineering

### *by*

### THEJASREE M (174G1A05A1)

### THRIVENI E (174G1A05A2)

### SAI SUDHA R (174G1A0574)

### RAVI KUMAR D (174G1A0565)

**Under the Guidance of**

## Mr. G. Chinna Pullaiah M.Tech.,( Ph.D)

Assistant Professor



## DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Srinivasa Ramanujan Institute of Technology: Anantapuramu

**(Affiliated to JNTUA ,Accredited by NAAC with ‘A’ Grade & Accredited by NBA, Approved by AICTE, New Delhi**

**& Accredited by NBA(CSE,ECE,EEE))**

**2020-21**

SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY: ANANTAPURAMU

(Affiliated to JNTUA, Accredited by NAAC with ‘A’ Grade, Approved by AICTE, New Delhi & Accredited by NBA (EEE, ECE & CSE)

****

**Certificate**

This is to certify that the project report entitled Classification of Acute Myeloid and Lymphocytic Leukemia using Machine Learning Algorithms based on Gene Expression Profiling is the bonafide work carried out by **THEJASREE M** bearing Roll Number 174G1A05A1, **THRIVENI E** bearing Roll Number 174G1A05A2, **SAI SUDHA R** bearing Roll Number 174G1A0574 and **RAVI KUMAR D** bearing Roll Number 174G1A0565 in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering** during the academic year 2020-2021.

**Signature of the Guide Head of the Department**

Mr. G. Chinna Pullaiah MTech, (Ph.D) Dr. G. K.V Narasimha Reddy,Ph.D

Assistant Professor Professor & HOD

Date:  **EXTERNAL EXAMINER**

Place: Ananthapuramu

**ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible, whose constant guidance and encouragement crowned our efforts with success. It is a pleasant aspect that we have now the opportunity to express my gratitude for all of them.

It is with immense pleasure that we would like to express my indebted gratitude to my Guide **Mr. G. Chinna Pullaiah,**MTech,(Ph.D)**, Assistant Professor, Computer Science & Engineering**, who has guided me a lot and encouraged me in every step of the project work. We thank him for the stimulating guidance, constant encouragement and constructive criticism which have made possible to bring out this project work.

We express our deep felt gratitude to **Dr. P. Chitralingappa, M.Tech, Ph.D, Associate Professor and Mrs. P.Soumya,M.Tech, Assistant Professor** project coordinator valuable guidance and unstinting encouragement enabled us to accomplish our project successfully in time.

We are very much thankful to **Dr. G. K. V. Narasimha Reddy,**Ph.D**, Professor and Head of Department, Computer Science & Engineering**, for his kind support and for providing necessary facilities to carry out the work.

We wish to convey my special thanks to **Dr. G. Balakrishna,**Ph.D **Principal** of **Srinivasa Ramanujan Institute of Technology** for giving the required information in doing my project work. Not to forget, we thank all other faculty and non-teaching staff, and my friends who had directly or indirectly helped and supported me in completing my project in time.

We also express our sincere thanks to the Management for providing excellent facilities**.**

Finally, we wish to convey our gratitude to our family who fostered all the requirements and facilities that we need.

**Project Associates**

**DECLARATION**

We**, Ms. THEJASREE M** bearing reg no**:174G1A05A1, THRIVENI E** bearing regno**:174G1A05A2, SAI SUDHA R** bearing reg no**:174G1A0574 and RAVI KUMAR D** bearing reg no**:174G1A0565** students of **SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY, Rotarypuram**, hereby declare that the dissertation entitled **“CLASSIFICATION OF ACUTE MYELOID AND LYMPHOCYTIC LEUKEMIA USING MACHINE LEARNING ALGORITHMS BASED ON GENE EXPRESSION PROFILING”** embodies the report of our project work carried out by us during IV Year **BACHELOR OF TECHNOLOGY** and this work has been submitted for the partial fulfillment of the requirements for the award of **BACHELOR OF TECHNOLOGY** degree.

The results embodied in this project report have not been submitted to any other Universities of Institute for the award of Degree.

**THEJASREE M Reg No: 174G1A05A1**

**THRIVENI E Reg No: 174G1A05A2**

**SAI SUDHA R Reg No: 174G1A0574**

**RAVI KUMAR D Reg No: 174G1A0565**

**ABSTRACT**

Health care industry plays a vital role in saving people’s life. Developing software which can be used in such a field increases the scope of engineering in larger context. Cancer has been portrayed as a heterogeneous disease comprising of a wide range of subtypes. The early diagnosis of a cancer type is very important to determine the course of medical treatment required by the patient. The significance of classifying cancerous cells into benign or malignant has driven many research studies. In the past years researchers have been encouraged to use different machine learning (ML) techniques for leukemia detection, as well as prediction of survivability and recurrence.

# Cancer is the second leading cause of mortality across the world. Thousands of people die and suffer due to cancers. Accurate and early prediction of cancer can assist the diagnosis earlier. So, given the growing trend on the application of ML techniques in precision medicine, here we present an ML techniques which classifies the leukemia Robust Cancers.

# In this project, we implement machine learning methods with the intent of classifying AML and ALL using the given patient’s genetic profile and hematological data of Gene Expression (GE) dataset. This Experiment shows that the proposed solution is capable to improve performance of Acute Myeloid and Lymphocytic Leukemia cancer classification

**Keywords**:  Leukemia, Acute Myeloid and Lymphocytic Leukemia, Machine Learning (ML), Gene Expression dataset.

**CONTENTS**

**List of abbreviations viii**

**List of figures ix**

**CHAPTER 1-INTRODUCTION 1-12**

1.1 Objective 1

1.2..Machine Learning 2

##### 1.1.1.Supervised Learning 2

##### 1.1.2.Unsupervised Learning 2

##### 1.3.Feature Selection 2

1.3.1.Feature selection techniques 3

##### 1.4.Wrapper Methods 3

##### 1.5.Classification 3

1.5.1.Binomial 4

1.5.2.Multi-Class Binomial 4

**CHAPTER 2-LITERATURE SURVEY 13-14**

## 2.1.Existing System 13

2.2 Proposed Model 13

# CHAPTER 3-ANALYSIS 15-23

# 3.1. Introduction 15

# 3.2. Software Requirement Specification 15

3.3. Hardware Requirements 16

3.4.Software Requirements 16

3.4.1 Anaconda 16

3.4.2.Languages Used 23

**CHAPTER 4-DESIGN 24-30**

4.1.UML Introduction 24

4.1.1.Usage of UML in Project 24

4.2.System Architecture 24

4.3. Data Flow Diagram 26

4.4..Use Case Diagram 27

4.5. State Chart Diagram 29

4.6.Steps involved in Design 30

**CHAPTER 5 –IMPLEMETATION 31-46**

5.1.Libraries Used 31

5.2 Implementation 33

5.2.1.Dataset description 33

5.2.2. Execution 34

# CHAPTER 6 –TESTING 47-48

6.1.Black-Box Testing 47

6.2.White-Box Testing 47

6.3.Performance Evaluation 48

# CONCLUSION 49

**BIBLIOGRAPHY 50**

**LIST OF ABBREVIATIONS**

CSV Comma-separated values

ALL Acute Myeloid Leukemia

AML Acute Lymphocytic Leukemia

SVM Support Vector Machine

PCA Principle Component Analysis

UML Unified Modeling Language

SMOTE Synthetic Minority Over-Sampling

ROC Receiver Operating Characteristic

AOC Area Under Curve

ACC Accuracy

ML Machine Learning

**viii**

**LIST OF FIGURES**

**Fig. No Description Page No**

Fig.1.1 Binomial Classification 4

Fig.1.2 Multi-class classification 4

Fig.1.3 SVM 6

Fig.1.4 Decision Tress 8

Fig.1.5 Trees in Random Forest 9

Fig.1.6 Linear Regression & Logistic Regression 11

Fig.4.1 System Architecture 21

Fig.4.2 Data Flow Diagram 22

Fig.4.3 Use Case Diagram 23

Fig.4.4 State Chart Diagram 24

Fig.5.1 Importing libraries 27

Fig.5.2 Output of first n rows 28

Fig.5.3 Checking for nulls 28

Fig.5.4 Dropping ‘call’ column 28

Fig.5.5 Joining all data 29

Fig.5.6 Transposing rows and columns 29

Fig.5.7 Plot of Class Distribution 30

Fig.5.8 Categorical variables 30

Fig.5.9 Output of categorical variables 31

Fig.5.10 Splitting the dataset 31

Fig.5.11 Dimensionality Reduction 32

Fig.5.12 3D plot of first 3 Principle components 32

Fig.5.13 Sampling 33

Fig.5.14 Hyperparameter optimization of SVM 33

Fig.5.15 SVM model 34

Fig.5.16 Output of SVM model 34

Fig.5.17 Hyperparameter optimization of Logistic regression 35

Fig.5.18 Logistic Regression Model 35

Fig.5.19 Output of logistic regression 36

Fig.5.20 Model Evaluation 36

**xi**

**CHAPTER-1**

**INTRODUCTION**

Leukemia arises from malfunctions in stem cells within the bone marrow that cause abnormal white blood cells to flood into the bloodstream. There are several types of leukemia, which are divided based mainly on whether the leukemia is acute (fast growing) or chronic (slower growing), and whether it starts in myeloid cells or lymphoid cells. Acute myeloid leukemia (AML) begins in early myeloid cells These are cells that become white blood cells (other than lymphocytes), red blood cells, or platelet making cells. It’s the most common type of leukemia in older people. Acute lymphocytic leukemia (ALL) starts in cells that become  lymphocytes white blood cells that are an important part of your immune system. About 80% of childhood leukemia cases are this type.

Research has shown correlations between AML and ALL and certain genetic mutations; these genetic mutations appear in an individual’s genome years before onset of the disease. These genetic changes can serve as biomarkers for cancer cell proliferation, allowing for early classification of AML and ALL if an accurate screening test can be developed.

The data science and machine learning algorithms are used for classification. For classification using machine learning algorithm all of the features present in the dataset might not be useful in building a machine learning model to make the necessary prediction. Using too many features might even make the predictions worse.

## 1.1 Objective:

The main goal of this project is build a model to obtain best possible accuracy with minimal set of features to classify AML and ALL based on which Molecular Classification of Cancer by Gene Expression Monitoring is focused. Feature selection plays a huge role in building a machine learning model. Feature selection can make use of less number of data features to give best possible accuracy. Especially when dealing with a large number of variables there is a need for dimensionality reduction. Feature Selection can significantly improve a learning algorithm's performance. In this project Gene Expression Dataset (Global et al.) dataset is employed and the dataset is available in *.csv* format.

## 1.2.Machine Learning:

Machine Learning is a study of algorithms and [statistical](https://en.wikipedia.org/wiki/Statistical_model) models that [computer](https://en.wikipedia.org/wiki/Computer_systems) [systems](https://en.wikipedia.org/wiki/Computer_systems) use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. It combines data with statistical tools to predict an output. This output is then used by corporate to makes actionable insights.

The types of machine learning algorithms differ in their approach, the type of data they input and output, and the type of task or problem that they are intended to solve.

##### **1.2.1.Supervised Learning:**

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as [training data](https://en.wikipedia.org/wiki/Training_data), and consists of a set of training examples.

Supervised learning algorithms include [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression.](https://en.wikipedia.org/wiki/Regression_analysis) Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range.

##### **1.2.2.Unsupervised Learning:**

Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms therefore learn from test data that has not been labeled, classified or categorized. Cluster analysis is the assignment of a set of observations into subsets (called *clusters*) so that observations within the same cluster are similar according to one or more predesignated criteria, while observations drawn from different clusters are dissimilar. Other methods are based on estimated density and graph connectivity.

## 1.3.Feature Selection:

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for several reasons:

* simplification of models to make them easier to interpret by researchers/users
* shorter training times,
* to avoid the curse of dimensionality,
* enhanced generalization by reducing overfitting.

The central premise when using a feature selection technique is that the data contains some features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information.

##### **1.3.1.Feature selection techniques:**

There are three general classes of feature selection algorithms: filter methods, wrapper methods and embedded methods. In our project we have used Wrapper Methods.

**1.3.1.1.Filter Methods**

Filter feature selection methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset.Some examples of some filter methods include the Chi squared test, information gain and correlation coefficient scores.

##### **1.3.1.2.Embedded methods**

This method complete the feature selection process within the construction of the machine learning algorithm itself. In other words, they perform feature selection during the model training, which is why we call them embedded methods.

**1.4.Wrapper Methods**

Wrapper methods are based on greedy search algorithms as they evaluate all possible combinations of the features and select the combination that produces the best result for a specific machine learning algorithm.

Categories:

* + - 1. Sequential feature selection(SFS),
      2. Sequential Backward Elimination(SBE) and
      3. Exhaustive feature selection.

In this project, SFS and SBE are used for selecting feature.

## 1.5.Classification:

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). Classification belongs to the category of supervised learning where the targets also provided with the input data. There are many applications in classification in many domains such as in credit approval, medical diagnosis, target marketing etc.

There are 2 types of Classification:

* + - Binomial
    - Multi-Class Binomial:

**1.5.1.Binomial**

Binary or binomial classification is the task of classifying the elements of a given set into two groups (predicting which group each one belongs to) on the basis of a classification rule. Our project is a Binomial Classification that is Benign(B) and Malignant(M).

M

B

**Fig. 1.1 Binomial Classification**

**1.5.2.Multi-Class Binomial:**

This classification is the task of classifying the elements of a given set into more than two groups (predicting which group each one belongs to) on the basis of a classification rule.

Class 1

class 3

Class 2

**Fig.1.2 Multi-class classification**

Some of the Classification Algorithms are:

* + - Random Forest
    - Decision Tree
    - Support Vector Machine
    - Naïve Bayes Classifier
    - Nearest Neighbour

##### **1.5.2.1.Naive Bayes Classifier:**

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.

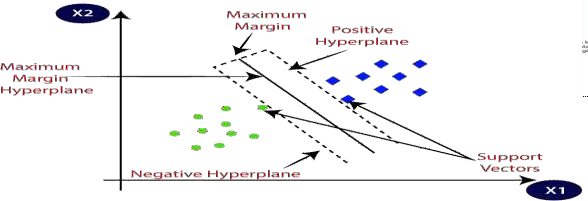
The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

**Naive**: 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.

##### Support Vector Machines:

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 a decision boundary or hyperplane:

**Fig.1.3 SVM.**

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.

**1.5.2.2.Decision Tree Classifier:**

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** Decision tree learning is a method commonly used in data mining. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram at right. Each [interior node](https://en.wikipedia.org/wiki/Interior_node) corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

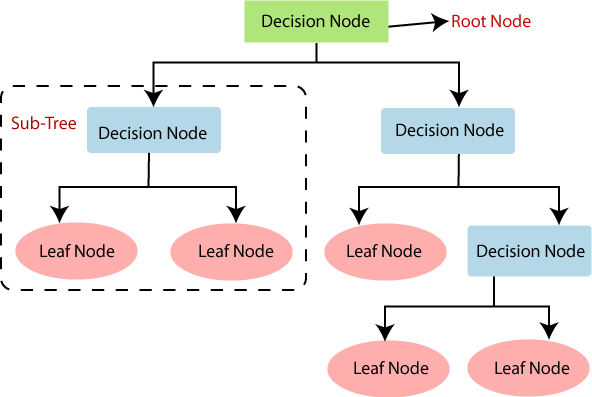
A decision tree is a simple representation for classifying examples. For this section, assume that all of the input [features](https://en.wikipedia.org/wiki/Feature_(machine_learning)) have finite discrete domains, and there is a single target feature called the "classification". Each element of the domain of the classification is called a class. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with an input feature are labeled with each of the possible values of the target or output feature or the arc leads to a subordinate decision node on a different input feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

A tree can be "learned” by splitting the source [set](https://en.wikipedia.org/wiki/Set_(mathematics)) into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called [recursive partitioning](https://en.wikipedia.org/wiki/Recursive_partitioning). See the examples illustrated in the figure for spaces that have and have not been partitioned using recursive partitioning, or recursive binary splitting. The [recursion](https://en.wikipedia.org/wiki/Recursion) is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees is an example of a [greedy algorithm](https://en.wikipedia.org/wiki/Greedy_algorithm), and it is by far the most common strategy for learning decision trees from data.

I[n data mining,](https://en.wikipedia.org/wiki/Data_mining) decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data. Data comes in records of the form:

(X,Y) = (x1, x2, x3,…..,xk, Y)

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector **X** is composed of the features, x1, x2, x3 etc., that are used for that task.



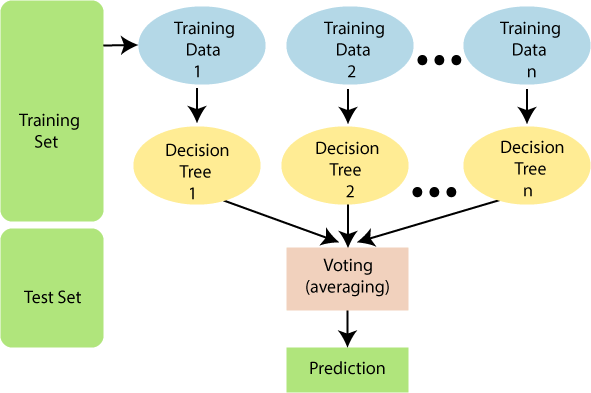
**Fig.1.4 Decision Tress**

In this project the Random Forest classification algorithm is used as its performance is quite effective.

##### **1.5.2.3.Random Forest Classifier:**

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.

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.



**Fig.1.5 Trees in Random Forest**

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. Two types of randomness are built into the trees.

1. Each tree is built on a random sample from the original data.
2. At each tree node, a subset of features are randomly selected to generate the best split.
3. Random sampling of training observations:

When training, each tree in a random forest learns from a random sample of the data points. The samples are drawn with replacement, known as *bootstrapping,* which means that some samples will be used multiple times in a single tree. The idea is that by training each tree on different samples, although each tree might have high variance with respect to a particular set of the training data, overall, the entire forest will have lower variance but not at the cost of increasing the bias.

At test time, predictions are made by averaging the predictions of each decision tree. This procedure of training each individual learner on different bootstrapped subsets of the data and then averaging the predictions is known as *bagging*, short for [*bootstrap*](https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/)[*aggregating*](https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/)*.*

**Random Subsets of features for splitting nodes:**

The other main concept in the random forest is that only a subset of all the features are considered for splitting each node in each decision tree. Generally this is set to sqrt(n- features) for classification meaning that if there are 16 features, at each node in each tree, only 4 random features will be considered for splitting the node. (The random forest can also be trained considering all the features at every node as is common in regression. These options can be controlled in the Scikit-Learn Random Forest implementation).

##### **Advantages**

* + 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.

##### **1.4.2.4 Logistic Regression:**

**Logistic regression** models a relationship between predictor variables and a categorical response variable. For example, we could use logistic regression to model the relationship between various measurements of a manufactured specimen (such as dimensions and chemical composition) to predict if a crack greater than 10 mils will occur (a binary variable: either yes or no). Logistic regression helps us estimate a probability of falling into a certain level of the categorical response given a set of predictors. We can choose from three types of logistic regression, depending on the nature of the categorical response variable:

**Binary Logistic Regression**:

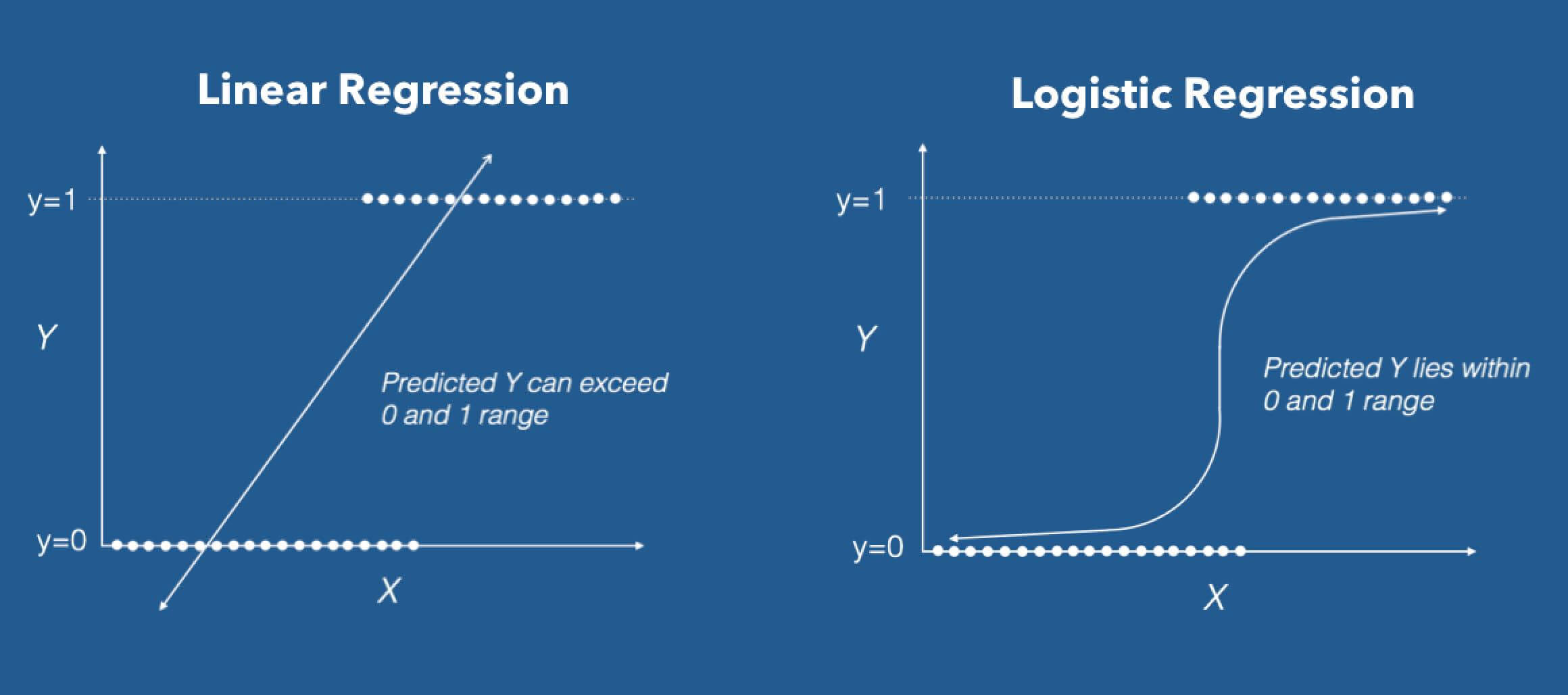
Used when the response is binary (i.e., it has two possible outcomes). The cracking example given above would utilize binary logistic regression. Other examples of binary responses could include passing or failing a test, responding yes or no on a survey, and having high or low blood pressure.

**Nominal Logistic Regression**:

Used when there are three or more categories with no natural ordering to the levels. Examples of nominal responses could include departments at a business (e.g., marketing, sales, HR), type of search engine used (e.g., Google, Yahoo!, MSN), and color (black, red, blue, orange).

**Ordinal Logistic Regression**:

Used when there are three or more categories with a natural ordering to the levels, but the ranking of the levels do not necessarily mean the intervals between them are equal. Examples of ordinal responses could be how students rate the effectiveness of a college course (e.g., good, medium, poor), levels of flavors for hot wings, and medical condition (e.g., good, stable, serious, critical).



**Fig1.6.Linear Regression & Logistic Regression**

**CHAPTER-2**

**LITERATURE SURVEY**

## 2.1.Existing System:

## Prediction using traditional methods and models involves various risk factors and it consists of various measures of algorithms such as datasets, programs and much more to add on. High-risk and Low-risk patient classification is done on the basis of the tests that are done in group. But these models are only valuable in clinical situations and not in big industry sector. So, to include the disease predictions in various health related industries, we have used the concepts of machine learning and supervised learning methods to build the predictions system.

After doing the research and comparison of all the algorithms and theorems of machine learning we have come to conclusion that all those algorithms such as Decision Tree, KNN, Naïve Bayes, Regression and Random Forest Algorithm all are important in building a model on classifying AML and ALL using gene Expression Profiling. After using various techniques such as neural networks to make predictions of the diseases and after doing that we come to conclusion that it can predicts up to 90% accuracy rate after doing the experimentation and verifying the results. The information of patient statistics, results, disease history in recorded in EHR, which enables to identify the potential data centric solution, which reduces the cost of medical case studies. Existing system can predict the disease but not the sub type of the disease and it fails to predict the condition of the people, the predictions of disease have been indefinite and non-specific.

**2.2 Proposed Model:**

In the Proposed Model ,We would like to work on finding more accuracy rate and classify AML and ALL based on Gene Expression profiling using Logistic regression And Support Vector Machines.The scientific community often uses logistic regression models on biomedical datasets with high-dimensionality, and continues to research how to tune and alter those models.All models needed significant regularization to prevent overfitting, So we wanted to try using Logistic regression to reduce overfitting and mitigate the variance . The Gene Expression Dataset was used as a training set to compare the performance of the various machine learning techniques in terms of key parameters such as accuracy, and precision. The results obtained are very competitive and can be used for detection and treatment.

Most of the models were build with utmost 30 features and the problems arise due to the large number of dimensions are

* Difficulty of models to interpret by researchers/users
* Larger training times,
* curse of dimensionality,
* problem of overfitting.

J. Laosai and K. Chamnongthai. "Acute leukemia classification by using SVM and K-Means clustering" Proceedings of the International Electrical Engineering Congress pp. 1-4. 2014 [2] ―Types of leukemia‖15 Nov. 2019 [1]

“Applications of Machine Learning in Acute Myeloid Leukemia Prediction” Joanna Bitton and Erica Dominic New York University jtb470@nyu.edu, [ehd255@nyu.edu](mailto:ehd255@nyu.edu) 2018 [2]

“Machine learning to predict 5-year survival among pediatric Acute Myeloid Leukemia patients and development of OSPAM-C online survival prediction tool” by Ashis Kumar Das1 , Shiba Mishra and Saji Saraswathy Gopalan [April,2020] [3]

“Scalable Prediction of Acute Myeloid Leukemia Using High- Dimensional Machine Learning and Blood Transcriptomics” by Stefanie WarnatHerresthal, Konstantinos Perrakis, Bernd Taschler.[Jan,2020] [3]

Beam A, Kohane I. Big data and machine learning in health care. *JAMA*. 2018: **319**:  1317- 1318.[Crossref](https://onlinelibrary.wiley.com/servlet/linkout?suffix=null&dbid=16&doi=10.1111%2Fijlh.13089&key=10.1001%2Fjama.2017.18391) [PubMed](https://onlinelibrary.wiley.com/servlet/linkout?suffix=null&dbid=8&doi=10.1111%2Fijlh.13089&key=29532063) [Web of Science®](https://onlinelibrary.wiley.com/servlet/linkout?suffix=null&dbid=128&doi=10.1111%2Fijlh.13089&key=000429035900010)[Google Scholar](http://scholar.google.com/scholar_lookup?hl=en&volume=319&publication_year=2018&pages=1317-1318&journal=JAMA&author=A+Beam&author=I+Kohane&title=Big+data+and+machine+learning+in+health+care) [4]

Machine Learning Algorithms For Diagnosis Of Leukemia by Italia Joseph Maria, T. Devi, D. Ravi Kantarjian et al. Blood Cancer Journal (2021) 11:41 [https://doi.org/10.1038/s41408-021-00425-3](https://doi.org/10.1038/s41408-021-00425-3%20%5b6) [5]

Leukemia Prediction Using Sparse Logistic Regression Tapio Manninen1 \*, Heikki Huttunen1 , Pekka Ruusuvuori1 , Matti Nykter May 3, 2013; Accepted July 12, 2013; Published August 30, 2013 [6]

# CHAPTER-3

# ANALYSIS

# 3.1. Introduction

## The Analysis Phase is where the project life cycle begins. This is the phase where you break down the deliverables in the high-level Project Charter into the more detailed business requirements. Gathering requirements is the main attraction of the Analysis Phase. The process of gathering requirements is usually more than simply asking the users what they need and writing their answers down. Depending on the complexity of the application, the process for gathering requirements has a clearly defined process of its own. This process consists of a group of repeatable processes that utilize certain techniques to capture, document, communicate, and manage requirements. This formal process, which will be developed in more detail, consists of four basic steps.

1. **Elicitation** – I ask questions, you talk, I listen
2. **Validation** – I analyze, I ask follow-up questions
3. **Specification** – I document, I ask follow-up questions
4. **Verification** – We all agree

Most of the work in the Analysis Phase is performed by the role of analyst.

## 3.2. Software Requirement Specification

SRS is a document created by system analyst after the requirements are collected. SRS defines how the intended software will interact with hardware, external interfaces, speed of operation, response time of system, portability of software across various platforms, maintainability, speed of recovery after crashing, Security, Quality, Limitations etc.

The requirements received from client are written in natural language. It is the responsibility of system analyst to document the requirements in technical language so that they can be comprehended and useful by the software development team.

**3.3. Hardware Requirements**

### Any Contemporary PC.

**3.4.Software Requirements**

Operating System **:**Windows 10

Tools **:** Anaconda

Dataset **:** Excel Sheet

Languages Used **:** Python

**3.4.1 Anaconda**

**Installation of Anaconda**

**Overview**

Anaconda is a package manager, an environment manager, and Python distribution that contains a collection of many open source packages (numpy, scikit-learn, scipy, pandas to name a few). Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS. If you need additional packages after installing Anaconda, you can use Anaconda’s package manager, conda or pip to install those packages. This is highly advantageous as you don’t have to ma manage dependencies between multiple packages.

1. Go to the [Anaconda Website](https://www.anaconda.com/download/#windows) and choose either a Python 3.x graphical installer (A) or a Python 2.x graphical installer (B). If you aren’t sure which Python version you want to install, choose Python 3. Do not choose both.

Graphical user interface, text, application

Description automatically generated

2. Locate your download.

Graphical user interface, text, application, email, website

Description automatically generated

Installing as administrator is for the case you don’t have permission install anaconda in the location you want or to add anaconda to your path.

When the screen below appears, click on Next.

Graphical user interface, text, application

Description automatically generated

3.Read the License Agreement and click on I Agree

Graphical user interface, text, application

Description automatically generated

5. Choose either Just Me (recommended) or All Users.

6. Please make a note of your installation location (1) and then click Next (2).

7. This is an important part of the installation process.

The recommended approach is to not check the box (1) to add Anaconda to your path. This means you will have to use Anaconda Navigator or the Anaconda Command Prompt (located in the Start Menu under “Anaconda”) when you wish to use Anaconda (you can always add Anaconda to your PATH later if you don’t check the box). If you want to be able to use Anaconda in your command prompt, please use the alternative approach and check the box. Click on Install (2). This is important. Consider what you are doing in this step.

 8. Click on Next.

Graphical user interface, text, application

Description automatically generated

9. Click on Finish.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**3.4.2.Languages Used**

The programming language that was used in our Breast Cancer Detection project is Python. The implementation of source code was done through python. Python is an interpreted, interactive, object-oriented programming language which is suitable for implementing machine learning algorithms in easier way.

##### **Features of Python**

Python provides lots of features that are listed below:

##### **Easy to Learn and Use:**

Python is easy to learn and use. It is developer-friendly and high-level programming language.

##### **Expressive Language:**

Python language is more expressive means that it is more understandable and readable.

##### **Interpreted Language:**

Python is an interpreted language i.e. interpreter executes the code line by line at a line. This makes debugging easy and thus suitable for beginners.

##### **Cross-platform Language:**

Python can run equally on different platforms such as Windows, Linux, Unix and Macintosh etc. So, we can say that Python is a portable language.

##### **Free and Open Source:**

Python language is freely available at official web address. The source-code is

also available. Therefore, it is open source.

##### **Object-Oriented Language:**

Python supports object-oriented language and concepts of classes and objects come into existence.

##### **Extensible:**

It implies that other languages such as C/C++ can be used to compile the code and thus it can be used further in our python code.

##### **Large Standard Library:**

Python has a large and broad library and provides rich set of module and functions for rapid application development.

##### **GUI Programming Support:**

Graphical user interfaces can be developed using Python.

##### **Integrated:**

It can be easily integrated with languages like C, C++ and JAVA etc.

**CHAPTER-4**

# DESIGN

## 4.1.UML Introduction:

The unified modeling language allows the software engineer to express an analysis model using the modeling notation that is governed by a set of syntactic, semantic and pragmatic rules. A UML system is represented using five different views that describe the system from distinctly different perspective.

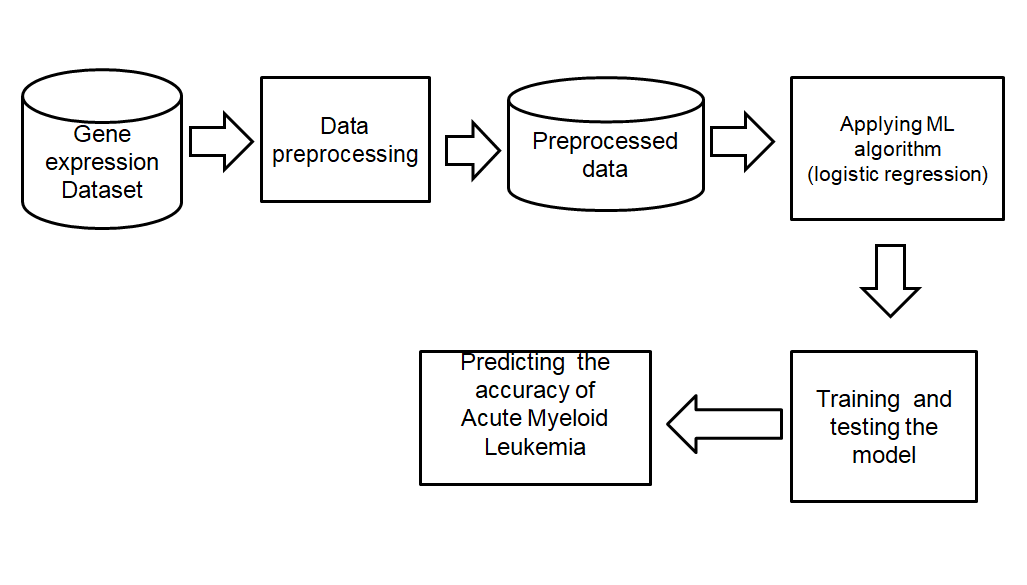
* UML is specifically constructed through two different domains, they are:
* UML Analysis modeling, this focuses on the user model and structural model views of the systems.
* UML Design modeling, which focuses on the behavioral modeling, implementation modeling and environmental model views.

**4.1.1 Usage of UML in Project**

As the strategic value of software increases for many companies, the industry looks for techniques to automate the production of software and to improve quality and reduce cost and time to the market. These techniques include component technology, visual programming, patterns and frameworks. Additionally, the development for the World Wide Web, while making some things simpler, has exacerbated these architectural problems. The UML was designed to respond to these needs. Simply, systems design refers to the process of defining the architecture, components, modules, interfaces and data for a system to satisfy specified requirements which can be done easily through UML diagrams.

**4.2.System Architecture:**

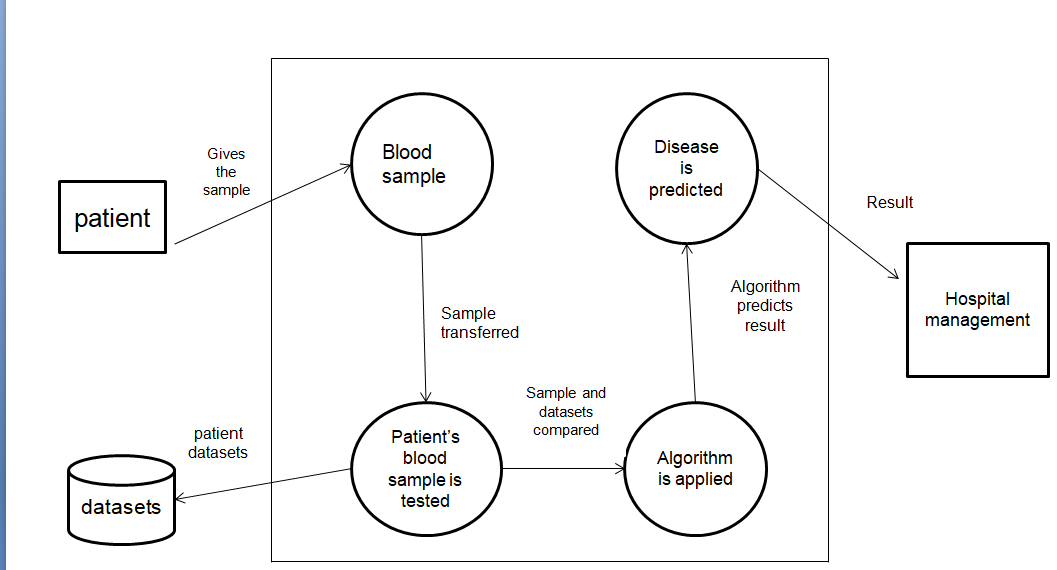
AML and ALL classification using machine learning classifies the patients who has AMl and ALL based on gene expression profiling .machine learning consist of various datasets through which we will compare the values of the user and predicts it, then the datasets are transformed into the smaller sets and from there it gets classified based on the classification algorithms later on the classified data is then processed into the machine learning technologies through which the data gets processed and goes in to the leukemia classification model using all the labels from the dataset taken. The preprocessed data combines and compares in the classification model of the system and finally classifies the leukemia. An architecture diagram is a graphical representation of a set of concepts, that are part of an architecture, including their principles, elements and components. The diagram explains about the system software in perception of overview of the system.



## Fig.4.1 System Architecture

## 4.3. Data Flow Diagram

## A data-flow diagram is a way of representing a flow of a data of a process or a system (usually an information system). This also provides information about the outputs and inputs of each entity and the process itself. A data-flow diagram has no control flow, there are no decision rules and no loops. Specific operations based on the data can be represented by a flowchart. The data-flow diagram is part of the structured-analysis modeling tools. When using UML, the activity diagram typically takes over the role of the data-flow diagram.



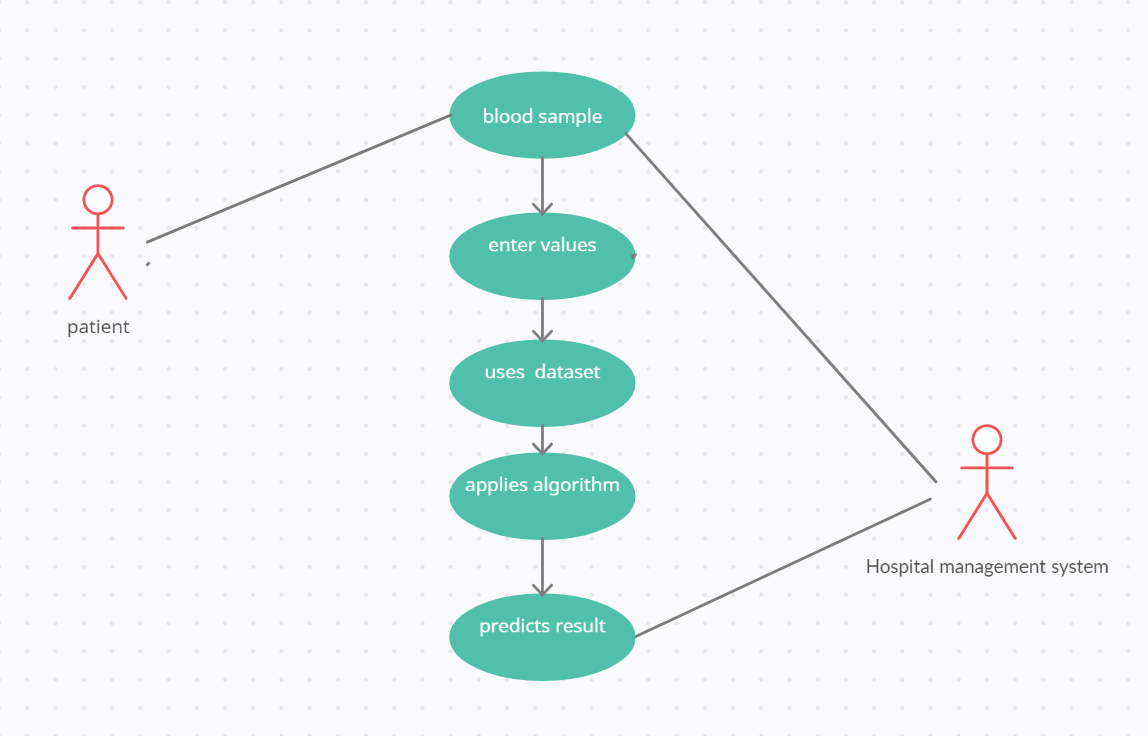
**Fig.4.2 Data Flow Diagram**

The above data flow diagram describes how breast cancer prediction can be done. Initially obtain the datasets from any websites or generate own datasets. After obtaining the datasets, perform data transformation to it in such a way that there shouldn’t be any integration problem or any redundancy issue.

Now, apply feature selection techniques to the Classification of ALL and AMl. Then a subset of features which are most important in the prediction of breast cancer to be malignant or benign are obtained. choose the classifier that gives the best possible accuracy with the subset of features obtained after feature selection. Applying the classifiers to the dataset actually mean that needs to train the model with the classifiers and test the data so that the model will be fit.

##### **4.4.Use Case Diagram**

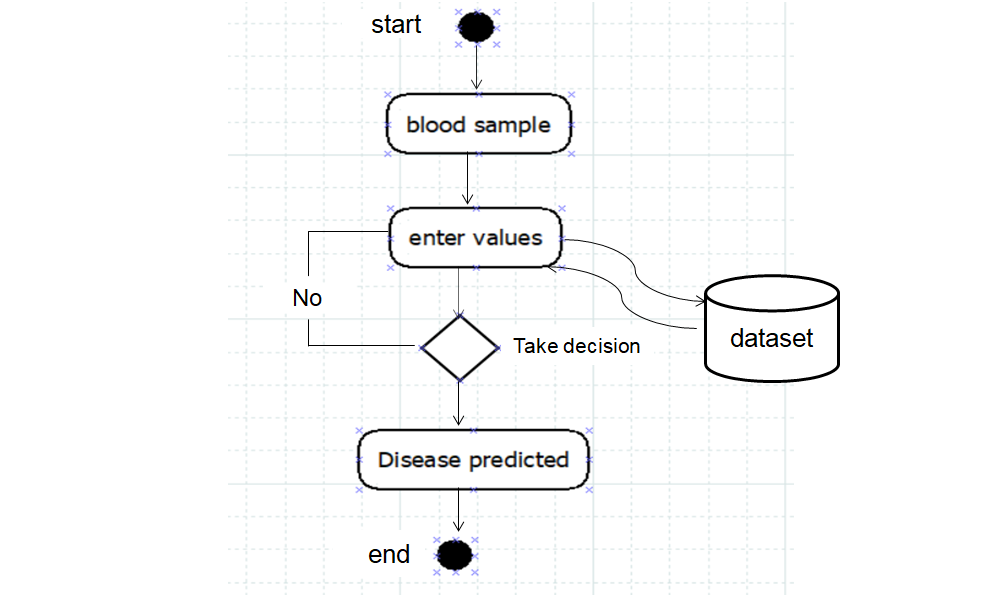
The Use Case diagram of the project classifying AML and ALL based on Gene Expression profiling using machine learning consist of all the various aspects a normal use case diagram requires. This use case diagram shows how from starting the model flows from one step to another, like patient has to give his blood sample and all the information’s and all other general information along with the values that goes into the system compares with the prediction model and if true is predicts the appropriate results otherwise it shows the details where the user if gone wrong while entering the information’s and it also shows the appropriate precautionary measure for the user to follow. Here the use case diagram of all the entities are linked to each other where the user gets started with the system.



**Fig.4.3 Use Case Diagram**

**4.5.State Chart Diagram**

A State chart diagram describes the behaviour of a single object in response to a series of events in a system. Sometimes it's also known as a Harel state chart or a state machine diagram. This UML diagram models the dynamic flow of control from state to state of a particular object within a system. It is similar to activity diagram but here there are only few rules like how it starts and how it end all are denoted with the help of the symbol given below, the system starts with blood sample of the patient.Taking the blood sample and identifying the gene description and gene access number,The patients will be classified into AML or ALL class by comparing the dataset values and calculations .



**Fig 4.4. State Chart Diagram**

**4.6.Steps involved in Design:**

* Data Collection
* Data Preprocessing
* Feature Selection
* Applying ML algorithms(Random forest classifier)

Each step has its own specific reason and plays prominent role in building up a model of the project. Each step has been explained in detail in implementation part.

**CHAPTER 5**

**IMPLEMETATION**

Here, in our project, feature selection is made from 7130 features of Gene Expression dataset by using different wrapper feature selection methods of Machine learning. Models are constructed with the selected features that are generated from different selection methods by using Llogistic Regression Algorithm and for classification comparative analysis Support Vector Machines is taken. Accuracy is calculated for each model and the best model with best features is selected as result based on high accuracy.

**5.1.Libraries Used**

Python is increasingly being used as a scientific language. Matrix and vector manipulation are extremely important for scientific computations. Both NumPy and Pandas have emerged to be essential libraries for any scientific computation, including machine learning, in python due to their intuitive syntax and high-performance matrix computation capabilities.

#### NumPy:

NumPy stands for ‘Numerical Python’ or ‘Numeric Python’. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Since, arrays and matrices are an essential part of the Machine Learning ecosystem, NumPy along with Machine Learning modules like Scikit-learn, Pandas, Matplotlib, TensorFlow, etc. complete the Python Machine Learning Ecosystem.

NumPy provides the essential multi-dimensional array-oriented computing functionalities designed for high-level mathematical functions and scientific computation. NumPy can be imported into the notebook using import numpy as np.

##### **Pandas**:

Similar to NumPy, Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Pandas provides in-memory 2d table object called Data frame. It is like a spreadsheet with column names and row labels.

Hence, with 2d tables, pandas are capable of providing many additional functionalities like creating pivot tables, computing columns based on other columns and plotting graphs. Pandas can be imported into Python using:import pandas as pd.

#### pip:

The pip command is a tool for installing and managing Python packages, such as those found in the Python Package Index. It's a replacement for easy install. The easiest way to install the nfl\* python modules and keep them up-to-date is with a Python-based package manager called [Pip](http://en.wikipedia.org/wiki/Pip_(package_manager)).

**Sklearn:**

Skikit-learn is a free software machine library for Python programming language.It features various classification , regression and clustering algorithms including support vector machine, random forest, k-means and gradient boosting. In our project we have used different features.

* from sklearn.model\_selection import train\_test\_split:

Used for Splitting the dataset into Training and Testing.

* from sklearn.svm import SVC

Used to import Support vector Machines.

* from sklearn.linear\_model
* import LogisticRegression
* from sklearn.metrics import recall\_score, precision\_score, classification\_report, accuracy\_score, confusion\_matrix, roc\_curve, auc,roc\_curve, accuracy\_score, plot\_confusion\_matrix
* from sklearn.preprocessing import StandardScaler, normalize
* from sklearn.decomposition import PCA
* from scipy import ndimage
* import seaborn as sns

**5.2 Implementation:**

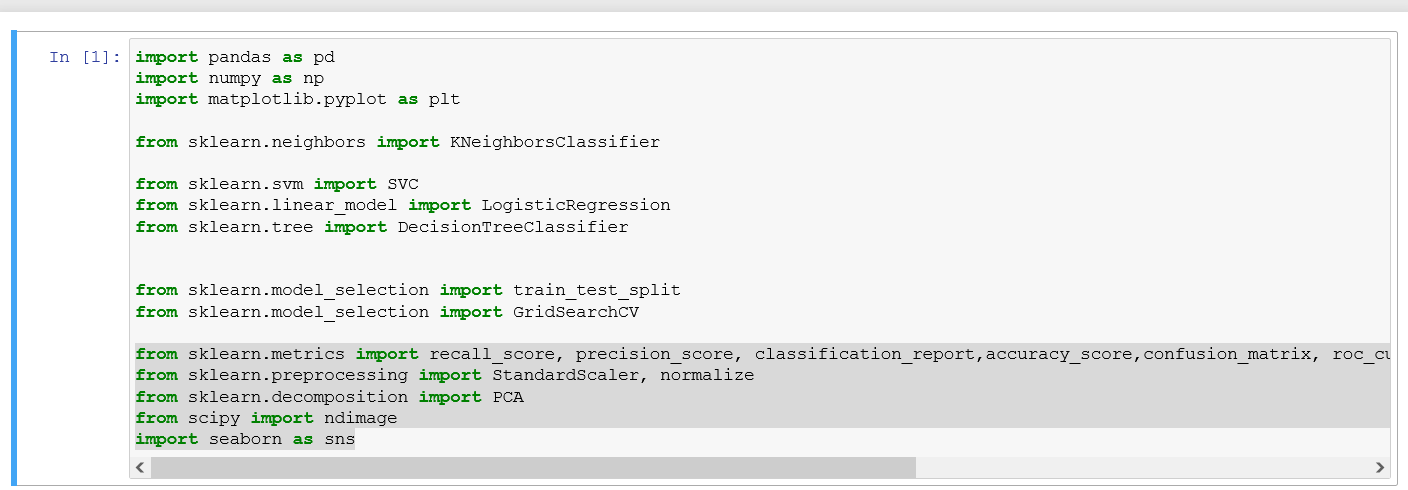
Required dataset should be present in a folder .The dataset we have taken is Gene Expression dataset.

**5.2.1.Dataset description:**

As mentioned, the dataset used in this project contains microarray gene expression data of patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The data used in the paper was originally split into two datasets of 38 and 34 samples, used respectively for training and testing. In addition, we had one more dataset with the labels (ALL or AML) for all samples. The first two columns of the dataset contain gene descriptions. The dataset contains microarray expression data from 7129 genes, where each gene is a row in the dataset. Intensity values have been re-scaled such that overall intensities for each chip are equivalent. The samples are the columns of the dataset. Next to the column for each sample there was a column of a variable named "Call", which took three values - "A", "P", and "M". These stand for "Absent", "Present", and "Marginal". They are based on the signal in the microarray for a gene at hand and they determine a gene's expression in a sample. In order for XGBoost to handle these categorical variables, these were transformed into numbers - 0 for "A", 1 for "P", and 2 for "M". We first trained classifiers only with the numerical variables, but they did not perform very well, and thus we trained the classifiers with the numerical and some of the categorical features.

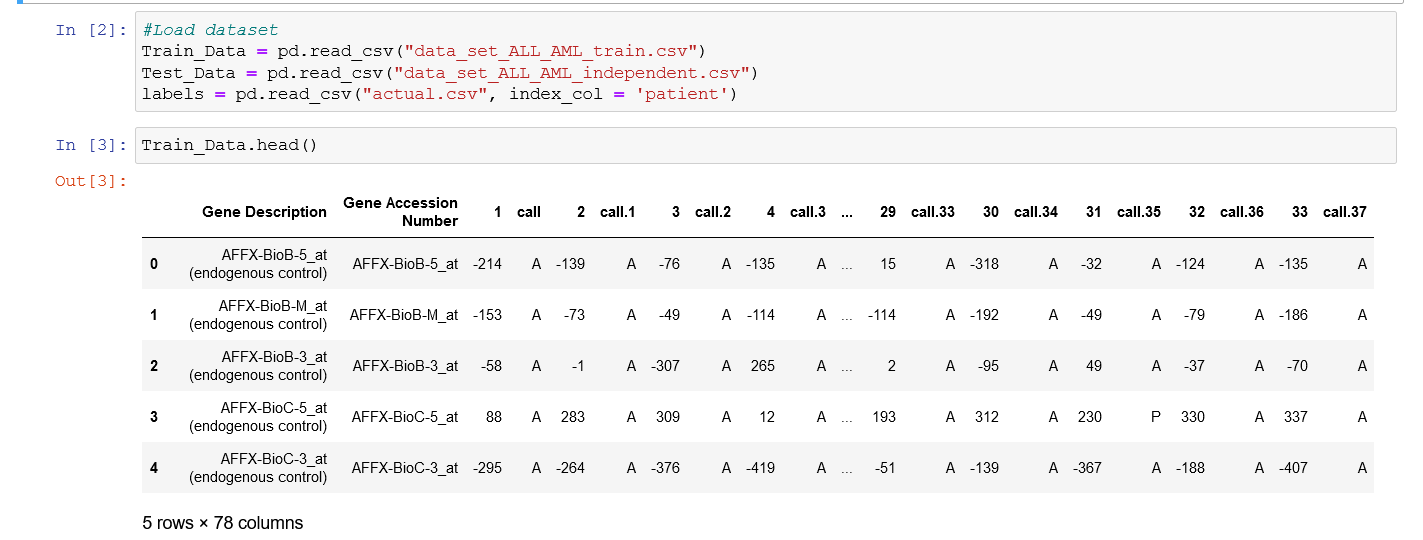
**5.2.2. Execution**

Firstly we have to load all the required fit libraries.



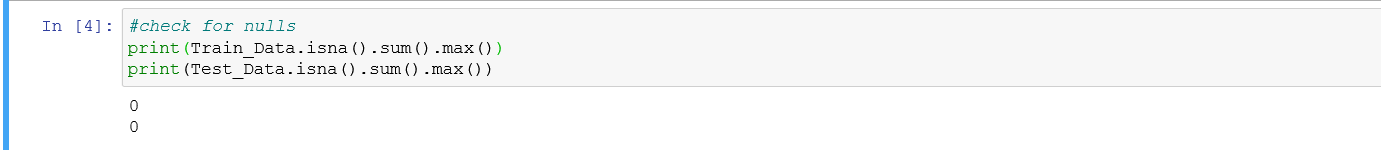
**Fig.5.1 importing libraries**

Next , We have to load the datset and train the data head to read first n rows of the dataset



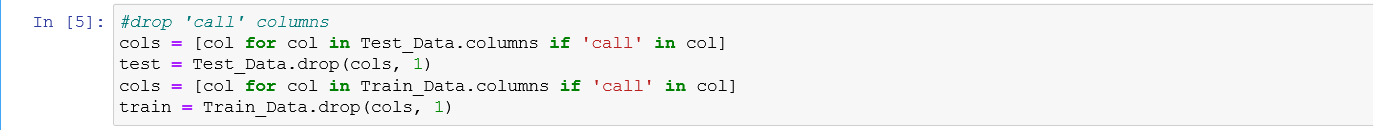
**Fig.5.2 output of first n rows**

Next step is the Preprocessing part .As part of this part firstly we check for null values in our dataset.



**Fig.5.3 checking for nulls**

# Next , we have to drop the column Call In the dataset .The “call” Column represents 3 alphabets such as “A” , “B”, “M” saying as Absent , Present and Marginal respectively.



**Fig.5.4 dropping “call” colum**

As per the next step , We join all the data .



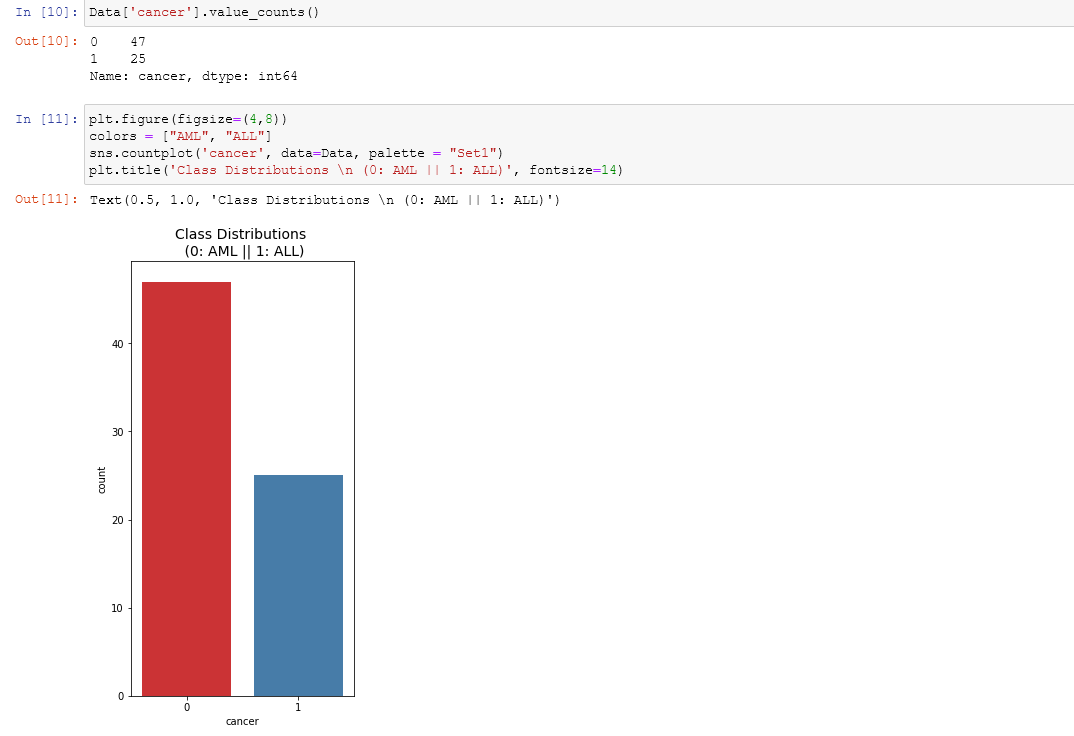
**Fig.5.5 joining all the data**

Next, Transpose the rows and columns , encode the categorical values and display the data .



**Fig.5.6 Tranpose of rows and couluns**

Next , lets try to count the patients of leukemia by representing “0” as AML and “1” as ALL and also try to plot the classification of AML and ALL .



**Fig.5.7 plot of class distribution**

We can clearly see that the labels are not balanced and we need to balance them before building our classification models.

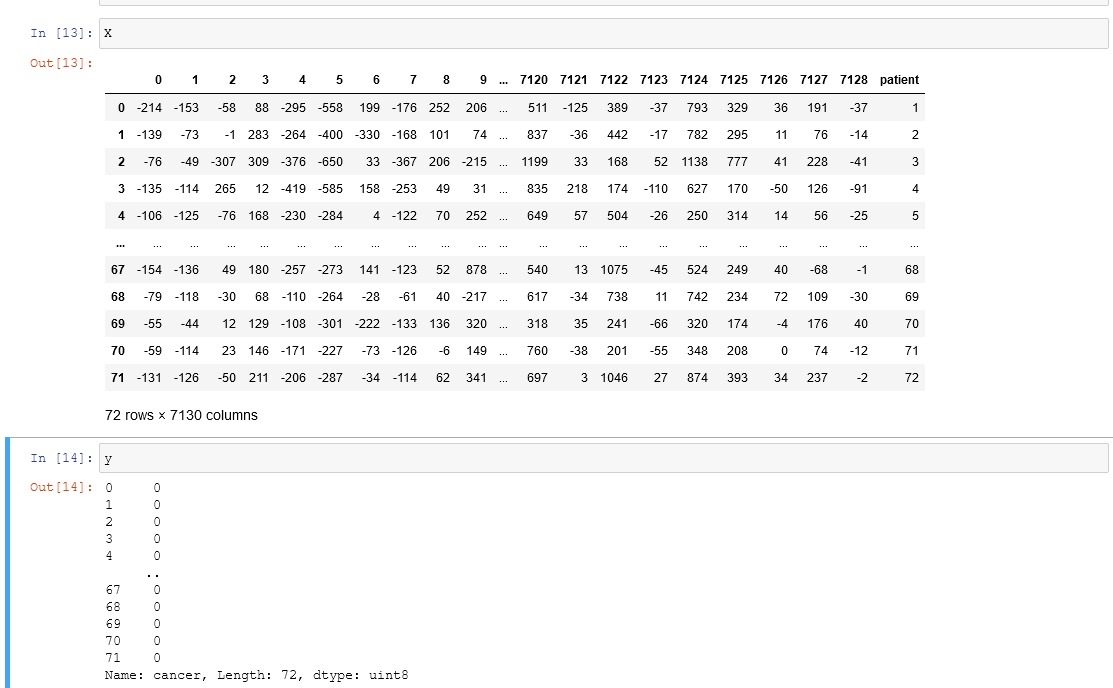
So, inorder to balance the classes and apply any Machine learning model we do the following steps.



**Fig.5.8 categorical variables**

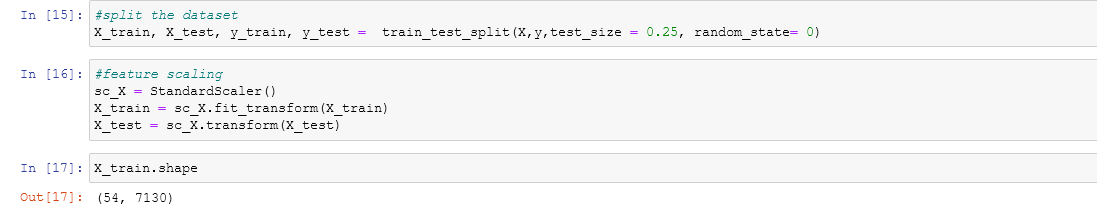
Next as per the implementation process, we should split the dataset into training data and testing data into 75% and 25% respectively. We have to take the categorical values x and y .

Displaying X (matrix independent Variable ) and y (dependent variable)

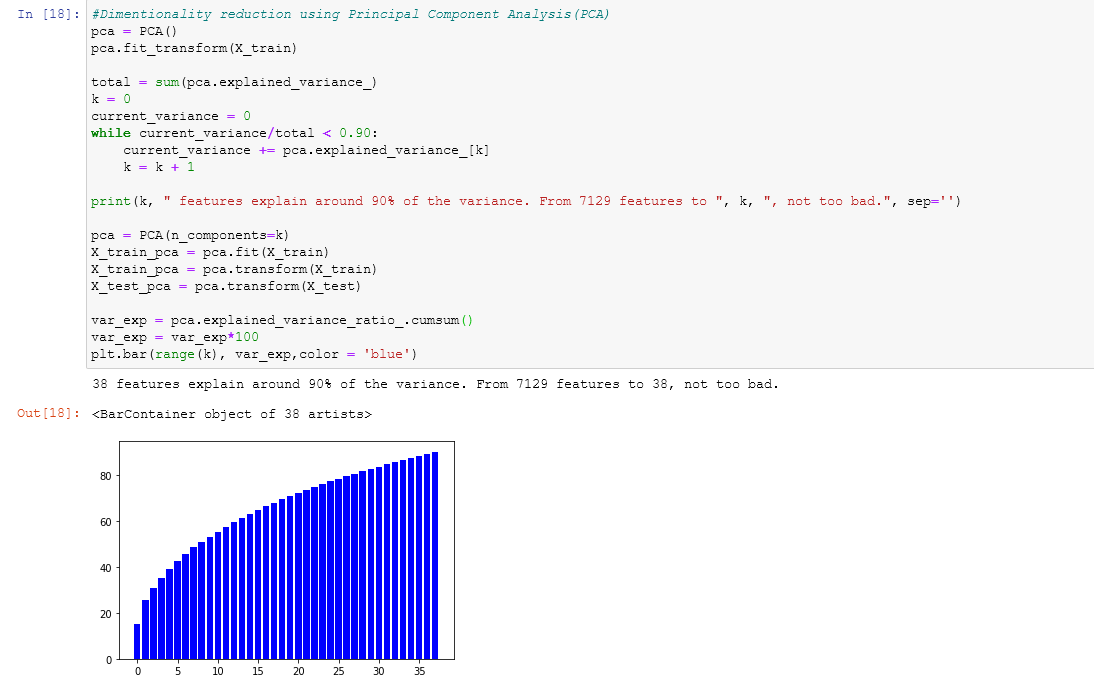


**Fig.5.9 output of categorical values**

Next, Splitting the dataset and feature scaling is done inorder to standardize the features present in the dataset into fixed range.

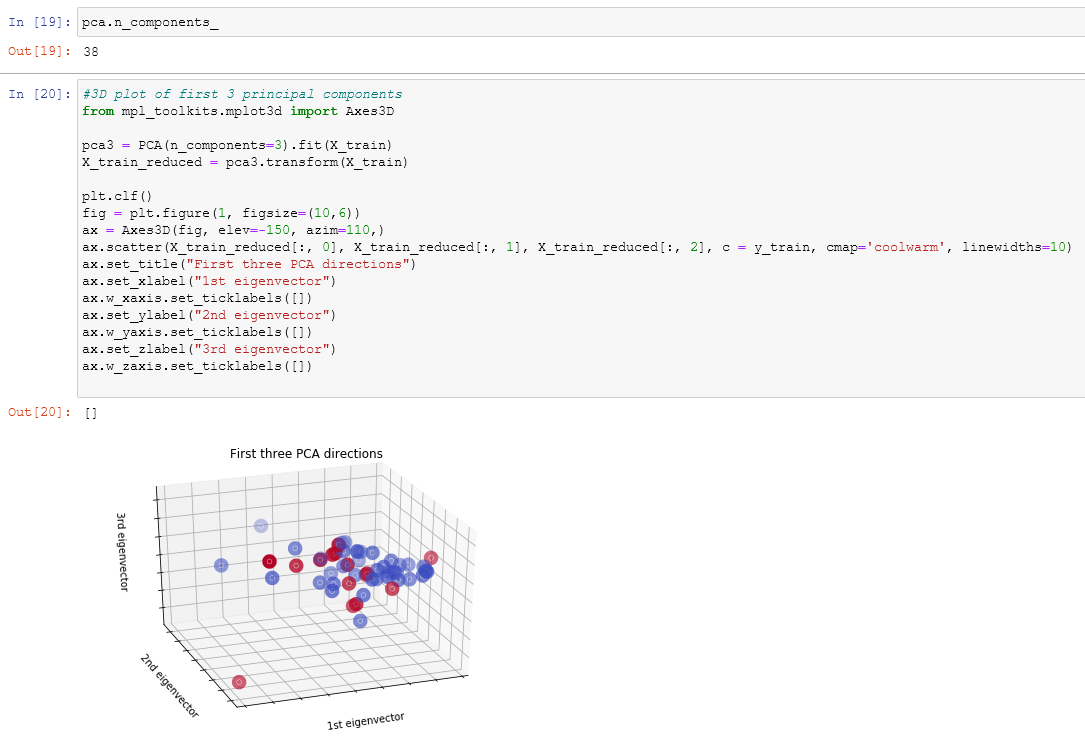


**Fig.5.10 spliting the dataset**

Next ,we are trying to do Dimensionality Reduction which is a Feature extraction technique to solve the problem such as over fitting.

**Fig.5.11 Dimensionality Reduction**

After that, lets just count the Principle components reduced from the 7130 features and plot a 3d graph to show first 3 principle components.



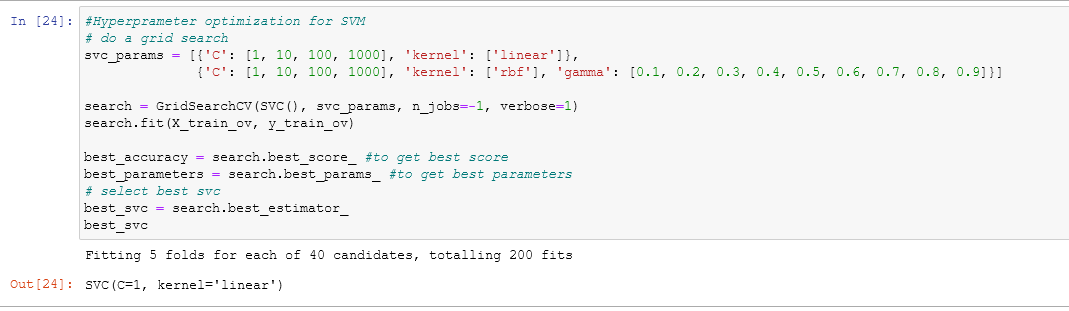
**Fig.5.12 3d plot of 3 pca**

Next , We have to sampling.Here We balance the labels using SMOTE analysis. SMOTE is Synthetic Minority Over-Sampling Technique .We can define a SMOTE instance with default parameters that will apply it in one step to create a transformed version of our dataset.



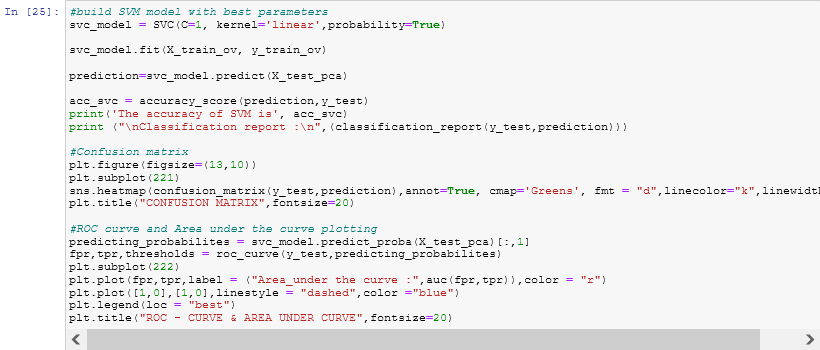
**Fig.5.13 sampling**

Now, We can Apply our required Model from Machine learning models .Here for comparative Analysis hyperparameter optimization is done using support vector machines. Hyper parameter optimization is for tuning a problem .Choosing the right set of values is normally known hyper parameter optimization .



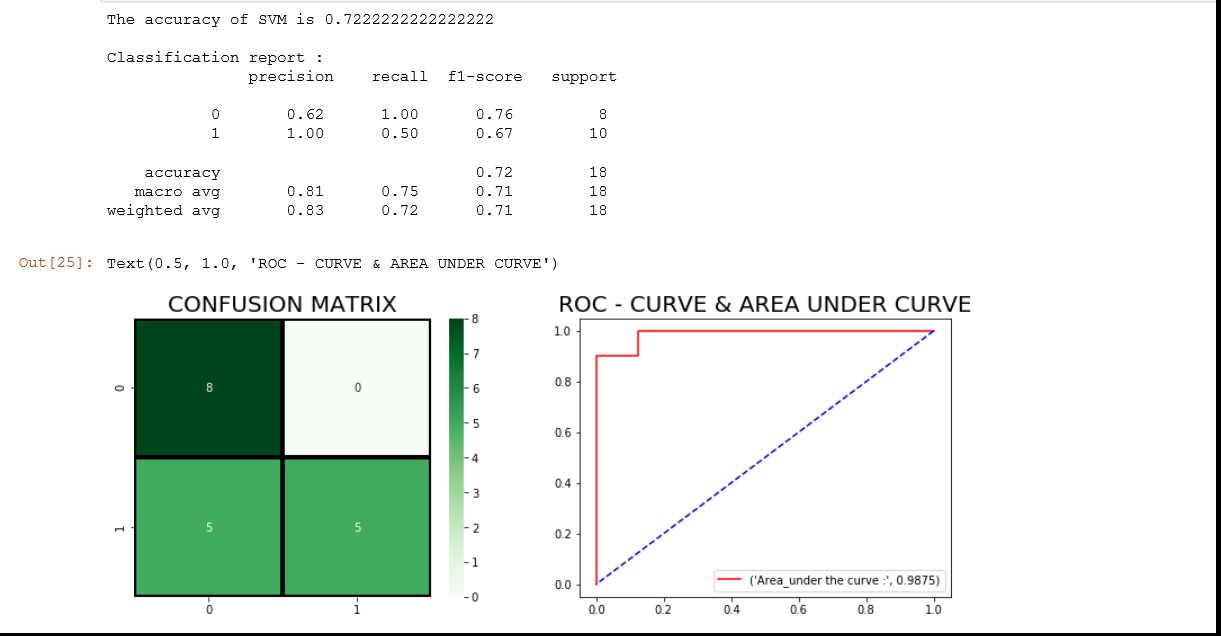
**Fig.5.14 hyperparameter optimization of SVM**

After choosing the best parameters we can build the Support Vector Machine model.



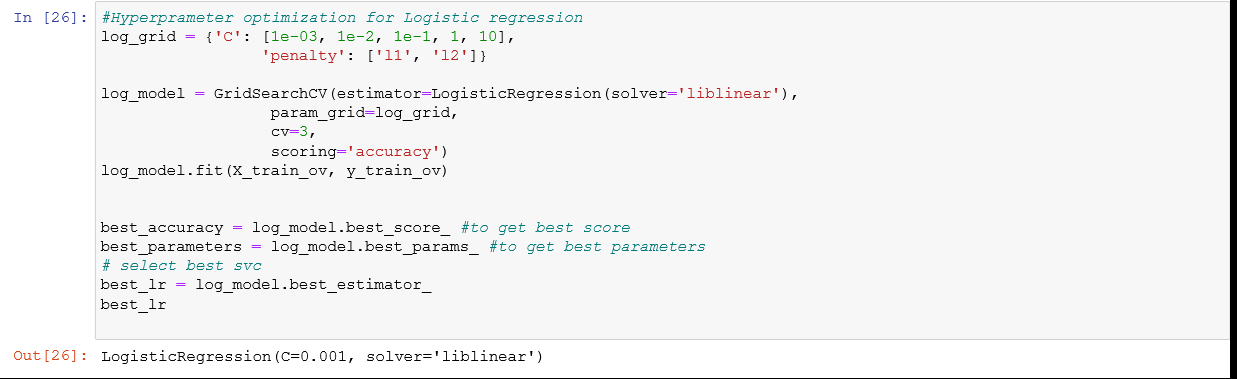
**Fig.5.15 SVM model**

The output will be displayed as figure below.



**Fig.5.16 output using SVM**

Next , Hyperparameter optimization for Logistic Regression .



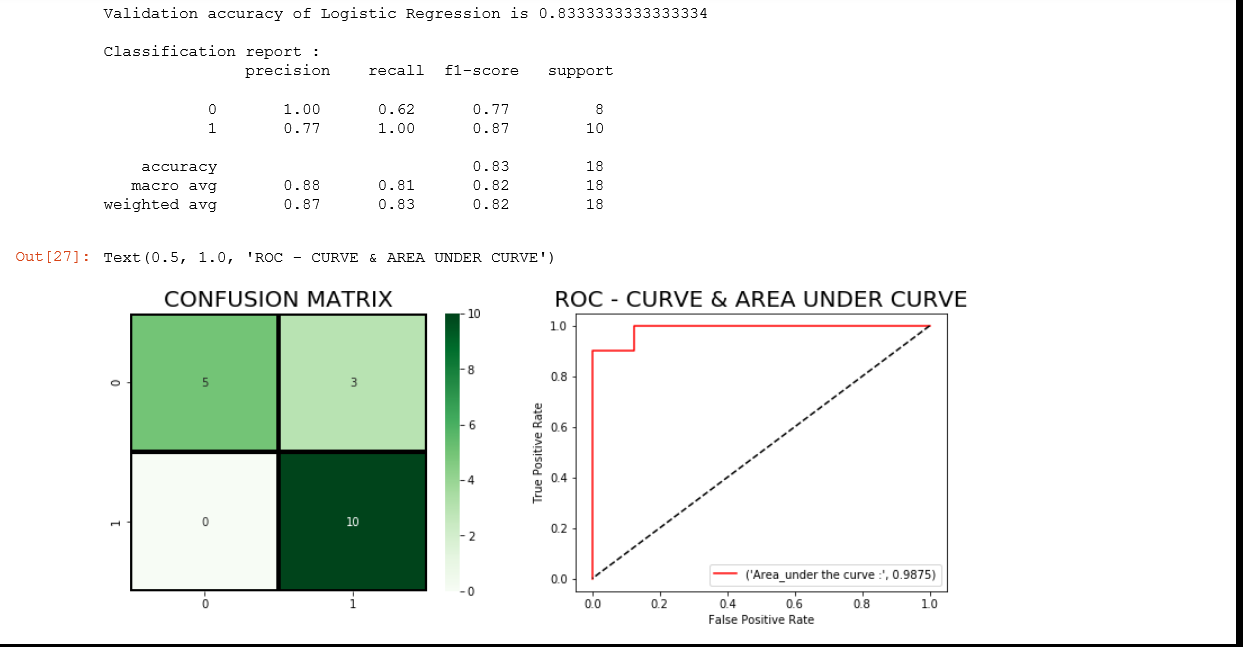
**Fig.5.17 hyperparameter optimization of Logistic regression**

After choosing the best parameters, we can build a model using logistic regression.



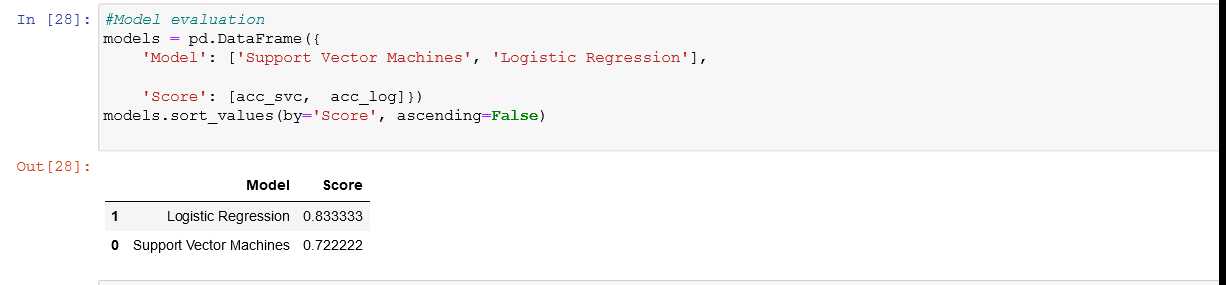
**Fig.5.18 Logistic Regression model**

The output is shown below in form of classification report ,confusion matrix, Roc curves etc.,



**Fig.5.19 output using Logistic regression**

Finally , the Model evaluation is done based on comparative analysis between Logistic Regression and Support Vector Machines.



**Fig.5.20 model evaluation**

Logistic regression gave 83% accuracy than Support vector Machines which gave 72.2%

# CHAPTER 6

# TESTING

The main objective of testing is to uncover a host of errors, systematically and with minimum effort and time. Stating formally, we can say,

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

* A successful test is one that uncovers an as yet undiscovered error.
* A good test case is one that has a high probability of finding error, if it exists.

The first approach is what known as Black box testing and the second approach is White box testing. We apply white box testing techniques to ascertain the functionalities top-down and then we use black box testing techniques to demonstrate that everything runs as expected.

**6.1.Black-Box Testing:**

This technique of testing is done without any knowledge of the interior workings of the application The tester is oblivious to the system architecture and does not have access to the source code. Typically, while performing a black-box test, a tester will interact with the system’s user interface by providing inputs and examining the outputs without knowing how and where the inputs are worked upon.

* Well suited and efficient for large code segments
* Code access is not required
* Clearly separates user’s perspectives from the developer’s perspective through visibly defined roles

##### **6.2.White-Box Testing:**

White-box testing is the detailed investigation of internal logic and structure of the code. It is also called “glass testing” or “open-box testing”. In order to perform white- box testing on an application, a tester needs to know the internal workings of the code.

The tester needs to look inside the source code and find out which part of the code is working inappropriately.

In this, the test cases are generated on the logic of each module. It has been uses to generate the test cases in the following cases:

* + - Guarantee that all independent modules have been executed.
    - Execute all logical decisions and loops.
    - Execute through proper plots and curves.

**6.3.Performance Evaluation**

This project has been successfully executed its source code. Initially there were some errors in the code. By resolving them, the code is fully free from errors and bugs.

After performing feature selection models are build using different classifiers. The classifier that gives best possible accuracy has been considered in this project. The classifier used to build a model in this project is Logistic Regression classifier which is gave the effective results. Hence, for the considered dataset, the Logistic regression has performed well with the accuracy of 83.3% compared to Support vector machines of accuracy 72.2%.

# CONCLUSION

Cancer has been portrayed as a heterogeneous disease comprising of a wide range of subtypes. The early diagnosis of a cancer type is very important to determine the course of medical treatment required by the patient. The significance of classifying cancerous cells into benign or malignant has driven many research studies. In the past years researchers have been encouraged to use different machine learning (ML) techniques for leukemia classification , as well as prediction of survivability and recurrence. In this project we successfully developed a model that can classify AML and ALL .

Here we have used machine learning algorithms like Logistic regression and Support Vector Machines Algorithm to classify Leukemia patients and also the type of cancer. Given a few parameters, our algorithms can classify a good model with comparative analysis of Logistic regression and Support ector Machines . During our comparative analysis ,we found the model using logistic regression is better than Support Vector Machines. We can fine and tune these models even further to get even better accuracy.

Integration of multi dimensional features can give more effective tools for detection of cancer. Other machine learning models like support vector machine, other models of neural networks (CNN or ANN) could be implemented. Other learning algorithms can be applied with using our chosen set of features. A dataset with more number of examples can be used.

**BIBLIOGRAPHY**

[1] J. Laosai and K. Chamnongthai. "Acute leukemia classification by using SVM and K-Means clustering" Proceedings of the International Electrical Engineering Congress pp. 1-4. 2014 [2] ―Types of leukemia‖15 Nov. 2019

[2] “Applications of Machine Learning in Acute Myeloid Leukemia Prediction” Joanna Bitton and Erica Dominic New York University jtb470@nyu.edu, [ehd255@nyu.edu](mailto:ehd255@nyu.edu) [2018]

[3] “Machine learning to predict 5-year survival among pediatric Acute Myeloid Leukemia patients and development of OSPAM-C online survival prediction tool” by Ashis Kumar Das1 , Shiba Mishra and Saji Saraswathy Gopalan [April,2020]

[4] “Scalable Prediction of Acute Myeloid Leukemia Using High- Dimensional Machine Learning and Blood Transcriptomics” by Stefanie WarnatHerresthal, Konstantinos Perrakis, Bernd Taschler.[Jan,2020]

[5] Beam A, Kohane I. Big data and machine learning in health care. *JAMA*. 2018; **319**: 1317- 1318.[Crossref](https://onlinelibrary.wiley.com/servlet/linkout?suffix=null&dbid=16&doi=10.1111%2Fijlh.13089&key=10.1001%2Fjama.2017.18391) [PubMed](https://onlinelibrary.wiley.com/servlet/linkout?suffix=null&dbid=8&doi=10.1111%2Fijlh.13089&key=29532063) [Web of Science®](https://onlinelibrary.wiley.com/servlet/linkout?suffix=null&dbid=128&doi=10.1111%2Fijlh.13089&key=000429035900010)[Google Scholar](http://scholar.google.com/scholar_lookup?hl=en&volume=319&publication_year=2018&pages=1317-1318&journal=JAMA&author=A+Beam&author=I+Kohane&title=Big+data+and+machine+learning+in+health+care)

[6] Machine Learning Algorithms For Diagnosis Of Leukemia by Italia Joseph Maria, T. Devi, D. Ravi Kantarjian et al. Blood Cancer Journal (2021) 11:41 [https://doi.org/10.1038/s41408-021-00425-3](https://doi.org/10.1038/s41408-021-00425-3%20%5b6)

[7] Leukemia Prediction Using Sparse Logistic Regression Tapio Manninen1 \*, Heikki Huttunen1 , Pekka Ruusuvuori1 , Matti Nykter May 3, 2013; Accepted July 12, 2013; Published August 30, 2013