Comparative study of Machine Learning for Classification of Cancer Types using Gene Expressions Dataset

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE

BACHELOR OF ENGINEERING IN
(Information Science and Engineering)

Submitted By:

Aditi N Sirigeri-01JST16IS002
Nandini V-01JST16IS055
Neha S-01JST16IS057
Siddharth Gokarn-01JST16IS064

Supervisor Name:

Prof. D S Vinod
Associate Professor

JSS MAHAVIDYAPEETHA

JSS SCIENCE AND TECHNOLOGY UNIVERSITY
MYSURU

2019-20

DECLARATION

We declare that this written submission represents our ideas in our own words and where other's ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not mispresented or fabricated or falsifies any idea/data/fact/source in our submission.

PROJECT GROUP

Aditi N Sirigeri
Nandini V
Neha S
Siddharth Gokarn

JSS SCIENCE AND TECHNOLOGY UNIVERSITY

Department of Information Science and engineering

CERTIFICATE

This is to certify that following students Aditi N Sirigeri, Nandini V, Neha S, Siddharth Gokarn of JSS SCIENCE AND TECHNOLOGY UNIVERSITY, Department of Information Science and Engineering have successfully completed the project work titled "Comparative study of Machine Learning for Classification of Tumor Types using Gene Expressions Dataset" during the year 2020.

Name of the Students

Aditi N Sirigeri	01JST16IS002
Nandini V	01JST16IS055
Neha S	01JST16IS057
Siddharth Gokarn	01JST16IS064

Internal Examiner External Examiner Head of the Department

ACKNOWLEDGEMENT

We would like to thank all the people whose support and encouragement made this project a possibility. We would like to especially thank our Final Year Project supervisors, D S Vinod and Project Representative C K Roopa, who gave us continuous guidance, assistance, and inspiration to continue efficiently working on our project and obtain promising results.

ABSTRACT

The classification of different tumor types is of great importance in cancer diagnosis. However, most previous cancer classification studies are clinical-based and have limited diagnostic ability. Cancer classification using gene expression data is known to contain the keys for addressing the fundamental problems relating to cancer diagnosis. Cancer classification based on molecular level investigation has gained the interest of researches as it provides a systematic, accurate and objective diagnosis for different cancer types. The recent advent of DNA microarray technique has made simultaneous monitoring of thousands of gene expressions possible. With this abundance of gene expression data, researchers have started to explore the possibilities of cancer classification using gene expression data.

Machine Learning (ML) techniques can be used to develop tools for physicians that can be used as an effective mechanism for early predication and diagnosis of cancer which will greatly enhance the survival rate of patients. This paper compares the most popular ML techniques commonly used for cancer prediction. The gene expression data set was used as a training set to evaluate and compare the performance of the different ML classifiers in terms of key parameters such as accuracy, recall, precision. We also introduce various proposed gene selection methods to get accurate cancer classification.

Table of Contents

Chapter 1. INTRODUCTION

1.1 Problem Statement

A comparative study of various algorithms for classification of different tumor types using different Gene Expression datasets.

1.2 General Introduction

Machine Learning (ML), is a subfield of Artificial Intelligence (AI) that allows machines to learn without explicit programming by exposing them to sets of data allowing them to learn a specific task through experience. Over the last few decades, ML methods have been widespread in the development of predictive models in order to support effective decision-making. In cancer research, these techniques could be used to identify different patterns in a data set.

Cancer research is one of the major research areas in the medical field. Accurate prediction of different tumor types has great value in providing better treatment and toxicity minimization on the patients. Previously, cancer classification has always been morphological, and clinical based. These conventional cancer classification methods are reported to have several limitations in their diagnostic ability. It has been suggested that specifications of therapies according to tumor types differentiated by pathogenetic patterns may maximize the efficacy of the. Also, the existing tumor classes has been found to be heterogeneous and comprises of diseases that are molecularly distinct and follow different clinical courses.

In order to gain a better insight into the problem of cancer classification, systematic approaches based on global gene expression analysis have been proposed. The expression level of genes is known to contain the keys to address fundamental problems relating to the prevention and cure of diseases, biological evolution mechanisms and drug discovery. The recent advent of microarray technology has allowed the simultaneous monitoring of thousands of genes, which motivated the development in cancer classification using gene expression data. Though still in its early stages of development, results obtained so far seemed promising.

Gene Expression and DNA Microarray Technology

For a classification of cancer by using gene expression data, first we have to know some fundamental knowledge in molecular biology. Cells are the fundamental working units of every living system. All the instructions needed to direct their activities are contained within the chemical deoxyribonucleic acid or DNA. A DNA molecule is a double-stranded polymer composed of four basic molecular units called nucleotides. DNA from all organisms is made up of the same chemical and physical components. Each nucleotide comprises a phosphate group, a deoxyribose sugar and one of the four nitrogen bases. The nitrogen bases are adenine (A), guanine (G), cytosine(C) and thymine (T). DNA sequence is a arrangement of the base pairs in the DNA strand. The entire DNA sequence that codes for a living thing is called its genome. The genome is an organism's complete set of DNAs. A gene is a small, defined section of the entire genomic sequence, each has a specific and unique purpose. There are three types of genes, namely the protein-coding genes, the RNA-specifying genes and the un transcribed genes.

DNA act as a template for making copies of itself but also as a blueprint for a molecule called RNA (ribonucleic acid). The genome provides a template for the synthesis of a variety of RNA molecules. The main types of RNA are messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA).The expression of the genetic information stored in the DNA molecule occurs in two stages: (i) transcription stage where the DNA molecule is transcribed into mRNA, (ii) translation stage where mRNA is translated into the amino acid sequences of the proteins that perform various cellular functions. The process of transcribing a gene's DNA sequence into RNA is called gene expression. A gene's expression level indicates the approximate number of copies of that gene's RNA produced in a cell and it is correlated with the amount of the corresponding proteins made. It has been shown that specific patterns of gene expression occur during different biological states such as embryogenesis, cell development, and during normal physiological responses in tissues and cells. Thus the expression of a gene provides a measure of activity of a gene under certain biochemical conditions. It is known that certain diseases, such as cancer, are reflected in the change of the expression values of certain genes. Normal cells can evolve into malignant cancer cells through a series of mutations in genes that control the cell cycle, apoptosis and genome integrity, etc.

Studies on the use of DNA microarrays have supported the effectiveness of gene expression patterns for identifying different gene functions and cancer diagnosis. Microarrays and serial analysis of gene expressions are two recent technologies for measuring the thousands of genome-wide expression values in parallel. The former, which consists of cDNA microarrays and high-density oligonucleotide arrays measures

the relative levels of mRNA abundance between different samples, while the latter measures the absolute level.

CDNA microarray analysis is a relatively new molecular biology method that expands on classic probe hybridization methods to provide access to thousands of genes at once. Therefore, allowing the recording of expression levels of thousands of genes simultaneously. CDNA microarrays consists of thousands of individual DNA sequences printed in a high-density array on a glass microscope. Each data point produced by a DNA microarray hybridization experiment represents the ratio of expression levels of a gene under two different experimental conditions. The result, from an experiment with m genes on a single chip, is a series of m expression level ratios. The numerator of the ratio is the expression level of the gene in the varying conditions of interest, and the denominator is the expression level of the gene in some reference condition.

Serial analysis of gene expression, or SAGE, is a technique designed to take advantage of high- Through put sequencing technology to obtain a quantitative profile of cellular gene expression. The SAGE technique does not measure the expression level of a gene but quantifies a “tag” which represents the transcription product of that gene. A tag in this case is a nucleotide sequence of defined length. The original length of the tag was nine bases, current SAGE protocols produce a ten to eleven base tag. The data product of the SAGE technique is a list of tags, with their corresponding count values, which is a digital representation of cellular gene expression. Most existing cancer classification methods uses DNA microarray expression data.

Gene expression level represents the amount of RNA produced in a cell under different biological states. So, during cell division process, if the cells suffer from diseases -i.e. cancer or malignant tumors- that cause alteration or mutations in genes, the uncontrollable behavior of the gene will be transmitted to daughter cells. Moreover, certain gene expression values will be affected and hence expression levels can be realized by monitoring the RNA. The expression levels of thousands of genes can be simultaneously measured under experimental environments and conditions due to the significant advancement of DNA microarray technology. This technology made it possible to understand life on the molecular level and enables to generate large-scale gene expression data. Besides, it has led to many analytical insights because it produced large amount of gene data ready to be analyzed rapidly and precisely by managing them using several statistical and machine learning processes.

A variety of different ML techniques and feature selection algorithms have been widely applied to disease prognosis and prediction Most of these works employ ML methods

for modelling the progression of cancer and identify informative factors that are utilized afterwards in a classification scheme. Different classification methods from statistical and machine learning area have been applied to cancer classification. The gene expression data is very different from any of other data. First, it has very high dimensionality, usually contains thousands to tens of thousands of genes. Second, publicly available data size is very small. Third, most genes are irrelevant to cancer distinction. It is obvious that those existing classification methods were not designed to handle this kind of data efficiently and effectively. Some researchers proposed to do gene selection prior to cancer classification. Performing gene selection helps to reduce data size thus improving the running time. More importantly, gene selection removes a large number of irrelevant genes which improves the classification accuracy. Due to the important role it plays in cancer classification, we also study the proposed gene selection methods.
In this paper, popular ML techniques applied to a gene expression data set are investigated and compared. These techniques are Support Vector Machine (SVM), Naïve Bayes, Gradient Boosting and many other.

1.3 Application

Cancer is one of the deadliest diseases for humans. According to the WHO (2015), cancer is the causes of the death number two in the world by 13 % after cardiovascular disease. Cancer often causes death if treatment is too late. Therefore, early detection of cancer is necessary to avoid the spread of cancer. Machine learning is particularly well-suited to medical applications, especially those that depend on complex proteomic and genomic measurements. As a result, machine learning has been applied to cancer prognosis and prediction. Various clinical or pathological factors have been evaluated as prognosis factors. For example, the treatment of cancer is often based on factors such as age, lymph node status, tumor size, etc. Although these factors provide valuable information about the risk of recurrence, they are generally considered to be insufficient to predict individual patient outcomes and determine an individual patient's need for systematic adjuvant therapy.

Recent advances in biotechnologies allow us to generate various types of molecular data for the same sample, e.g. copy number aberrations as measured by array CGH, mRNA expression, SNPs, methylation, etc. Each of these distinct data types provides one view of the molecular machinery of the cancer cell. Molecular data allows for adding information to the analysis of biological phenotypes. Studying the characteristics of thousands of genes simultaneously offered a deep insight into cancer classification problem. It introduced an abundant amount of data ready to be explored. It has also been applied in a wide range of applications such as drug discovery, cancer prediction and

diagnosis which is a very important issue for cancer treatment. Besides, it helps in understanding the function of genes and the interaction between genes in normal and abnormal conditions.

1.4 Challenges

There have been extensive studies done in the past on the classification problem by the statistical, machine learning and database research community. But gene classification as a new area of research poses new challenges due to its unique problem nature. Here we elaborate on some of these challenges.

First challenge comes from the unique nature of the available gene expression data set. Though the successful application of cDNA microarrays and the high-density oligonucleotides have made fast simultaneous monitoring of thousands of gene expressions possible and inexpensive, the publicly available gene expression data set size remains small.

Second challenge involves dealing with a huge number of irrelevant attributes (genes). Though irrelevant attributes are present in almost every kind of data sets researchers have dealt with previously, but the ratio of irrelevant attributes to the relevant attributes is not as huge as that in the gene expression data. In most gene expression data set, the number of relevant genes only occupy a small portion of the total number of genes. Most genes are not cancer related. The presence of these irrelevant genes interferes with the discrimination power of those relevant attributes. This not only incurs extra computation time in both the training and testing phase of the classifier, but also increases the classification difficulty. One way to handle this is to incorporate a gene selection mechanism to select a group of relevant genes. Then cancer classifiers can be built on top of these selected genes.

Third challenge arises from the application domain of cancer classification. Accuracy is important in cancer classification, but it is not the only goal we want to achieve. Biological relevancy is another important criterion, since any biological information revealed during the process can help in further gene function discovery and other biological studies. Some useful information can be gained from the classification process is the determination of the genes that work as a group in determining the cancerous tissues or cells or the genes that are under-expressed or over-expressed in certain tissues or cells. All these would help biologists in gaining more understanding about the genes and how they work together and interact with each other. Therefore, biologists are more interested in classifiers that not only produce high classification accuracy but also reveal important biological information.

1.6 Motivation

The World Cancer Report described cancer as a global problem because it affects the whole greater population. The classification of different types of tumor is of great importance in cancer diagnosis and drug discovery. Earlier studies on cancer classification have limited diagnostic ability and cancer prediction has always been clinical based and morphological. Cancer classification problem has been extensively studied by researchers in the area of statistics, machine learning and databases. Many classification algorithms have been proposed in the past, such as the decision tree methods, the linear discrimination analysis, the Bayesian network, etc. For the last few years, researchers have started paying attention to the cancer classification using gene expression.

Systematic approaches based on global gene expression have been proposed, in order to understand the problem of cancer classification. Cancer classification using gene expression data stands out from the other previous classification data due to its unique nature and application domain. The recent development of microarray technology has motivated the simultaneous monitoring of genes and cancer classification using gene expression data. In its early stage of development, result obtained so far is promising. Through this survey, we hope to gain some insight into the problem of cancer classification in aid of further developing more effective and efficient classification algorithms.

1.7 Objectives

- Performs the pre-processing in order to get cleaned data.
- Gene selection is performed by using many feature selection techniques.
- To classify the cancer dataset by using different algorithms and make comparative study on those algorithms.
- To evaluate model by using confusion matrix, precision, recall and F1 score.

Chapter 2. LITERATURE SURVEY

In this article [1] they explain how machine learning helpful for cancer classification. They present a review of recent ML approaches employed in the modelling of cancer progression. The predictive models discussed in this are based on various supervised ML techniques as well as on different input features and data samples. Given the growing trend on the application of ML methods in cancer research, they present the most recent publications that employ many ML techniques as an aim to model cancer risk or patient outcomes. And explain survey of ML application in cancer. Which tells as how machine learning techniques would be help us in cancer classification. After knowing some knowledge on uses ML on cancer classification we focused on type of dataset.

In this paper [2] they explain how earlier studies on cancer classification have limited diagnostic ability. And the uses of recent development of DNA microarray technology. This technology has made monitoring of thousands of gene expression simultaneously and helps researchers to exploring the possibilities of cancer classification. They also present an overview of various cancer classification methods and evaluate these proposed methods based on their classification accuracy, computational time and ability to reveal gene information.

[3] In this paper, they used gene expression dataset with an effective ensemble approach. They show Ensemble classifiers increase not only the performance of the classification,

but also the confidence of the results. Decision tree, Random forest are some of ensemble classifier. These classifiers can be help us for cancer classification.

[4] They tried to evaluate effectiveness of machine learning algorithms in the task of lung cancer classification based on gene expression levels. They processed four publicly available data sets and used the k-nearest neighbor algorithm, naive Bayes classifier and C4.5 decision tree. They conclude that Machine learning algorithms can be used for lung cancer morphology classification and similar tasks based on gene expression level evaluation.

[5] In this paper they try to find the smallest subset of features that can guarantee highly accurate classification of breast cancer as either benign or malignant. Then a relative study on different cancer classification approaches. Naïve Bayes (NB), Logistic Regression (LR), and Decision Tree (DT) classifiers are conducted where the time complexity of each of the classifier is also measured. Here, Logistic Regression classifier is concluded as the best classifier with the highest accuracy as compared to the other two classifiers

[6] This paper proposes an effective feature selection method, combining double RBF-kernels with weighted analysis, to extract feature genes from gene expression data, by exploring its nonlinear mapping ability.

[7] Here 8 different machine learning algorithms are applied to the data, without applying any feature selection methods. Then two different feature selection methods are applied. The results of the classifications are compared with each other and with the results of the first case. After applying the two different feature selection methods with the best 50 features are applied, SVM gave the best results. MLP is applied using different number of layers and neurons to examine the effect of the number of layers and neurons on the classification accuracy. It is determined that the increase in the number of layers sometimes decreased, sometimes didn't change the accuracy.

[8] This paper reports the state of the art of ensemble classification methods in lung cancer detection. They observed that homogeneous ensembles and decision trees were the most frequently adopted for constructing ensembles and the majority voting rule was the predominant combination rule. Few studies considered the parameter tuning of the techniques used.

[9] This paper explores the possible diagnosis of breast cancer using Radial Basis Function (RBF) for the data set. They analyse the breast Cancer data available from the Wisconsin Breast Cancer WBC, WDBC from UCI machine learning with the aim of

developing accurate prediction models for breast cancer using RBF. Overall, the RBF neural network technique has proved better performance than that of the BPN technique.

[10] This paper aims to explore the problems associated in solving the classification of cancer in gene expression data using deep learning model. They examine the genes based on their significance related to cancer types through the heat map and associate them with biomarkers. They used CNN for the classification task, fosters the deep learning framework in the cancer genome analysis and leads to better understanding of complex features in cancer disease.

Chapter 3: SYSTEM REQUIREMENTS

Requirement specification encompasses the needs of this project. It aims at providing a full description of the requirements based on the concepts defined in the Problem Domain. The system requirement specification is produced at the culmination of the analysis task. The function and performance allocated to software as part of system engineering are refined by establishing a complete information description, a detailed functional description, representation of system behavior, an indication of performance requirements and design constraints, appropriate validation criteria and other information pertinent to requirements.

3.1 Requirement Analysis and System Specification

- Intel Processor
- RAM 2 GB or above.

- Jupiter Notebook or Python IDLE.

3.2 Feasibility Study

3.2.1 Technical Feasibility

The development of this project requires time for comparison and classification of different algorithms being tested against the gene expression dataset. With a simple python environment and operating system, this project can be implemented. The main challenge is to study gene expression dataset and select important features among the vast features present in the dataset.

3.2.2 Operational Feasibility

In the proposed system we worked on three big datasets: PANCAN dataset that consists of RNA-Seq gene expressions of patients having different types of tumor: BRCA, KIRC, COAD, LUAD and PRAD have been collected from UCI Repository. The tumors are considered as class labels and the second data set is the Colon cancer data (http://microarray.princeton.edu/oncology). This data consists of 62 samples of colon epithelial cells from colon-cancer patients. The samples consists of tumors biopsies collected from tumors, and normal biopsies collected from healthy part of the colons of the same patient. The number of genes in the data set is more than 2000.

The third data set is the Leukemia data (http://www.genome.wi.mit.edu/MPR). This data consists of 72 samples. The samples consist of two types of leukemia, 25 of AML and 47 of ALL. The samples are taken from 63 bone marrow samples and 9 peripheral blood samples. There are 7192 genes in the data set

3.3 Software Requirement Specifications

Python 3

Python is an interpreted, high-level, general-purpose programming language. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Hence, this is the language we use to code because it is one of the best for working in ML because it has various libraries to work on ML.

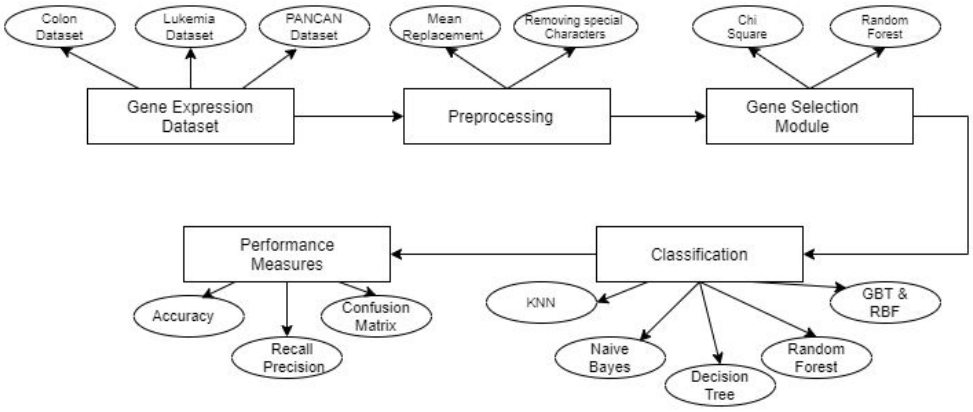
Libraries used

- Scikit learn
- Pandas
- NumPy

Chapter 4. SYSTEM DESIGN

Chapter 5. PROPOSED METHOD

A total of 6 classification algorithms have been used in this comparative study. The classifiers are both supervised and unsupervised learning algorithms. Machine learning uses two types of techniques: supervised learning, which trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data. Block diagram of the proposed system is presented below. The following points explain the function of each module.



5.1 Data Pre-Processing

Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from raw data. Raw data (real world data) is always incomplete and that data cannot be sent through a model. That would cause certain errors. That is why we need to pre-process data before sending through a model. Data Pre-processing can directly affect the ability of the model to learn; therefore, it is extremely important that pre-process the data before feeding it into a model.

Some of data pre-processing steps are importing the libraries and dataset, checking out the missing and categorical values and splitting the dataset into training and test set. We have brief introduction on dataset in future section. The concept of missing values is important to understand in order to successfully manage data. There are two basic ways to handle missing values:

1. Replacing missing values with null.
2. Replacing missing values with mean, median and mode.

The first way can be done by deleting a row if it has a null value for a feature and a particular column if it has more than 75% of missing values. This method is advised only when there are enough samples in the data set. One must make sure that after we have deleted the data, there is no addition of bias. Removing the data will lead to loss of information which will not give the expected results while predicting the output.

The second way can be applied on a feature which has numeric data. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be

negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values. This method is also called as leaking the data while training. Another way is to approximate it with the deviation of neighboring values. This works better if the data is linear.

Therefor we are using second way to replace the missing values. First, we find out the all special characters or null values in the dataset and replace it by the mean values. This works by calculating the mean of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. Replacing median can be help more if dataset has more outliers.

Uses of replacing noisy data with mean are,

- It is simple and fast.
- Works well with small numerical datasets.

5.2 Feature Selection

Feature selection methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable. It is desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model.

There are two main types of feature selection algorithms:

- Wrapper Feature Selection Methods.
- Filter Feature Selection Methods.

Wrapper feature selection methods: Create many models with different subsets of input features and select those features that result in the best performing model according to a performance metric. These methods are unconcerned with the variable types, although they can be computationally expensive.

Filter feature selection methods: Use statistical techniques to evaluate the relationship between each input variable and the target variable, and these scores are used as the basis to choose (filter) those input variables that will be used in the model. It is common to use correlation type statistical measures between input and output variables as the basis for filter feature selection.

5.2.1 Chi-Square as a Feature selection

The **Chi Square** statistic is commonly used for testing relationships between categorical variables. The null hypothesis of the Chi-Square test is that no relationship exists on the categorical variables in the population; they are independent.

Steps to perform the Chi-Square Test:

Step 1: Define Hypothesis.

Step 2: Build a Contingency table.

Step3: Find the expected values.

Step 4: Calculate the Chi-Square statistic.

Step 5: Accept or Reject the Null Hypothesis.

1. Define Hypothesis

Null Hypothesis (H0): Two variables are independent.

Alternate Hypothesis (H1): Two variables are not independent.

2. Contingency table

Contingency table can be built by considering one variable in rows and another in columns. It is used to study the relation between two variables. For example considering cancer type as rows and gene type as a columns. We can calculate Degrees of freedom for contingency table by (r-1) * (c-1) where r, c are rows and columns. From this table we have figured out all observed values and our next steps are to find expected values, get the Chi-Square value and check for relationship.

3. Find the Expected Value

Based on the null hypothesis that the two variables are independent. We can say if A, B are two independent events.

$$P(A \cap B) = P(A) * P(B)$$

Let's E1 be the expected value for the first cell in table, it can be find by using below equation.

$$E1 = n * p$$

Where n is the probability of column variable.

p is the probability of row variable.

In similar, we calculate remaining Expected values.

4. Calculate Chi-Square value

A chi-square test is used in statistics to test the independence of two events. Given the data of two variables, we can get observed count O and expected count E. Chi-Square measures how expected count E and observed count O deviates each other.

The Formula for Chi Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

c = degrees of freedom

O = observed value(s)

E = expected value(s)

In feature selection, we aim to select the features which are highly dependent on the response. When two features are independent, the observed count is close to the expected count, thus we will have smaller Chi-Square value. So high Chi-Square value indicates that the hypothesis of independence is incorrect. In simple words, higher the Chi-Square value the feature is more dependent on the response and it can be selected for model training.

5. Accept or Reject the Null Hypothesis

With 95% confidence that is alpha = 0.05, we will check the calculated Chi-Square value falls in the acceptance or rejection region.

Pros and Cons of Chi square

Pros

- The reason beyond using chi square as a feature selection is, the data type of the feature to be tested and the target variable are both categorical attributes.

Cons

- Chi-Square is sensitive to small frequencies in cells of tables. Generally, when the expected value in a cell of a table is less than 5, chi-square can lead to errors in conclusions.

5.2.2 Random forest as a feature selection

Random Forests are often used for feature selection in a data science workflow. The reason is because the tree-based strategies used by random forests naturally ranks by how well they improve the purity of the node. This mean decrease in impurity over all trees (called Gini impurity). Nodes with the greatest decrease in impurity happen at the start of the trees, while notes with the least decrease in impurity occur at the end of trees. Thus, by pruning trees below a particular node, we can create a subset of the most important features.

5.3 Algorithms Used

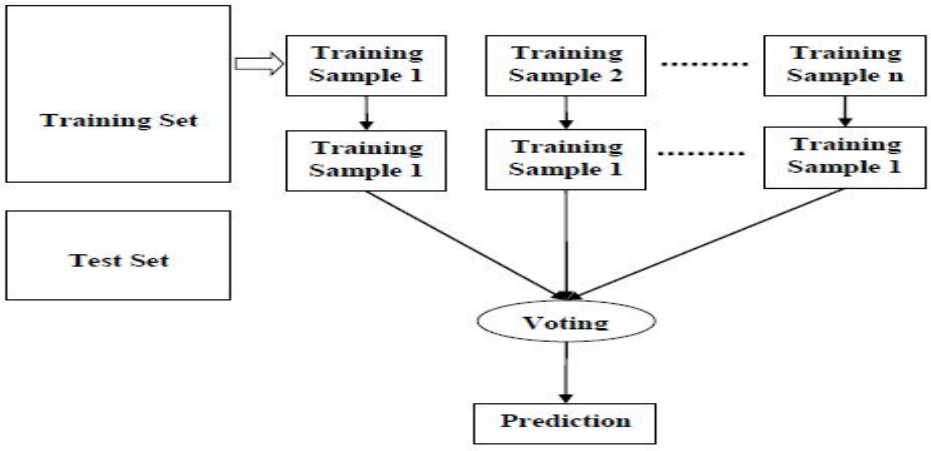
5.3.1 Random Forest Algorithm

1. First, start with the selection of random samples from a given dataset.

2**.** Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree. The decision tree is a decision support tool. Decision contains set of rules. These rules are constructed by using information gain and Gini index calculations.

3. In this step, voting will be performed for every predicted result.

4. At last, select the most voted prediction result as the final prediction result.



Pros and Cons of Random Forest

Pros

- It overcomes the problem of overfitting by averaging or combining the results of different decision trees.
- Random forests work well for a large range of data items than a single decision tree does.
- Random forests are very flexible and possess very high accuracy.
- Random Forest algorithms maintains good accuracy even a large proportion of the data is missing.

Cons

- Complexity is the main disadvantage of Random forest algorithms.
- More computational resources are required to implement Random Forest algorithm.
- It is less intuitive in case when we have a large collection of decision trees.
- The prediction process using random forests is very time-consuming in comparison with other algorithms.

5.3.2 Classification Algorithms- Naïve Bayes

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence. The below Bayes' equation helps to find conditional probability.

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Where P(y/x) is posterior probability.

P(x/y) is likelihood probability.

P(y) is class prior probability.

P(x) is predicator prior probability.

Naïve Bayes Algorithm

1. Calculate the prior probability for given class labels

2. Find Likelihood probability with each attribute for each class

3. Put these value in Bayes Formula and calculate posterior probability.

4. See which class has a higher probability, given the input belongs to the higher probability class.

Mathematical representation

The algorithm uses Bayes theorem. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred.

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Where, y is class variable and x are a dependent feature vector (of size n) where:

x = (x1, x2……. Xn)

Here x1, x2……. Xn represent the features, by substituting for X and expanding using the chain rule we get,

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y).....P(x_n|y)P(y)}{P(x_1)P(x_2)....P(x_n)}$$

Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible values of the class variable y and pick up the output with maximum probability. This can be expressed mathematically as:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Using the above function, we can obtain the class, given the predictors.

Pros and Cons of Naïve Bayes algorithm

Pros

- Naive Bayes classifiers are linear classifier hence it's simple to implement.
- Can solve problems involving both categorical and continuous valued attributes?
- For very high-dimensional data, when model complexity is less important.
- For very well-separated categories, when model complexity is less important

The last two points seem distinct, but they actually are related: as the dimension of a dataset grows, it is much less likely for any two points to be found close together (after all, they must be close in every single dimension to be close overall). This means that clusters in high dimensions tend to be more separated, on average, than clusters in low dimensions, assuming the new dimensions actually add information. For this reason, simplistic classifiers like naive Bayes tend to work as well or better than more complicated classifiers as the dimensionality grows.

Cons

- The assumption that all features are independent is not usually the case in real life so it makes Naive Bayes algorithm less accurate than complicated algorithms.

5.3.3 K-Nearest Neighbor

K-nearest neighbor (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well,

- **Lazy learning algorithm**: KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- **Non-parametric learning algorithm**: KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

KNN Algorithm

1. Initialize K to your chosen number of neighbours and N will be the training data samples.

2. For i=0 to N: Calculate the Euclidean distance between test samples to all training samples. And add the distance to an ordered collection

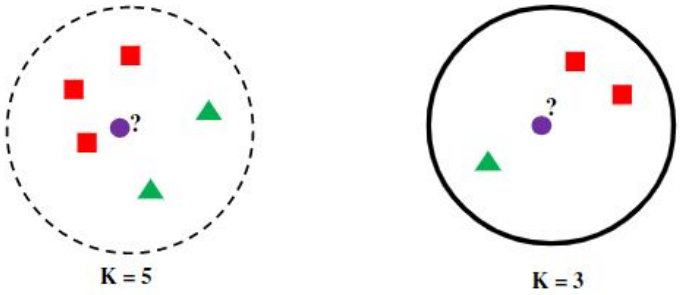
3. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances

4. Pick the first K entries from the sorted collection

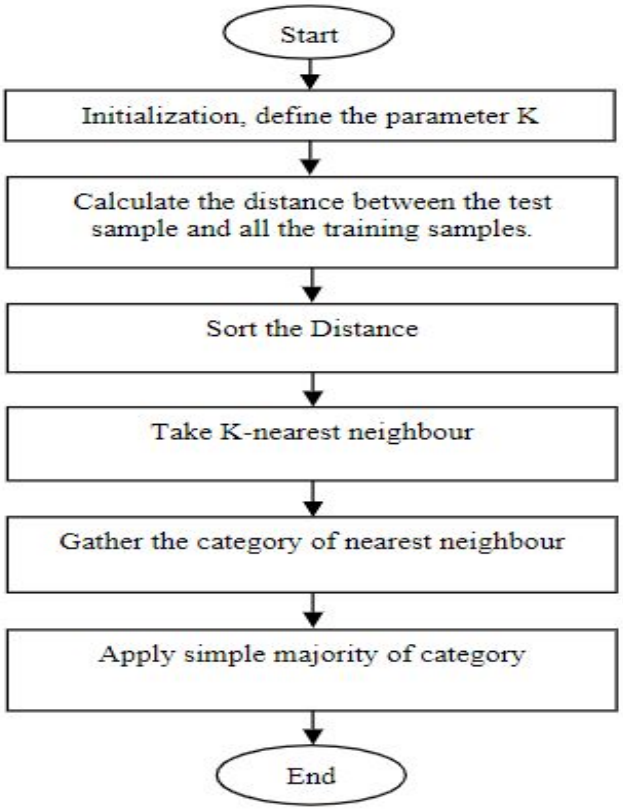
5. Get the labels of the selected K entries

Working of KNN with example

In K-NN Classification, the output is a class membership. Classification is done by a majority vote of neighbours. If K = 1, then the class is single nearest neighbour.



A test sample classification is shown in the above Fig. Consider the test sample is a big dot located inside the circles which is classified either to the first class of triangles or to the second class of squares. If K=5 (dashed line circle) it is assigned to the second class because there are 3 squares and 2 triangles inside that circle. If K=3 (solid line circle) it is assigned to the second class because here 2 squares and 1 triangle inside that circle. It can be useful if the weight contributions of the neighbours are considered because the nearer neighbours contribute more than the distant ones. For example, in a common weighting scheme, individual neighbour is assigned to a weight of 1/d if d is the distance to the neighbour. The shortest distance between any two neighbours is always a straight line and the distance is known as Euclidean distance.



Pros and Cons of KNN algorithm

Pros

- The algorithm is simple and easy to implement.
- There's no need to build a model, tune several parameters, or make additional assumptions.
- Naturally handles multi class cases.

Cons

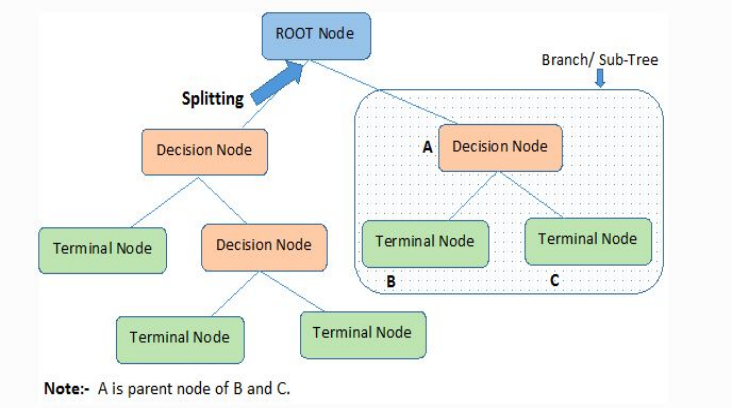
- The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.

5.3.4 Decision Tree

The decision tree classifier creates the classification model by building a decision tree. Decision Tree algorithms are referred to as CART (Classification and Regression Trees). Decision tree is constructed from given attributes on the basis of Gini index impurity. Decision tree can contract for both continuous and categorical target variables.

Decision Tree Algorithm

1. Place the best attribute of the dataset at the root of the tree.
2. Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until it finds leaf nodes in all the branches of the tree.



Assumptions while creating Decision Tree

- At the beginning, the whole training set is considered as the root.
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- Records are distributed recursively on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

Attributes Selection

If dataset consists of "n" attributes then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. If it selected by random approach, it may give low accuracy. For solving this attribute selection problem, there are some criterion like information gain, Gini index, etc. These criterions will calculate values for every attribute. The values are sorted, and attributes are placed in the tree by following the order i.e., the attribute with a high value (in case of information gain) is placed at the root. While using information Gain as a criterion, we assume attributes to be categorical, and for Gini index, attributes are assumed to be continuous.

Pros and Cons of Decision Tree

Pros

1. Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.
2. A decision tree does not require scaling of data as well.
3. Decision tree can use for both continuous and categorical target variables.

Cons

1. A small change in the data can cause a large change in the structure of the decision tree causing instability.
2. For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
3. Decision tree often involves higher time to train the model.
4. Decision tree training is relatively expensive as complexity and time taken is more.

5.3.5 Gradient Boosting classifier

Gradient boosting classifier is a machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets.

Theory behind Gradient Boost

Gradient boosting involves three elements:

1. A loss function to be optimized.
2. A weak learner to make predictions.
3. An additive model to add weak learners to minimize the loss function.

The Gradient Boosting Classifier depends on a loss function. A custom loss function can be used, and many standardized loss functions are supported by gradient boosting classifiers, but the loss function has to be differentiable. Classification algorithms frequently use logarithmic loss, while regression algorithms can use squared errors. Gradient boosting systems don't have to derive a new loss function every time the boosting algorithm is added, rather any differentiable loss function can be applied to the system.

Gradient boosting systems use decision trees as their weak learners. Regression trees are used for the weak learners, and these regression trees output real values. Because the outputs are real values, as new learners are added into the model the output of the regression trees can be added together to correct for errors in the predictions.

The additive component of a gradient boosting model comes from the fact that trees are added to the model over time, and when this occurs the existing trees aren't manipulated, their values remain fixed.

Steps to perform Gradient Boosting

In order to implement a gradient boosting classifier, we'll need to carry out a number of Different steps. We'll need to:

1. Fit the model

2. Tune the model's parameters and hyper parameters

3. Make predictions

4. Interpret the results

Fitting models by using the fit () command after setting up the model.

However, tuning the model's hyper parameters requires some active decision. There are various arguments/hyper parameters we can tune to try and get the best accuracy for the model. One of the ways we can do this is by altering the learning rate of the model. We'll want to check the performance of the model on the training set at different learning rates, and then use the best learning rate to make predictions.

Predictions can be made in by using predict () function after fitting the classifier. When we want to predict on the features of the testing dataset, and then compare the predictions to the actual labels. The process of evaluating a classifier typically involves checking the accuracy of the classifier and then tweaking the parameters/hyper parameters of the model until the classifier has an accuracy that the user is satisfied with.

Pros and Cons of Gradient Boosting

Pros

- Lots of flexibility - can optimize on different loss functions and provides several hyper parameter tuning options that make the function fit very flexible.
- No data pre-processing required, often works great with categorical and numerical values as is.

Cons

- Computationally expensive - GBMs often require many trees (>1000) which can be time and memory exhaustive.

Random Forest

Random Forest is an ensemble tree-based learning algorithm. The Random Forest Classifier is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object. The working method is same as, what we explained earlier. Random forest runs efficiently on large datasets but is comparatively slower than other algorithms. It can effectively estimate missing values and hence is suitable for handling datasets with large number of missing values.

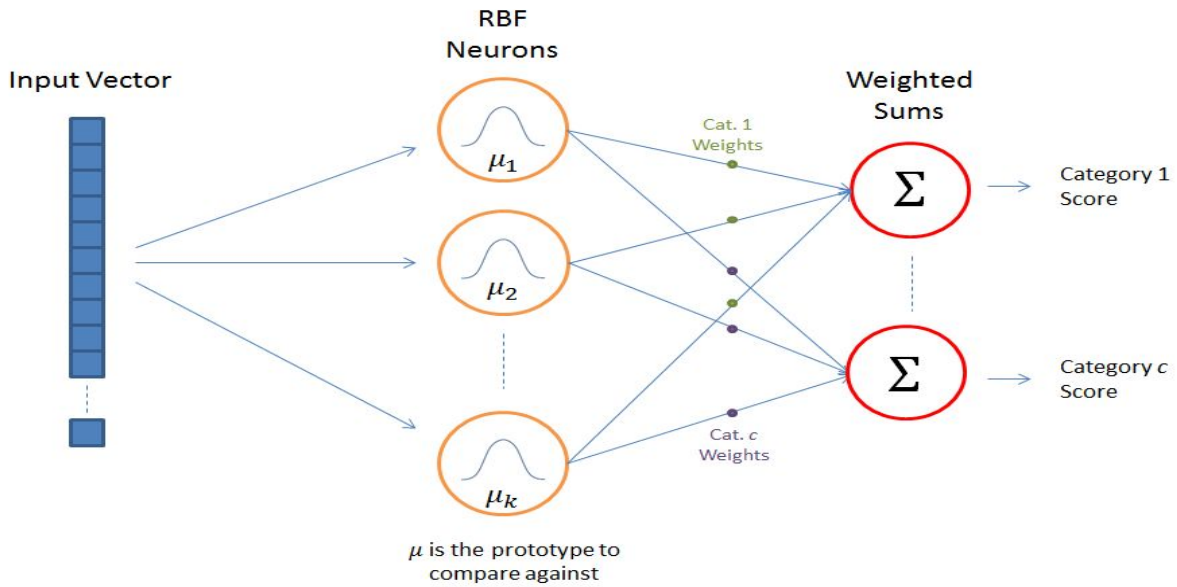
5.3.6 Neural Networks-Radial Basis Function:

RBF network is an artificial neural network with an input layer, a hidden layer, and an output layer. The Hidden layer of RBF consists of RBF neurons, and activation function of these neurons is a Gaussian function.

An RBFN performs classification by measuring the input's similarity to examples from the training set. Each RBFN neuron stores a "prototype", which is just one of the examples from the training set. When we want to classify a new input, each neuron computes the Euclidean distance between the input and its prototype. Roughly speaking,

if the input more closely resembles the class A prototypes than the class B prototypes, it is classified as class A.

RBF Network Architecture



The above illustration shows the typical architecture of an RBF Network. It consists of an input vector, a layer of RBF neurons, and an output layer.

The Input Vector

The input vector is the n-dimensional vector that what we're trying to classify. The entire input vector is shown to each of the RBF neurons.

The RBF Neurons

Each RBF neuron stores a "prototype" vector which is just one of the vectors from the training set. Each RBF neuron compares the input vector to its prototype, and outputs a value between 0 and 1 which is a measure of similarity. If the input is equal to the prototype, then the output of that RBF neuron will be 1. As the distance between the input and prototype grows, the response falls off exponentially towards 0. The shape of the RBF neuron's response is a bell curve, as illustrated in the network architecture

diagram. The neuron's response value is also called its "activation" value. The prototype vector is also often called the neuron's "centre", since it's the value at the centre of the bell curve.

The Output Nodes

The output of the network consists of a set of nodes, one per category that we are trying to classify. Each output node computes a sort of score for the associated category. Typically, a classification decision is made by assigning the input to the category with the highest score.

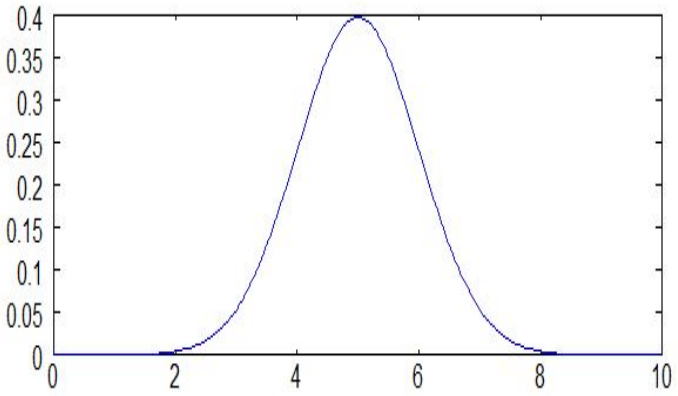
The score is computed by taking a weighted sum of the activation values from every RBF neuron. By weighted sum we mean that an output node associates a weight value with each of the RBF neurons, and multiplies the neuron's activation by this weight before adding it to the total response. Because each output node is computing the score for a different category, every output node has its own set of weights. The output node will typically give a positive weight to the RBF neurons that belong to its category, and a negative weight to the others.

RBF Neuron Activation Function

Each RBF neuron computes a measure of the similarity between the input and its prototype vector (taken from the training set). Input vectors which are more similar to the prototype return a result closer to 1. There are different possible choices of similarity functions, but the most popular is based on the Gaussian. Below is the equation for a Gaussian with a one-dimensional input.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where x is the input, μ is the mean, and σ is the standard deviation. This produces the familiar bell curve shown below, which is cantered at the mean, σ (in the below plot the mean is 5 and sigma is 1).



The RBF neuron activation function is slightly different, and is typically written as:

$$\phi(X) = e^{\frac{-||X-\mu||^2}{\sigma^2}}$$

In the Gaussian distribution, σ refers to the mean of the distribution. Here, it is the prototype vector which is at the centre of the bell curve.

RBF Algorithm

- Define the number of hidden neurons "K".
- Set the positions of RBF centres using K-means clustering algorithm.
- Calculate σ (Standard deviation) using equation.
- Calculate actions of RBF node using equation.
- Train the output using equation.

Pros and Cons of RBF

Pros

- Training in RBNN is faster than in Multi-layer Perceptron (MLP) takes many interactions in MLP.

- We can easily interpret what is the meaning of the each node in hidden layer of the RBNN. This is difficult in MLP.

Cons

- Classification will take more time in RBNN than MLP.

Model Evaluation

Validation is done by using confusion matrix, Accuracy, precision, recall and F1 score.

Confusion Matrix

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**.**

	Predicted No	Predicted Yes
Actual No	TN	FP
Actual Yes	FN	TP

- **True positives** (TP): Classifier predicted yes, and it's correct prediction.
- **True negatives** (TN): Classifier predicted no, and it's correct prediction.
- **False positives** (FP): Classifier predicted yes, but actually it's not correct prediction.
- **False negatives** (FN): Classifier predicted no, but actually it's not correct prediction.

Accuracy

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right.

Accuracy = (TP+TN) / (TN+FP+FN+TP)

Precision

Precision quantifies the number of positive class predictions that actually belong to the positive class.

Precision = TP / (TP+FP)

Recall

Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.

Recall = TP / (FP + FN)

F1-Measure

F-Measure provides a single score that balances both the concerns of precision and recall in one number. Basically, it is weighted harmonic mean of precision and recall.

F1- Measure = (2 * precision * recall) / (precision + recall)

Chapter 6 IMPLEMENTATION AND EXPERIMENTAL ANALYSIS

6.1 Information on Dataset

First dataset is The PANCAN dataset that consists of RNA-Seq gene expressions of patients having different types of tumors: BRCA, KIRC, COAD, LUAD and PRAD have been collected from UCI Repository. The tumors are considered as class labels.

Number of instances: 801

Number of attributes: 16384

The class labels BRCA, KIRC, COAD, LUAD, PRAD are 5 different cancer/tumour types.

Type	Name	Organ
BRCA	Breast Invasive Carcinoma	Breast
KIRC	Kidney renal clear cell carcinoma	Kidney

COAD	Colon adenocarcinoma	Large intestine
LUAD	Lung adenocarcinoma	Lungs
PRAD	Prostate adenocarcinoma	Prostate gland

The second data set is the Colon cancer data (http://microarray.princeton.edu/oncology). This data consists of 62 samples of colon epithelial cells from colon-cancer patients. The samples consist of tumors biopsies collected from tumors, and normal biopsies collected from healthy part of the colons of the same patient. The number of genes in the data set is more than 2000.

The third data set is the Leukemia data (http://www.genome.wi.mit.edu/MPR). This data consists of 72 samples. The samples consist of two types of leukemia, 25 of AML and 47 of ALL. The samples are taken from 63 bone marrow samples and 9 peripheral blood samples. There are 7192 genes in the data set.

6.2 Results and Discussions

RESULTS AND DISCUSSIONS

(i) PANACAN Data Set

Confusion Matrix

Algorithm	Without feature Selection	With Random Forest	With Chi-Square
Naive bayes	[[76 0 4 1 0] [1 9 0 8 0] [2 0 25 2 0] [11 0 2 27 0]	[[81 0 0 0 0] [0 18 0 0 0] [0 0 28 1 0] [0 0 0 40 0]	[[62 0 0 0 0] [0 14 0 0 0] [0 0 23 0 0] [1 0 0 31 0]

	[1 0 0 0 31]]	[0 0 0 0 32]]	[0 0 0 0 29]]
Random Forest	[[80 0 0 1 0] [0 18 0 0 0] [0 0 28 1 0] [1 0 0 39 0] [1 0 0 0 31]]	[[80 0 0 1 0] [0 18 0 0 0] [0 0 29 0 0] [0 0 0 40 0] [1 0 0 0 31]]	[[62 0 0 0 0] [0 14 0 0 0] [1 0 22 0 0] [0 0 0 32 0] [0 0 0 0 29]]
KNN	[[81 0 0 0 0] [0 18 0 0 0] [0 0 29 0 0] [0 0 0 40 0] [0 0 0 0 32]]	[[81 0 0 0 0] [0 18 0 0 0] [0 0 29 0 0] [0 0 0 40 0] [0 0 0 0 32]]	[[62 0 0 0 0] [0 14 0 0 0] [0 0 23 0 0] [0 0 0 32 0] [0 0 0 0 29]]
Decision Tree	[[80 0 0 1 0] [0 18 0 0 0] [0 0 29 0 0] [0 2 0 38 0] [0 0 0 0 32]]	[[81 0 0 0 0] [0 18 0 0 0] [0 0 29 0 0] [1 2 0 37 0] [0 0 0 0 32]]	[[59 0 0 2 1] [0 14 0 0 0] [0 0 23 0 0] [0 0 0 32 0] [1 0 0 0 28]]
GBT	[[80 0 0 1 0] [1 17 0 0 0] [1 0 27 1 0]	[[81 0 0 0 0] [0 18 0 0 0] [0 0 28 1 0]	[[62 0 0 0 0] [0 14 0 0 0] [0 0 23 0 0]

	[1 0 1 38 0] [1 0 0 0 31]]	[1 0 0 39 0] [1 0 0 0 31]]	[0 0 0 32 0] [0 0 0 0 29]]
RBF	[[81 0 0 0 0] [0 18 0 0 0] [0 0 29 0 0] [0 0 0 40 0] [0 0 0 0 32]]	[[51 30 0 0 0] [3 15 0 0 0] [3 26 0 0 0] [4 36 0 0 0] [1 31 0 0 0]]	[[49 13 0 0 0] [0 14 0 0 0] [0 23 0 0 0] [0 32 0 0 0] [0 29 0 0 0]]

Accuracy

Algorithm	Without feature Selection	With Random Forest	With Chi-Square
Naive bayes	0.84	0.995	0.99
Random Forest	0.98	0.99	0.99375
KNN	1.0	1.0	1.0
Decision Tree	0.985	0.985	0.975
GBT	0.965	0.985	1.0

RBF	0.33	0.33	0.395

(ii)Gene Expression Leukemia Data Set

Confusion Matrix

Algorithm	Without feature Selection	With Random Forest	With Chi-Square
Naive bayes	[[11 1] [1 9]]	[[8 1] [1 8]]	[[17 1] [1 3]]
Random Forest	[[11 1] [3 7]]	[[8 1] [2 7]]	[[17 1] [1 3]]
KNN	[[11 1] [6 4]]	[[8 1] [1 8]]	[[17 1] [1 3]]
Decision Tree	[[10 2] [2 8]]	[[9 0] [2 7]]	[[17 1] [1 3]]
GBT	[[10 2] [8 2]]	[[8 1] [0 9]]	[[17 1] [1 3]]

RBF	[[12 0] [8 2]]	[[9 0] [3 6]]	

Accuracy

Algorithm	Without feature Selection	With Random Forest	With Chi-Square
Naive bayes	0.90	0.88	0.90
Random Forest	0.81	0.83	0.90
KNN	0.68	0.88	0.90
Decision Tree	0.81	0.88	0.90
GBT	0.54	0.94	0.90
RBF	0.63	0.83	0.90

(iii) Colon Cancer Gene Expression dataset

Confusion Matrix

Algorithm	Without feature Selection	With Random Forest	With Chi-Square
Naive bayes	[[11 4] [1 3]]	[[12 3] [1 3]]	[[10 4] [0 2]]
Random Forest	[[13 2] [0 4]]	[[13 2] [0 4]]	[[12 2] [0 2]]
KNN	[[15 0] [3 1]]	[[14 1] [3 1]]	[[14 0] [0 2]]
Decision Tree	[[10 5] [1 3]]	[[10 5] [2 2]]	[[10 4] [1 1]]
GBT	[[14 1] [0 4]]	[[13 2] [1 3]]	[[8 6] [0 2]]
RBF	[[13 2] [3 1]]	[[15 0] [4 0]]	[[13 1] [1 1]]

Accuracy

Algorithm	Without feature Selection	With Random Forest	With Chi-Square

Naive bayes	0.736	0.789	0.75
Random Forest	0.894	0.894	0.875
KNN	0.842	0.789	1.0
Decision Tree	0.684	0.631	0.6875
GBT	0.947	0.842	0.625
RBF	0.73	0.78	0.875

Chapter 7. CONCLUSION AND FUTURE SCOPE

We observed that, algorithms perform better when the number of instances in the dataset is more. Based on the results obtained,when compared with all the dataset,we observed Random Forest has performed well. When Chi-Square feature selection is applied,we observed considerable increase in accuracy compared to random forest.
In future we would like to work upon different gene dataset and like to explore more algorithms.

REFERENCES

1. Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis , Dimitrios I. Fotiadis. “Machine learning applications in cancer prognosis and prediction”. Computational and Structural Biotechnology Journal 13:8–17. November 2014.

2. Amit Bhola1 and Arvind Kumar Tiwari. “Machine learning based approaches for cancer classification using gene expression data”. An International Journal (MLAIJ) Vol.2, No.3/4.December 2015.

3. Sara Tarek, Reda Abd Elwahab, Mahmoud Shoman. “Gene expression based cancer classification”. Egyptian Informatics Journal 18:151–159. December 2016.

4. Maxim D Podolsky, Anton A Barchuk, Vladimir I Kuznetcov, Natalia F Gusarova, Vadim S Gaidukov, Segrey A Tarakanov,”Evaluation of Machine Learning Algorithm Utilization for Lung Cancer Classification Based on Gene Expression Levels”. Asian Pacific Journal of Cancer Prevention, Vol 17, 2016.

5. Subrata Kumar Mandal.” Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree”. International Journal Of Engineering And Computer Science ISSN: 2319-7242Volume 6, Issue 2 Feb 2017.

6. Shenghui Liu, Chunrui Xu1, Yusen Zhang1, Jiaguo Liu, Bin Yu, Xiaoping Liu and Matthias Dehmer.” Feature selection of gene expression data for Cancer classification using double RBF kernels”. Liu et al. BMC Bioinformatics 19:396, 2018.

7. Siyabend Turgut ; Mustafa Dağtekin ; Tolga Ensari.” Microarray breast cancer data classification using machine learning methods”. Published in Electric Electronics, Computer Science, and Biomedical Engineering' Meeting (EBBT), IEEE: 10.1109/EBBT.2018.8391468,2018.

8. Mohamed Hosni, Ginés García-Mateos, Juan M. Carrillo-de-Gea, Ali Idri & José Luis Fernández-Alemán. “A mapping study of ensemble classification methods

in lung cancer decision support systems". Medical & Biological Engineering & Computing (2020). https://doi.org/10.1007/s11517-020-02223-8.

9. Vijayalakshmi S and Priyadarshini J. "Breast Cancer Classification using RBF and BPN Neural Networks". International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 15 (2017).

10. Joseph M. de Guia , Madhavi Devaraj , Carson K. Leung." Deep Learning Using Gene Expression for Cancer Classification". IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) , 2019.

11. K. Arutchelvanand R. Periasamy, "Analysis of Cancer Detection System Using Datamining Approach", International Journal of Innovative Research in Advanced Engineering, vol. 2, no. 11, (2015).

12. H. Li, G. Hong and Z.Guo, "Reversal DNA methylation patterns for cancer diagnosis", 2014 8th International on ference on Systems Biology (ISB), IEEE, (2014).

13. K. Balachandran and R. Anitha, "Ensemble based optimal classification model for pre-diagnosis of lung cancer",2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), IEEE,(2013).

14. Kathija, Shajun Nisha,"Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques", International Journal of Innovative Research in Computer and Communication Engineering- Vol. 4,Issue 12, December 2016.