

GENE EXPRESSION ANALYSIS ON CANCER DATASET

Submitted in partial fulfillment of the requirements for
the award of
Bachelor of Engineering degree in Computer Science and Engineering

By

M.Vinod Kumar Reddy (38110316)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600 119**

MARCH - 2020



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with “A” grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600
119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide
work of **M.Vinod Kumar Reddy (38110316)**
who carried out the project entitled “**GENE EXPRESSION ANALYSIS ON CANCER
DATASET**” under my supervision from **June 2019** to **March 2020**.

Internal Guide

Dr.T.Sasikala, M.E.,

Ph.d

Head of the Department

DR. VIGNESHWARI S, M.E., Ph. D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I, **M.Vinod Kumar Reddy (38110316)** hereby declare that the Project Report entitled "**GENE EXPRESSION ANALYSIS ON CANCER DATASET**" done by me under the guidance of Dr.T.Sasikala (Internal) is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering / Technology degree in **Computer Science and Engineering**.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr.T.Sasikala M.E., Ph. D, Dean**, School of Computing, **Dr.S.Vigneshwari M.E., Ph.D.**, and **Dr.L.Lakshmanan M.E., Ph.D.**, Heads of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.T.Sasikala, M.E, Ph.D**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

Genes are the basis of tumor formations around the body, better known as cancer. They inhibit basic processes such as cell death (apoptosis) and promote cell division to an unhealthy extent. The expression of every gene provides a baseline as to how far a cancer has progressed, the organ or tissue it originated from and its approximated course of action. The analysis of such gene expression values using traditional machine learning methods provides a higher efficiency and accuracy at finding relationships between genes and may serve as a future for diagnosis for cancer using these values. The main challenge is to use the bases created to efficiently compute the highly effective genes for specific types of cancer using their expression values and thus, raise the question of a potential relationship between them for each type. A Random Forest Model has been used to perform Feature Selection over the dataset to extract the important features (i.e.) the most influential genes. They are then visualized using traditional packages in Python (i.e. Scikit-plot, Matplotlib, Seaborn) and using a data visualization tool called Tableau to project the result of the analysis.

TABLE OF CONTENTS

Chapter No.	Title	Page No.
	Declaration	iii
	Acknowledgement	iv
	Abstract	v
	Table of Contents	vi
	List of Figures	vii
1	INTRODUCTION	01
	1.1 Outline	01
	1.2 Model IDE	01
	1.3 Problem Statement	01
	1.4 Objective	03
2	LITERATURE SURVEY	04
3	RANDOM FOREST ALGORITHM	06
	3.1 Introduction to Machine Learning	06
	3.2 Training the data	07
	3.3 Methods in Supervised Learning	08
	3.4 Approaches in Classification	09
	3.5 Packages	13
	3.6 Background Study	16
	3.7 The Involvement of Genes	17
	3.8 Types of Cancer and their probable genes	20
4	RESULTS AND DISCUSSION	26
	4.1 Results	26
	4.2 Analysis	37
	4.3 System Requirements	38
5	SUMMARY AND FUTURE SCOPE	40
	5.1 Summary	40
	5.2 Future Scope	40
	References	41
	APPENDIX - Source Code	45
	Screenshots	50

LIST OF FIGURES

S.No	Figures	Page No.
1	Classification vs Regression	8
2	k-Nearest Neighbour	10
3	Random Forest Simplified	12
4	The Epigenetics of Cancer	19
5	Distribution of dataset	27
6	Density plot of gene_1 vs gene_2	28
7	Cumulative density plot of 4 features	29
8	Confusion Matrix	30
9	Classification Report	31
10	Precision formula	31
11	Recall formula	32
12	F1-Score formula	32
13	Box Graph of Important Features	34
14	10 th Decision Tree	35
15	20 th Decision Tree	36
16	Line graph – Averaged important features vs Classes	37

CHAPTER – 1

INTRODUCTION

1.1. OUTLINE:

Cancers being the second leading cause of death have taken a massive toll on the population of Earth. With many cases being reported even with no prior history of risks or unhealthy habits, the race to find a cure or a preventive measure is increasing. Due to the unreliable causes present and the limited biological study, there exists the need for the intervention of other fields to help fasten the study and discovery of precision medicine and quicker diagnosing tools. This project acts as a base which may be further developed to act as a pre-diagnostic tool for the early detection of cancers.

This project uses Python 3.6 for the programming in a scientific development environment called the Jupyter Notebook. Various data manipulation, machine learning and visualization packages are used to create and analyse the dataset using a traditional machine learning model. A Data Visualization tool called Tableau is used to interpret the results provided by the model after analysis to represent and act as a proof for the intended result.

1.2. MODEL IDE:

Jupyter Notebook (formerly IPython Notebooks) is a web-based interactive computational environment for creating Jupyter notebook documents. The IDE was installed in Anaconda, an open-source distribution for the languages Python and R used to perform Data Science and Machine Learning. The IDE's UI in which the model was developed is given in Fig 1.1.

1.3. PROBLEM STATEMENT:

Tumors are groups of abnormal cells that form lumps or growths. They can start in any one of the trillions of cells in our bodies. Tumors grow and behave differently, depending on whether they are cancerous (malignant), non-cancerous (benign) or precancerous.

1.3.1 Cancerous tumors (Malignant)

Cancer can start in any part of the body. When cancer cells form a lump or growth, it is called a cancerous tumor. A tumor is cancerous when it:

- grows into nearby tissues
- has cells that can break away and travel through the blood or lymphatic system and spread to lymph nodes and distant parts of the body.

Cancer that spreads from the first place it started (called the primary tumor) to a new part of the body is called Metastatic cancer. When cancer cells spread and develop into new tumors, the new tumors are called Metastases.

1.3.2 Types of malignant cancer:

There are several types of cancer with their rates of occurrence differing based on different criteria such as gender, age, lifestyle, habits etc. In the Machine Learning model created, we discuss the occurrence of specifically 5 different types of cancer based on the gene expression values of 16382 genes. Namely,

❖ **BRCA – Breast Cancer**

❖ **LUAD – Lung Adenocarcinoma (Lung Cancer)**

❖ **PRAD – Prostate Adenocarcinoma (Prostate Cancer)**

❖ **KIRC – Kidney Renal Clear Cell Carcinoma (Kidney Cancer)**

❖ **COAD – Colon Adenocarcinoma (Colon Cancer)**

Cancerous tumors or Malignant tumors are present in 4 stages, where the 4th stage is the final or the most critical stage. During this stage, the cancer has spread throughout the affected person's body and so, the time present for treatment is decreased. Hence, a biopsy may waste time, posing a threat to the patient's life. The presence of detection of reception by gene expression aids in an earlier and faster detection and treatment towards cancer

1.4. OBJECTIVE:

The main objective of this model is to ensure earlier, faster and a more reliable detection and treatment for cancer. This model concentrates on 5 sub types of malignant cancer and not all the types of cancers and the different types of cells they may occur on. But if needed, this model can be extended to support other types of cancer or tissue types under the circumstance that the required data is prevalent, proper and reliable for further processing and application of the model to provide accurate results as any un-reliability may pose a risk if applied for practical usage.

CHAPTER – 2

LITERATURE SURVEY

In the years gone by, research on the topic of exploiting machine learning and deep learning algorithms to classify cancer-related data into the specific types of cancer and their analysis have taken place widely. This is due to the demand in understanding the deeper relationship between cancer and human genes, and also the relationship between the genes involved themselves.

Joseph M. De Guia et al proposed using microarray gene expression related to different types of cancer genes to classify them as accurately as possible. (2018)

TaeJin Ahn et al proposed a Deep Neural Network (DNN), a part of Deep Learning, to classify or identify normal cells from cancerous cells using gene expression data integrated from various datasets. (2018)

Qun-Xiong Zhu et al proposed a system using MMI for feature selection and ELM as the classifier, to perform cancer classification computationally using gene expression data. (2018)

Comparison between two machine learning models (Logistic Regression and Artificial Neural Network) and a statistical algorithm (ANOVA) was performed by **Behrouz Shamsei and Cuilan Gao** to determine their costs, advantages and drawbacks. The dataset used was animal gene expression values for Medulloblastoma cancer type. (2016)

Gene expression data for Hepatocellular carcinoma was analysed using clustering and classification techniques (Network Topology and SVM) by **Chen Shen and Zhi-Ping Liu** to identify possible module biomarkers. (2017)

Feature Extraction methods such as Chi-Square, F-Score, PCA and MRMR and an ensemble-SVM classifier were used in collaboration to classify Colon cancer by **Saima Rathore et al.** (2014)

Stacked Denoising Autoencoder, a type of Neural Network, was used as a Feature Extraction method to learn the highly influential genes from a generated gene expression dataset by **Vitor Teixeira et al.** (2017)

Muxuan Liang et al proposed a new machine learning model, called multimodal Deep Belief Network. It was trained using Contrastive Divergence in an unsupervised manner to find correlations and key genes that play a role in the pathogenesis of cancer. (2015)

Relief-F is a filtering algorithm used for Feature Selection. Linear-SVM and k-NN classifiers were used to classify using the filtering algorithm by **Yuhang Wang and F. Makedon** to compare its output efficiency when compared to other feature filtering methods. This proved that the use of the Relief filtering algorithm provided better performance in a gene expression (microarray) dataset. (2004)

Ujjwal Maulik et al proposed a method where forward greedy search algorithm was used together with Transductive SVMs to selectively procure genes that could be used to predict cancer subtypes. (2013)

Ambrosio Hernandez et al published an article to assess the possibilities of gene expression patterns playing a vital role in cancers (specifically Colon Cancer). Genomic approach was used to determine the effects of such patterns and their reliability during different metastatic stages in Colon Cancer in humans. (2000)

Sara Alghunaim and Heyam H. Al-Baity performed a comparative study of the efficiency of 3 machine learning models (SVM, Decision Tree and Random Forest), with gene expression and DNA Methylation datasets to analyse the performance and drawbacks of traditional machine learning methods on such datasets. (2019)

A Deep Neural Forest Model was used by **Jayadeep Pati** on a gene expression dataset for Lung Cancer to analyse the genes and provide targeted set of genes that may cause Lung Cancer. (2019)

Jig Xu et al proposed a novel deep flexible Neural Network Model to classify the type of cancer present using gene expression data.

CHAPTER – 3

RANDOM FOREST ALGORITHM

3.1 INTRODUCTION TO MACHINE LEARNING:

Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) with data, without being explicitly programmed.

The name *machine learning* was coined in 1959 by Arthur Samuel. Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data – such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or unfeasible; example applications include email filtering, detection of network intruders, and computer vision.

Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter sub-field focuses more on exploratory data analysis and is known as unsupervised learning.

Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

3.2 TRAINING THE DATA:

There are basically two widely-used types of training that can be done to create a model:

- i. Supervised Learning
- ii. Un-supervised Learning

3.2.1 Supervised Learning:

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from *labeled training data* consisting of a set of *training examples*. In supervised learning, each example is a *pair* consisting of an input object (typically a vector) and a desired output value (also called the *supervisory signal*). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

3.2.2 Unsupervised Learning:

Unsupervised machine learning is the machine learning task of inferring a function that describes the structure of "unlabeled" data (i.e. data that has not been classified or categorized). Since the examples given to the learning algorithm are unlabeled, there is no straightforward way to evaluate the accuracy of the structure that is produced by the algorithm—one feature that distinguishes unsupervised learning from supervised learning and reinforcement learning.

The type of training used in this model is **SUPERVISED LEARNING**.

3.3. METHODS IN SUPERVISED LEARNING:

Supervised Learning mainly consists of two methods,

- Classification
- Regression

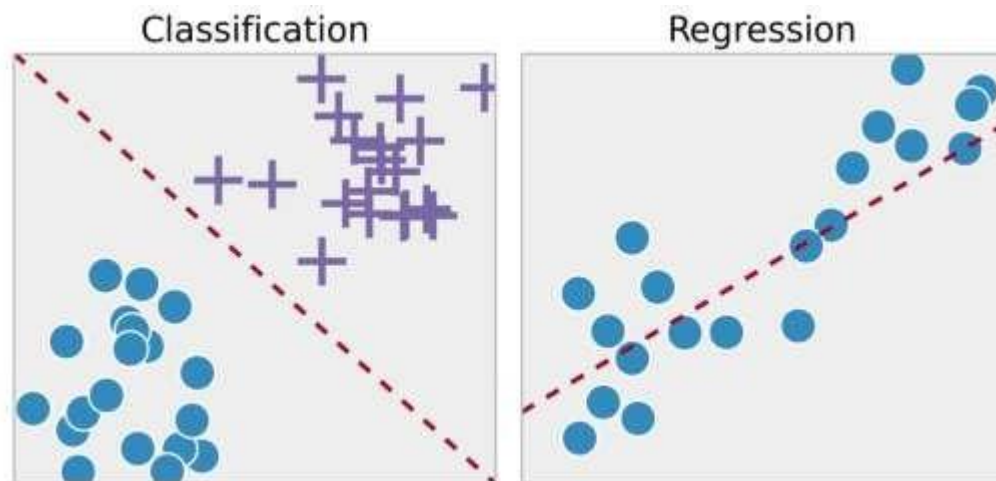


Fig 3.1 Classification vs Regression

3.3.1 Classification:

In machine learning, **classification** is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Examples are assigning a given email to the "spam" or "non-spam" class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition. An algorithm that implements classification, especially in a concrete implementation, is known as a **classifier**. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance.

3.3.2 Regression:

Regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, a function of the independent variables called the **regression function** is to be estimated.

Regression analysis is widely used for prediction and forecasting. It is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However, this can lead to illusions or false relationships, so caution is advisable; for example, correlation does not prove causation.

The type used in this model is **CLASSIFICATION** and so, more focus will be given on it.

3.4 APPROACHES IN CLASSIFICATION:

There are different approaches when it comes to the Classification method, they are:

1. Linear classifiers
 - 1.1. Fisher's linear discriminant
 - 1.2. Logistic regression
 - 1.3. Naive Bayes classifier
 - 1.4. Perceptron
2. Support vector machines
 - 2.1. Least squares support vector machines
3. Quadratic classifiers
4. Kernel estimation
 - 4.1. k-nearest neighbor

5. Boosting (meta-algorithm)
6. Decision trees
 - 6.1. Random forests
7. Neural networks
8. Learning vector quantization

The most common approaches used are **k-nearest neighbor** and **decision trees**.

3.4.1 k-Nearest Neighbour (k-NN):

In pattern recognition, the **k-nearest neighbours algorithm (k-NN)** is a nonparametric method used for classification. The output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small).

If $k = 1$, then the object is simply assigned to the class of that single nearest neighbour.

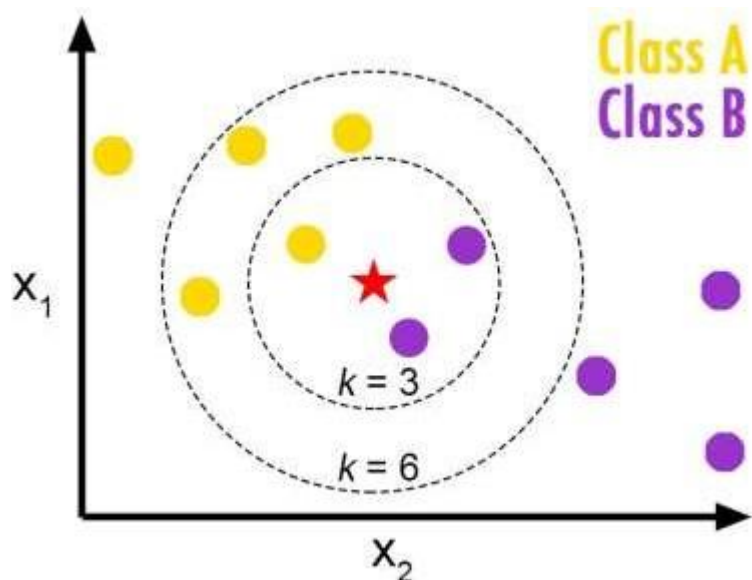


Fig 3.2 k-Nearest Neighbour

k -NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k -NN algorithm is among the simplest of all machine learning algorithms. A useful technique is the assigning of weight to the contributions of the neighbours, so that the nearer neighbours contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbour a weight of $1/d$, where d is the distance to the neighbour. The neighbours are taken from a set of objects for which the class is known.

This can be thought of as the training set for the algorithm, though no explicit training step is required. A peculiarity of the k -NN algorithm is that it is sensitive to the local structure of the data.

3.4.2 Decision Trees

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called **classification trees**; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data (but the resulting classification tree can be an input for decision making).

This page deals with decision trees in data mining.

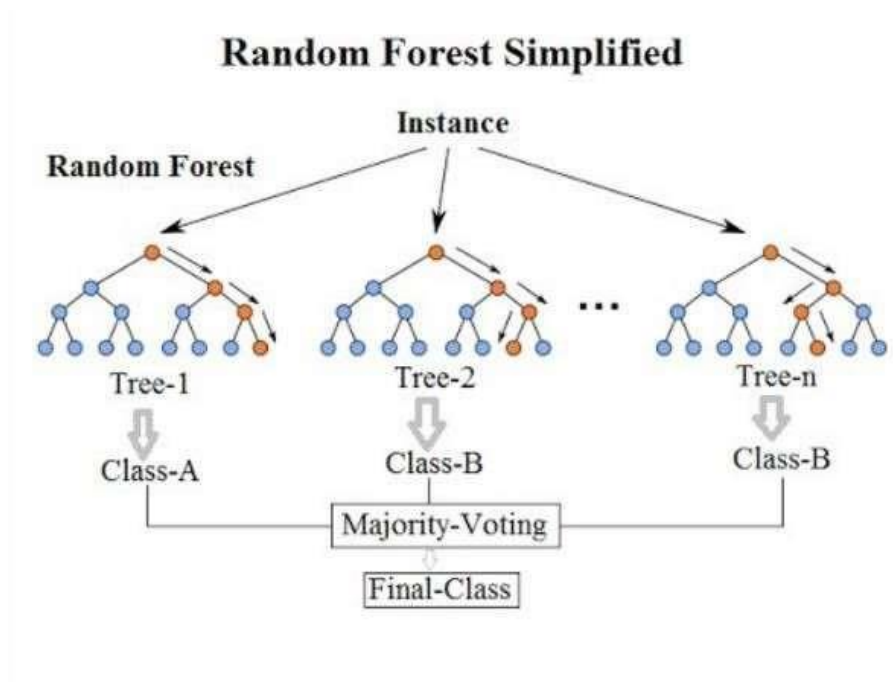


Fig 3.3 Random Forest Simplified

The most common and versatile application of decision trees come in the Random Forests approach/ algorithm. **Random forests** or **random decision forests** are an ensemble learning method for classification, that operates by constructing a multitude of decision trees at training time and outputting the class that is the most commonly appearing of the classes of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

The approach used in this model is the **RANDOM FOREST** approach which has already been discussed.

3.5. PACKAGES:

The packages used in this model include:

- Pandas
- NumPy
- Scikit-learn
- Scikit-plot
- Matplotlib's pyplot
- Seaborn
- Pydot

3.5.1 DATA MANIPULATION PACKAGES:

➤ **Pandas**

- **pandas** is a Python package providing fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, **real world** data analysis in Python. Additionally, it has the broader goal of becoming **the most powerful and flexible open source data analysis / manipulation tool available in any language**. It is already well on its way toward this goal.
- The two primary data structures of pandas, Series (1-dimensional) and DataFrame (2-dimensional), handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering. For R users, DataFrame provides everything that R's data.frame provides and much more. pandas is built on top of NumPy and is intended to integrate well within a scientific computing environment with many other 3rd party libraries.
- Written in: Python, Cython and C

➤ **NumPy**

- **NumPy** is a library for Python, adding support for large, multidimensional arrays and matrices, along with a large collection of highlevel mathematical functions to operate on these arrays. NumPy is an open-source software and has many contributors.
- NumPy targets the CPython reference implementation of Python, which is a non-optimizing bytecode interpreter. Mathematical algorithms written for this version of Python often run much slower than compiled equivalents. NumPy addresses the slowness problem partly by providing multidimensional arrays and functions and operators that operate efficiently on arrays, requiring rewriting some code, mostly inner loops using NumPy.

- Using NumPy in Python gives functionality comparable to MATLAB since they are both interpreted, and they both allow the user to write fast programs as long as most operations work on arrays or matrices instead of scalars.
- In comparison, MATLAB boasts a large number of additional toolboxes, notably Simulink, whereas NumPy is intrinsically integrated with Python, a more modern and complete programming language. Moreover, complementary Python packages are available; SciPy is a library that adds more MATLAB-like functionality and Matplotlib is a plotting package that provides MATLAB-like plotting functionality.
- Written in: Python and C

3.5.2 MODEL BUILDING PACKAGE:

➤ **Scikit-learn**

- **Scikit-learn** (formerly **scikits.learn**) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines(svm), random forest, gradient boosting, kmeans and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
- It has been built on top of NumPy, SciPy and matplotlib.
- Written in: Python, Cython, C and C++

3.5.3 DATA VISUALIZATION PACKAGES:

➤ **Scikit-plot**

- Scikit-plot is the result of an unartistic data scientist's dreadful realization that visualization is one of the most crucial components in the data science process, not just a mere afterthought.
- 38Gaining insights is simply a lot easier when you're looking at a colored heatmap of a confusion matrix complete with class labels rather than a single-line dump of numbers enclosed in brackets. Besides, if you ever need to present your results to someone

(virtually any time anybody hires you to do data science), you show them visualizations, not a bunch of numbers in Excel.

- All in all, it is an intuitive library used to add plotting functionality to a scikit-learn object.
- Written in: Python, Cython, C and C++.

➤ **Matplotlib's pyplot**

- Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.
- It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt or GTK+.
- `matplotlib.pyplot` provides a MATLAB-like plotting framework.
- `pylab` combines `pyplot` with `numpy` into a single namespace. This is convenient for interactive work, but for programming it is recommended that the namespaces be kept separate.
- Written in: Python

➤ **Pydot**

- is an interface to GraphViz
- can parse and dump into the DOT language used by GraphViz,
- is written in pure Python,
- and `networkx` can convert its graphs to `pydot`.
- GraphViz package facilitates the creation and rendering of graph descriptions in the DOT language of the GraphViz graph drawing software from Python.

3.6 BACKGROUND STUDY:

3.6.1 Overview:

- Cancer is a broad term. It describes the disease that results when cellular changes cause the uncontrolled growth and division of cells.
- Some types of cancer cause rapid cell growth, while others cause cells to grow and divide at a slower rate.
- Certain forms of cancer result in visible growths called tumors, while others, such as leukemia, do not.
- Cancerous tumors are malignant, which means they can spread into, or invade, nearby tissues. In addition, as these tumors grow, some cancer cells can break off and travel to distant places in the body through the blood or the lymph system and form new tumors far from the original tumor.
- Unlike malignant tumors, benign tumors do not spread into, or invade, nearby tissues. Benign tumors can sometimes be quite large, however. When removed, they usually don't grow back, whereas malignant tumors sometimes do. Unlike most benign tumors elsewhere in the body, benign brain tumors can be life threatening.

3.6.2 Cancer

3.6.3.1 *The How:*

- Cancer is a genetic disease—that is, it is caused by changes to genes that control the way our cells function, especially how they grow and divide.
- Genetic changes that cause cancer can be inherited from our parents. They can also arise during a person's lifetime as a result of errors that occur as cells divide or because of damage to DNA caused by certain environmental exposures. Cancer-causing environmental exposures include substances, such as the chemicals in tobacco smoke, and radiation, such as ultraviolet rays from the sun.

- In general, cancer cells have more genetic changes, such as mutations in DNA, than normal cells. Some of these changes may have nothing to do with the cancer; they may be the result of the cancer, rather than its cause.

3.6.3.2 The When:

- As scientists have learned more about the molecular changes that lead to cancer, they have found that certain mutations commonly occur in many types of cancer. Because of this, cancers are sometimes characterized by the types of genetic alterations that are believed to be driving them, not just by where they develop in the body and how the cancer cells look under the microscope.
- Metastatic cancer has the same name and the same type of cancer cells as the original, or primary, cancer. For example, breast cancer that spreads to and forms a metastatic tumor in the lung is metastatic breast cancer, not lung cancer.
- Under a microscope, metastatic cancer cells generally look the same as cells of the original cancer. Moreover, metastatic cancer cells and cells of the original cancer usually have some molecular features in common, such as the presence of specific chromosome changes.
- Treatment may help prolong the lives of some people with metastatic cancer. In general, though, the primary goal of treatments for metastatic cancer is to control the growth of the cancer or to relieve symptoms caused by it. Metastatic tumors can cause severe damage to how the body functions, and most people who die of cancer die of metastatic disease.

3.7 THE INVOLVEMENT OF GENES:

- The genetic changes that contribute to cancer tend to affect three main types of genes—proto-oncogenes, tumor suppressor genes, and DNA repair genes. These changes are sometimes called “drivers” of cancer.
- Genes carry the instructions to make proteins, which do much of the work in our cells. Certain gene changes can cause cells to evade normal growth

controls and become cancer. For example, some cancer-causing gene changes increase production of a protein that makes cells grow. Others result in the production of a misshapen, and therefore nonfunctional, form of a protein that normally repairs cellular damage.

- Genetic changes that promote cancer can be inherited from our parents if the changes are present in **germ cells**, which are the reproductive cells of the body (eggs and sperm). Such changes, called **germline changes**, are found in every cell of the offspring.
- Cancer-causing genetic changes can also be acquired during one's lifetime, as the result of errors that occur as cells divide or from exposure to **carcinogenic** substances that damage DNA, such as certain chemicals in tobacco smoke, and radiation, such as ultraviolet rays from the sun. Genetic changes that occur after conception are called **somatic (or acquired) changes**.
- There are many different kinds of DNA changes. Some changes affect just one unit of DNA, called a **nucleotide**. One nucleotide may be replaced by another, or it may be missing entirely. Other changes involve larger stretches of DNA and may include rearrangements, deletions, or duplications of long stretches of DNA.
- Sometimes the changes are not in the actual sequence of DNA. For example, the addition or removal of chemical marks, called **epigenetic modifications**, on DNA can influence whether the **gene is “expressed”**—that is, whether and how much messenger RNA is produced. (Messenger RNA in turn is translated to produce the proteins encoded by the DNA.)
- In general, cancer cells have more genetic changes than normal cells. But each person's cancer has a unique combination of genetic alterations. Some of these changes may be the result of cancer, rather than the cause. As the cancer continues to grow, additional changes will occur. Even within the same tumor, cancer cells may have different genetic changes.

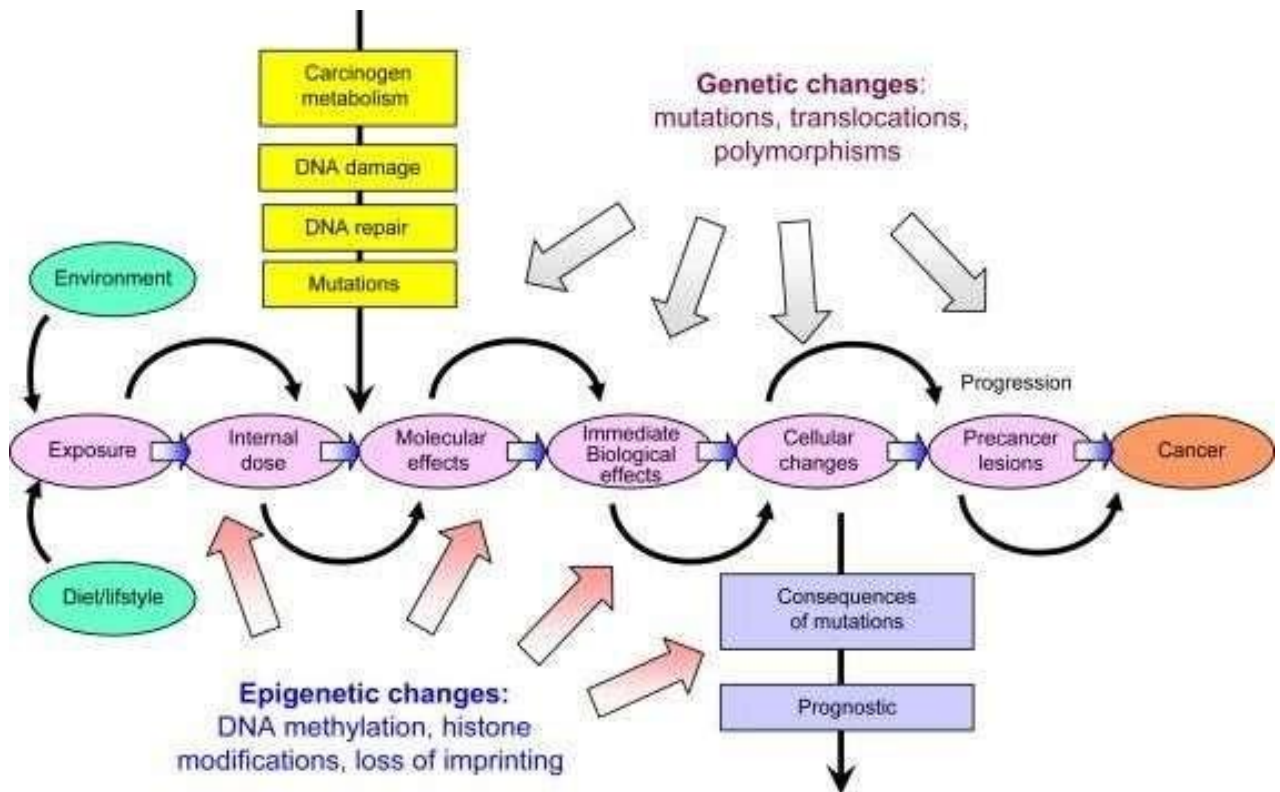


Fig 3.4 The Epigenetics of Cancer

3.7.1 Gene Expressions:

□ A **gene** is a small piece of genetic material written in a code and called DNA. Each gene has within it a set of instructions for making molecules that organisms need to survive. Genes themselves cannot be used by an organism. Instead, they must be turned into a gene product. **Gene expression** is the process by which the information contained within a gene becomes a useful product.

□ Genes can be expressed as either RNA or protein. However, not every gene product is needed all the time, nor are they needed in the same amounts. In fact, it would be very energetically demanding for a cell to express every gene all the time. Further, some gene products are harmful to the cell and thus only expressed at the end of a cell's life. Because of these reasons, we regulate the expression of our genes.

□ Human cells have a very complicated system of mechanisms to monitor themselves and their environment. They take internal and external signals,

analyze them, and then decide if a gene product is needed. Once it is decided that a gene product is needed, cells have a process to create the gene product.

- When there is damage to the processing element of the cell (i.e.) the instructor, DNA, the gene product production is misled.
- This is when the situation of higher, lower or unnecessary production of a gene product occurs, affecting all-natural processes of a cell like Cell Regeneration, general Cell Functions and timely Cell Death (Apoptosis).
- This in-turn disables the necessary actions taken to avoid cancer (i.e.) the effects of the presence of essential enemies of cancer like Tumor-Suppressor genes are disabled, giving the cell now full control to turn into a cancerous cell.
- The turn of one normal cell then multiplies into several cells with the help of cell multiplication. But unlike normal cells where the cell duplication or multiplication occurs in 2-fold, the cancer cells multiplicative process is n-fold or at least twice the normal fold, making cancer spread faster than a normal cell.

3.8 TYPES OF CANCER AND THEIR PROBABLE GENES:

3.8.1 Types of Cancer to be seen:

○ **BRCA:**

- **BRCA**, named so by The Cancer Genome Atlas Program made by the National Cancer Institute.
- It consists of two human genes, **BRCA1** and **BRCA2**, that are responsible for producing the tumor proteins.
- These proteins help repair damaged DNA and, therefore, play a major role in the stability of each cell's genetic material (i.e.) avoiding mutation.
- When either of these genes is mutated or altered, such that its protein is not made or does not function correctly, DNA damage may

not be repaired properly. As a result, cells are more likely to develop additional genetic alterations that can lead to cancer.

- Specific inherited mutations in *BRCA1* and *BRCA2* are most notably the increase in the risk of female breast and ovarian cancers, but they have also been associated with increased risks of several additional types of cancer.
- People who have inherited mutations in *BRCA1* and *BRCA2* tend to develop breast and ovarian cancers at younger ages than people who do not have these mutations.
- About 12% of women in the general population will develop breast cancer sometime during their lives. By contrast, a recent large study estimated that about 72% of women who inherit a harmful *BRCA1* mutation and about 69% of women who inherit a harmful *BRCA2* mutation will develop breast cancer by the age of 80.
- Like women from the general population, those with harmful *BRCA1* or *BRCA2* mutations also have a high risk of developing a new primary cancer in the opposite (contralateral) breast in the years following a breast cancer diagnosis. It has been estimated that, by 20 years after a first breast cancer diagnosis, about 40% of women who inherit a harmful *BRCA1* mutation and about 26% of women who inherit a harmful *BRCA2* mutation will develop cancer in their other breast.

○ **LUAD:**

- **LUAD** depicts Lung Adenocarcinoma which is a form of common lung cancer.
- It is the most common form of Lung Cancer in young women and Asians.
- The pathophysiology of adenocarcinoma is complicated, but generally follows a histologic progression from cells found in healthy lungs to distinctly dysmorphic, or irregular, cells.

- As like Breast Cancer, Lung Cancer or specifically Lung Adenocarcinoma also has its own proto-oncogenes or genes which are more probable to cause this particular type of cancer in a human.
- But unlike Breast Cancer, there is no specific gene targeted towards the occurrence of this cancer but various genes which may occur together, in a combination, paired or individually based on every human being's individual genetic structure. The high number of non-specific genes is due to the fact that Lung Cancer is more viable through smoking, which is very common. This in turn creates a strong mutagenic factor, meaning multiple genes may be affected at a given time.
- Some commonly mutated proteins producing genes are TP53 – Tumor Protein 53 (found to be mutated in 46% of all cases), EGFR – Epidermal Growth Factor Receptor (found to be mutated in 27% of all cases), KRAS (found to be mutated in 32% of all cases) and minor few such as KEAP1 – Kelch-like ECH-Associated Protein 1, STK11 - Serine/Threonine Kinase 11, NF1 – Neurofibromin 1 and CDKN2A – Cycline-Dependent Kinase Inhibitor 2A.
- In rarer cases where there are no genetic mutations but only chromosomal rearrangements (i.e.) presence of a set of genes(chromosomes) in an improper place, the genes ALK – Anaplastic Lymphoma Kinase and ROS1 are found to be most active. In particular, ALK gene is said to bind or partner with EML4 where EGFR or KRAS gene mutations seem to be inactive.

○ **PRAD:**

- **PRAD** or Prostate Adenocarcinoma is the development of cancer in the prostate, a gland in the male reproductive system. Most prostate cancers are slow growing; however, some grow relatively quickly.
- The cancer cells may spread from the prostate to other areas of the body, particularly the bones and the lymph nodes. It may initially cause no symptoms but in later stages can lead to difficulty in

urinating, blood in the urine or pain in the pelvis, back or when urinating.

- Factors that increase the risk of prostate cancer include older age, a family history of the disease and race. About 99% of cases occur in males over the age of 50. Having a first-degree relative with the disease increases the risk over two to three-fold.
- The biggest problem that arises with Prostate Cancer is the absence of any single gene that would be responsible for the occurrence of this cancer. Many different genes have been implicated with respect to this cancer.
- The genes BRCA1 and BRCA2 which are majorly related to breast and ovarian cancers in women have been found to play a role in the occurrence of prostate cancer in men.
- Other linked genes include the Hereditary Prostate Cancer gene 1 (HPC1), the androgen receptor, the vitamin D receptor and the TMPRSS2-ETS gene family fusion (TransMembrane Protease, Serine 2-Erythroblast Transformation Specific).
- A wide variety of genomes which were capable of linking Single Nucleotide Polymorphisms (SNPs) were found to play a major role in Prostate Cancer in 2008.

○ **KIRC:**

- **KIRC** or also known as Kidney or Renal Clear Cell Carcinoma (RCC) is a common type of Kidney Cancer occurring in humans, occurring in approximately 90-95% of all Kidney Cancer cases.
- The initial symptoms of Renal Cell Carcinoma often include blood in the urine (occurring in 40% of cases), flank pain (40%), a mass in the abdomen or flank (25%), weight loss (33%), fever (20%), high blood pressure (20%), night sweats and generally feeling unwell.

- When RCC metastasizes (i.e.) spreads, the most common areas it can spread to are the lymph nodes, lungs, liver, adrenal glands, the brain and bones.
- The general five-year-survival rate is 65-90%, but this is lowered considerably when the cancer has spread or has metastasized.
- The factor heredity is minor in this type of cancer or have major impact on the ability to procure this cancer with people having immediate relatives diagnose with RCC having a two to four-fold risk of developing the condition.
- The genes that were found to cause an increased risk of kidney cancer when mutated were the FH (Fumarate Hydratase) gene, the FLCN (Folliculin) gene, the SDHB (Succinate DeHydrogenase complex subunit B) gene and the SDHD (Succinate DeHydrogenase complex subunit D) gene.

○ **COAD:**

- **COAD** or Colon Adenocarcinoma is a common form of Colon Cancer (also known as Colorectal Cancer or Bowel Cancer) and develops in the colon or rectum (parts of the large intestine).
- General signs and symptoms include blood in the stool, a change in bowel movements, weight loss and feeling tired all the time.
- Most of the time, this particular type of cancer occurs due to old age and lifestyle factors, with only a small number of cases due to underlying genetic disorders. Other risk factors include diet, obesity, smoking and lack of physical activity.
- Those with a family history in two or more first-degree relatives have a two to three-fold greater risk of obtaining this type of cancer and this type accounts for about 20% of all cases.

- Mutations in genes such as POLE (DNA Polymerase Epsilon) and POLD1 (Gene Polymerase Delta 1) have been associated with familial colon cancer.
- Other genes that are targeted towards colon cancer include primarily MACC1 (Metastasis Associated in Colon Cancer-1) and secondarily include MSH2 (MutS Homolog 2), MSH6 (MutS Homolog 6) and MLH1 (MutL Homolog 1).

CHAPTER – 4

RESULTS AND DISCUSSION

4.1 Results:

The result of the classification model prepared using Python may be divided into three parts in itself:

- I. Pre-model Data Visualization
- II. Model Information Visualization
- III. Post-model Data Visualization

4.1.1 Pre-Model Data Visualization

The pre-model data visualization section consists of all the data visualizations or example visualizations made with the pure data set to understand the features and the spread of data in the data set.

Data Set:

The data set which was used to train and test the model was taken from UCI Machine Learning repository. In the repository, it is known as the “Gene Expression Cancer RNA-Seq Data”. Below is the screenshot of the data set currently being used in the model.

The data set used has 16384 columns out of which the first column alone being the label column with the classes and the rest 16383 being the features of gene expressions.

It consists of 16383 gene expressions named as gene_0, gene_1, gene_2 and so on till gene_16382.

The gene expressions play a vital role in the classification basis of the model and cannot be reduced or formatted as their originality and integrity will give the most accurate result.

Visualizations:

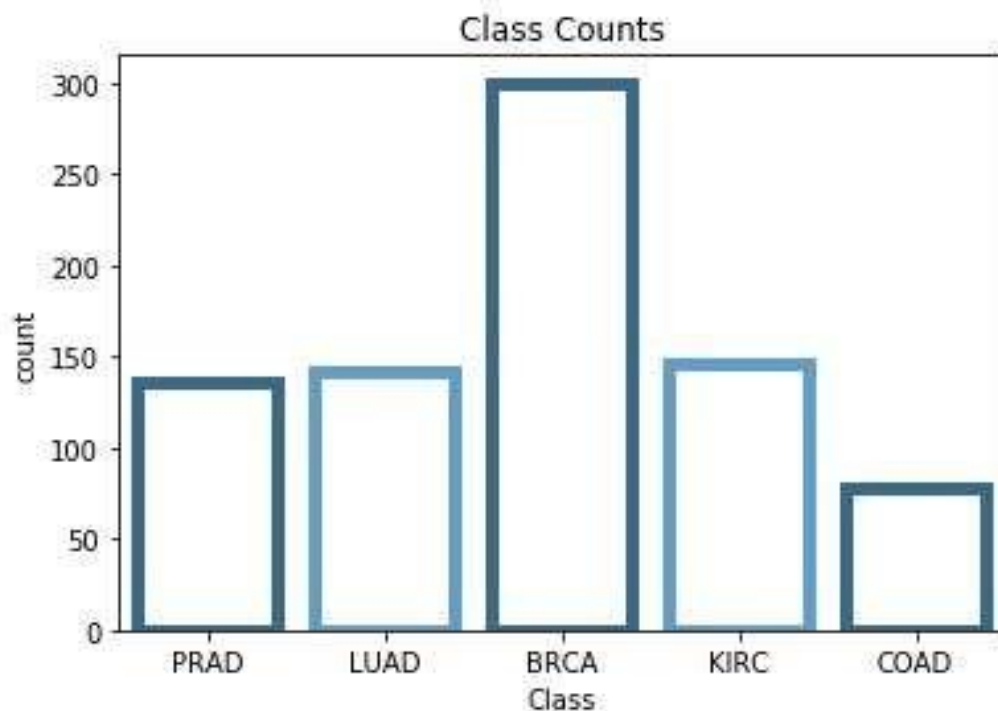


Fig 4.1 Distribution of dataset

The above figure shows the distribution of data amongst the 5 classes present. It is evident that the number of Breast Cancer cases in the data set is high which is directly due to the fact that Breast Cancer is the most commonly occurring cancer out of the taken 5.

The distribution cannot be made even by simulated data, though it is possible. This is because of the fact that the dataset is biological data and simulated data will not and does not conform to the forms, structures and patterns that any biological real-time data would conform to. This is why no efforts to even the distribution has been taken.

As a result of this uneven distribution, there occurs higher bias towards the class BRCA. This can be fixed or rectified to a level without changing the distribution of the data by the fixing the hyperparameters of the model to be used accordingly such that the weights that may be put are distributed and suggests that all the classes present in this dataset are equal viable and possible to occur with no partiality towards one.

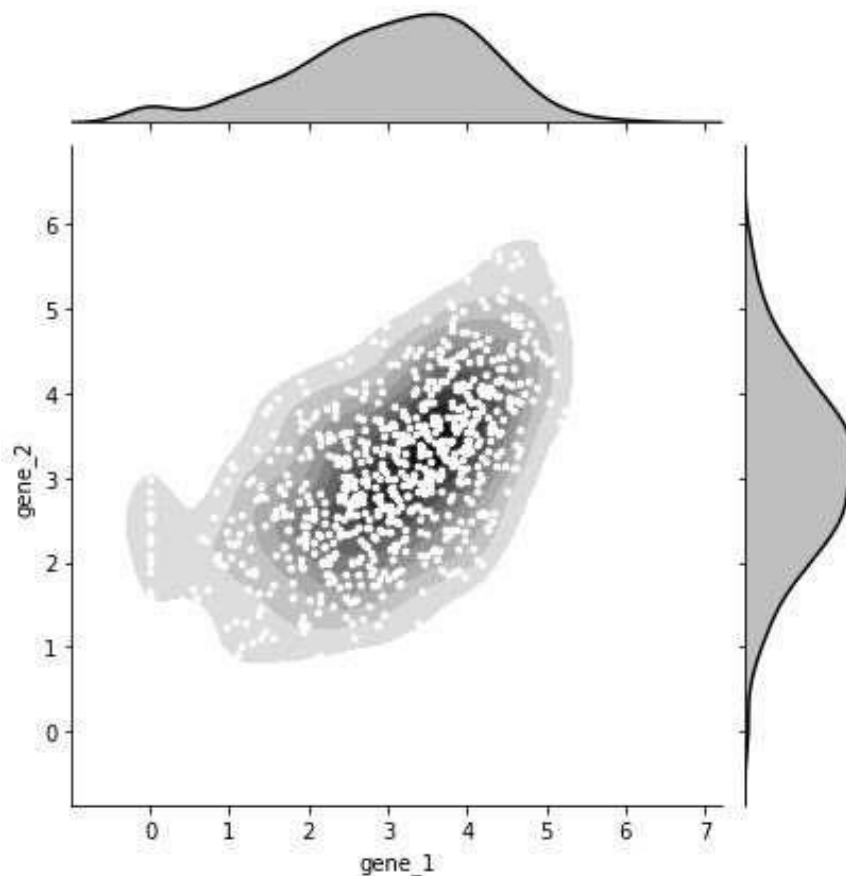


Fig 4.2 Density plot of gene_1 vs gene_2

The above figure shows a comparison between two genes 1 and 2 as an example of how genetic expression of one gene may affect or correlate with the other.

The various shades in the box indicate the regions where the correlation of the data of the gene expressions of gene 1 and 2 are low (lighter shade of grey) and the areas or values where the values are more compactly packed or dense in nature indicating possible correlation, but not causation (darker shades of grey).

The outer distribution graphs for each gene show how the genes are distributed in the data as some genes may be normally distributed by default.

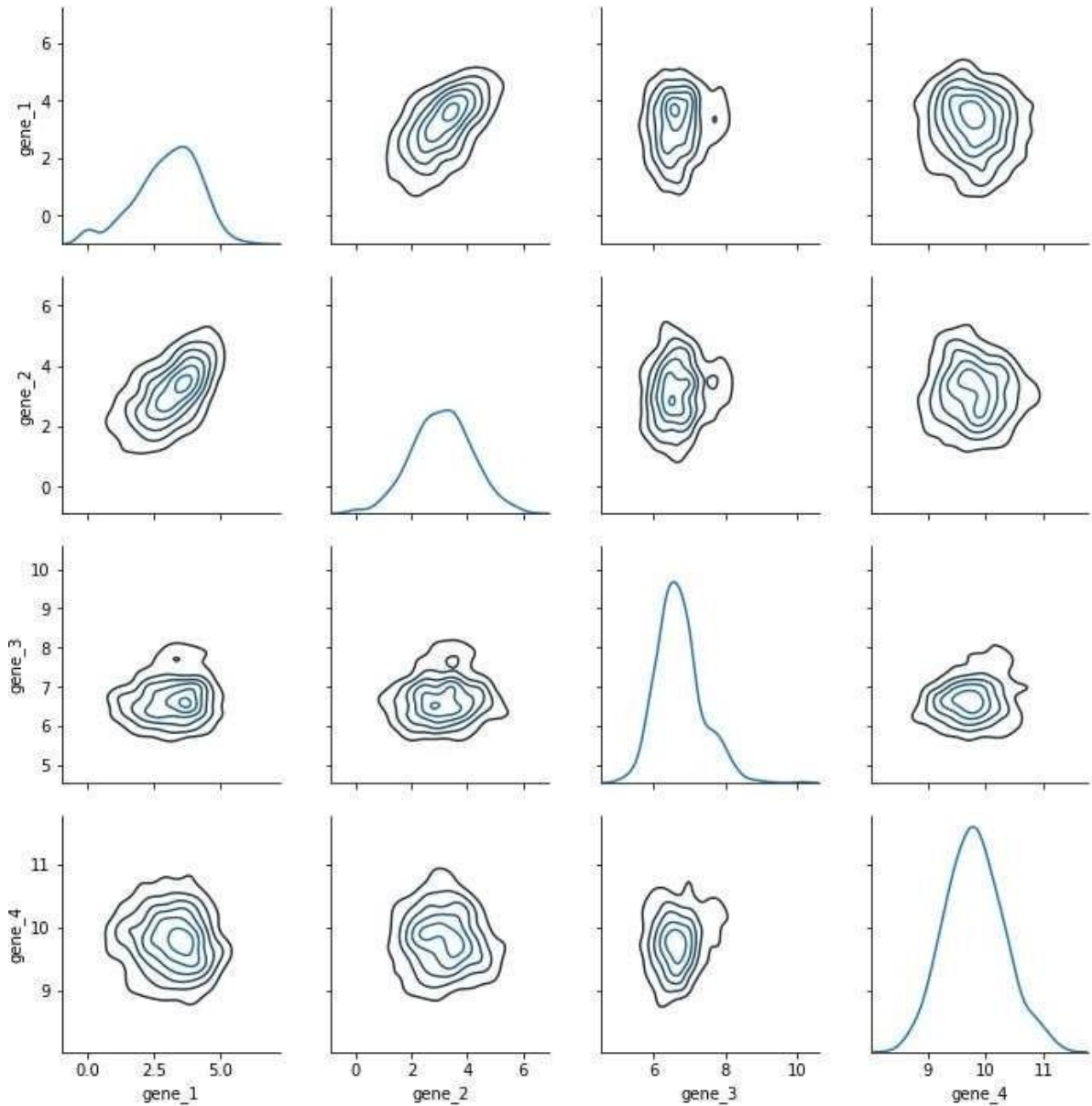


Fig 4.3 Cumulative density plot of 4 features

The above figure is a cumulative of the previous figure, meaning, it contains the comparative density plots of various genes at the same time.

Here, it is notable that gene 4 presents with a near-identical representation of normal distribution. This refers that some genes in the data set do conform to normal distribution and so the dataset cannot be duplicated or simulated as the distribution may be lost or altered with.

All the above representations, graphs and comparative studies can be performed over any set of genes/ features. The above are given to establish that such studies can be performed in this dataset.

4.1.2 Model Information Visualization

In this part of the chapter, the visualizations with respect to the model will be discussed.

Confusion Matrix:

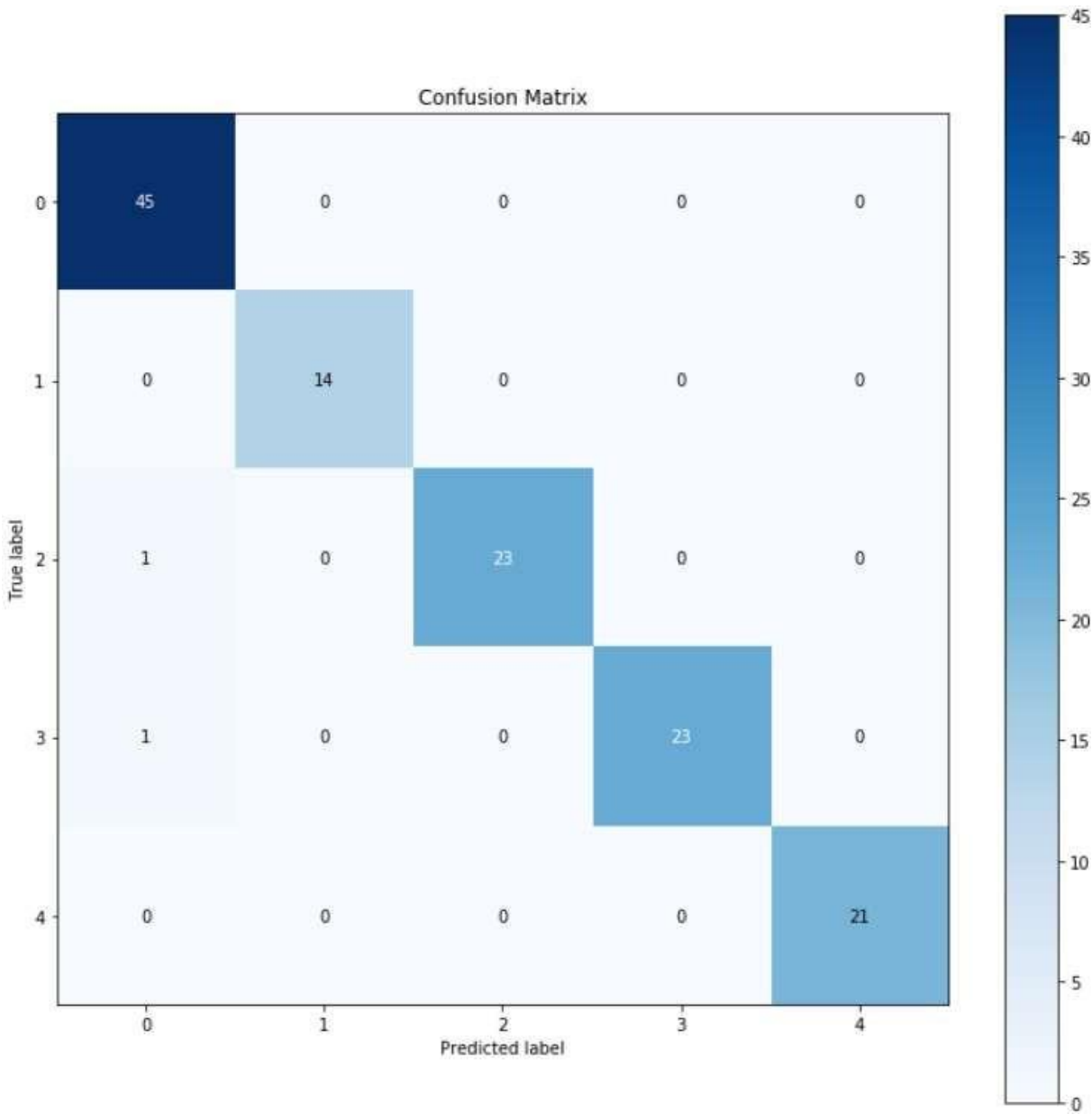


Fig 4.4 Confusion Matrix

BRCA-0 LUAD-1 PRAD-2 KIRC-3 COAD-4

The above figure provides data about the correct and misclassifications. Out of the training data used, it has misclassified under 5 instances all as BRCA. This

is clearly due to the bias that has been initially discussed. No other misclassifications have occurred hence proving the integrity and reliability of the model.

Classification Report:

	precision	recall	f1-score	support
PRAD	1.00	1.00	1.00	45
KIRC	1.00	1.00	1.00	14
COAD	1.00	0.96	0.98	24
LUAD	1.00	0.96	0.98	24
BRCA	1.00	1.00	1.00	21
micro avg	1.00	0.98	0.99	128
macro avg	1.00	0.98	0.99	128
weighted avg	1.00	0.98	0.99	128
samples avg	0.98	0.98	0.98	128

Fig 4.5 Classification Report

The above figure represents all the necessary factors taken to determine whether a model is good or bad. They are technically known as ***Model Evaluation Metrics***.

○ Precision:

$$\begin{aligned}
 \text{Precision} &= \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \\
 &= \frac{\text{True Positive}}{\text{Total Predicted Positive}}
 \end{aligned}$$

Fig 4.6 Precision formula

Precision is defined as the fraction of the examples which are actually positive among all the examples which we predicted positive.

It is also defined as the number of true positives divided by the number of true positives plus the number of false positives. False positives are

cases the model incorrectly labels as positive that are actually negative, or in our example, individuals the model classifies as terrorists that are not.

○ **Recall:**

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$
$$= \frac{\text{True Positive}}{\text{Total Actual Positive}}$$

Fig 4.7 Recall formula

The metric our intuition tells us we should maximize is known in statistics as recall, or the ability of a model to find all the relevant cases within a dataset.

The precise definition of recall is the number of true positives divided by the number of true positives plus the number of false negatives. True positives are data point classified as positive by the model that actually are positive (meaning they are correct), and false negatives are data points the model identifies as negative that actually are positive (incorrect).

○ **F1-Score:**

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Fig 4.8 F-1 Score formula

Traditionally, it is defined to be the harmonic mean of precision and recall.

It is also called the F Score or the F Measure. Put another way, the F1 score conveys the balance between the precision and the recall.

It is considered to be a better measure than Precision and Recall individually as the trade-off between the two is hard to achieve.

○ **Support:**

Support is pretty simply put just the number of data rows supporting the particular class in the trained set.

It is generally given to show out of what number the other metrics have been calculated.

4.1.3 Post-Model Data Visualization

The Post – Model Data Visualization section consists primarily of two visualizations:

- i. Important Features
- ii. A decision tree from the ensemble of Random Forest model

i. Important Features:

The important features section is an inbuilt feature of random forest in python's scikit-learn where the required 'n' number of top features playing a major role in the decision-making process of the model is given.

Here, the top 10 features or the top 10 genes affecting the output the most have been shown:

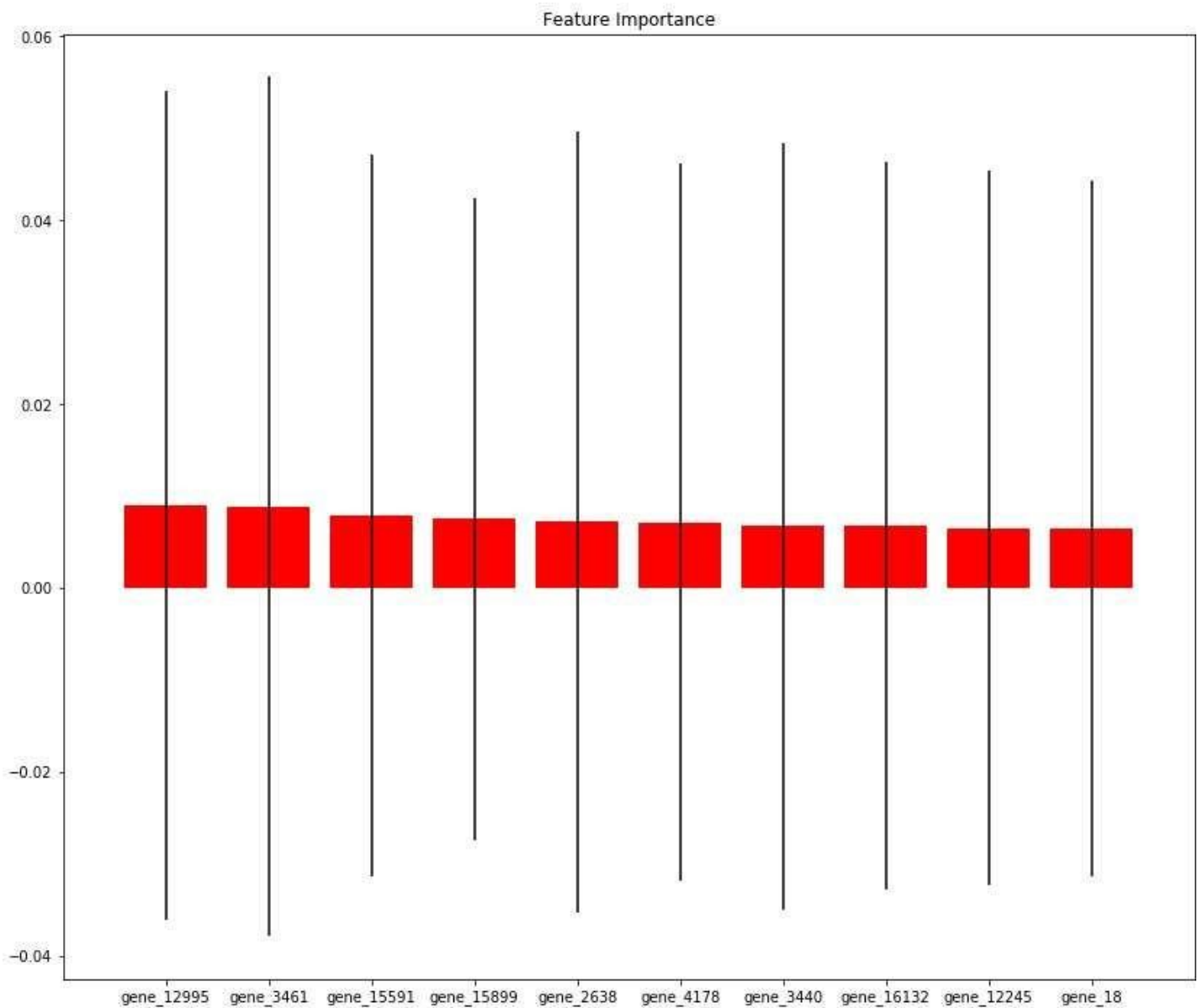


Fig 4.9 Box Graph of Important Features

ii. A Decision Tree from the Random Forest Model ensemble:

This section consists of an extraction of few decision trees from the Random Forest ensemble for visualization.

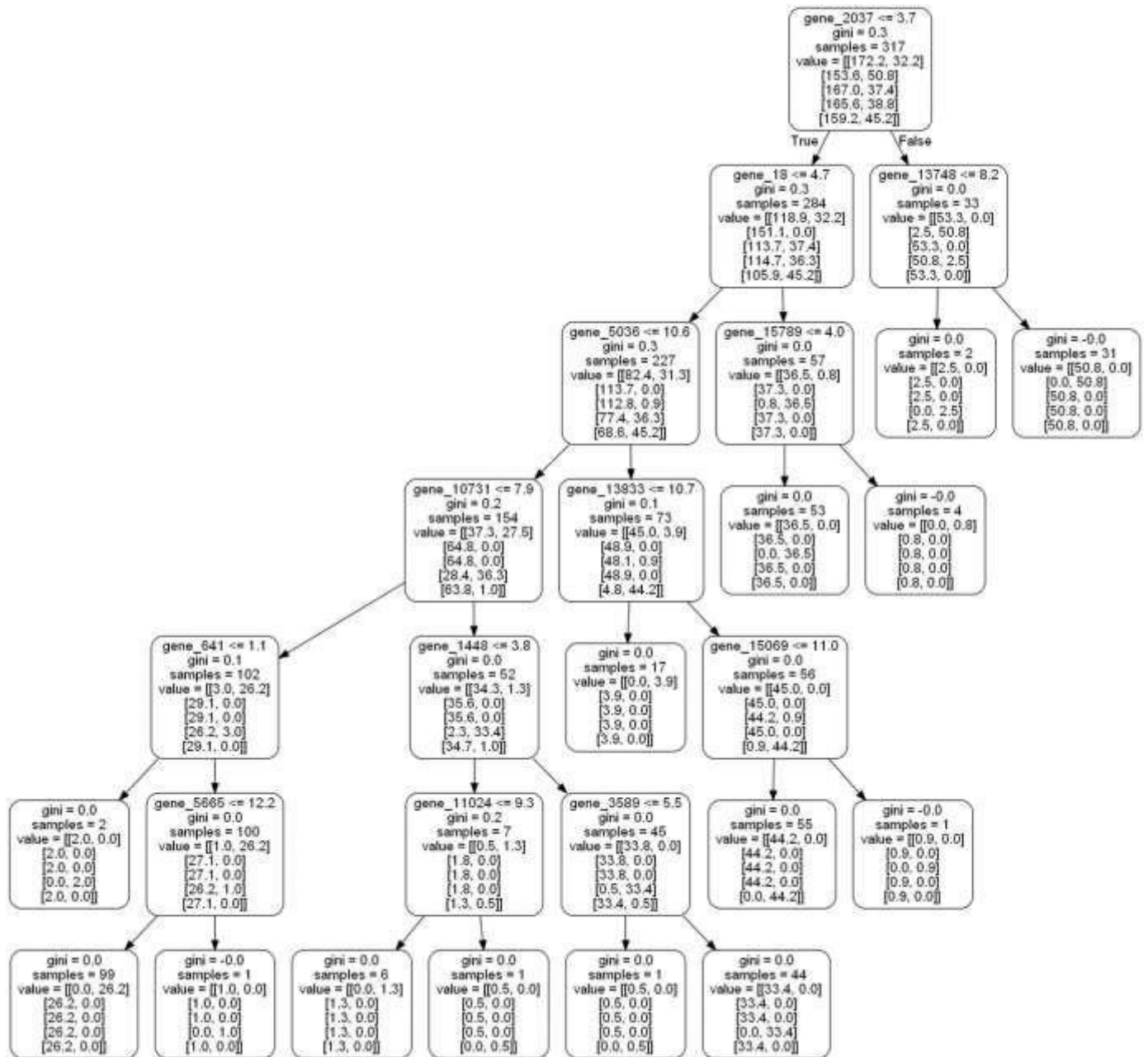


Fig 4.10 10th decision tree

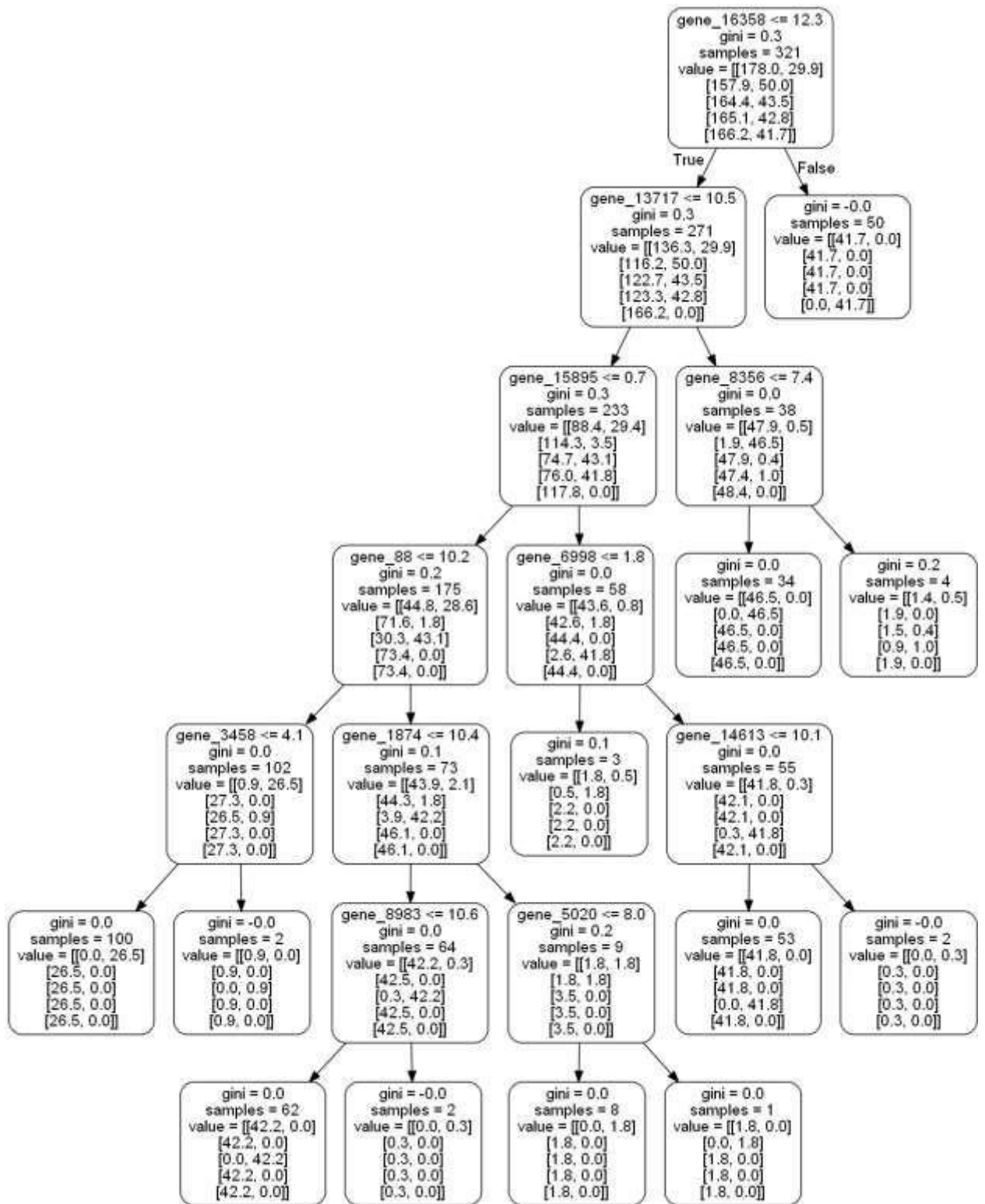


Fig 4.11 – 20th decision tree

4.2 ANALYSIS:

4.2.1 Research

There has been research done in this topic, but due to the decreased availability of real-time gene expression that may require Next Generation DNA Sequencing the research has been limited. Several methods have suggested the use of targeted a single gene for such purposes but have proved to be inefficient in front of the human body's complex biological functioning systems.

4.2.2 Proposed System

The proposed system provides an accuracy of 98.4%, an Out of the Bag (OOB Score), used for Random Forest models, of 0.99688 which is equivalent to 99.69% accuracy over model produced inputs and an R^2 score of 0.97963, which when nearer to 1 confirms of a stable model. The accuracy can be increased once reliable data with a more balanced dataset has been made.

The important features as predicted by the model are the genes that the model finds to be the most influential in its decision making. Some of these genes (due to spatial constraints) when plotted as their averages show the below graph as the result for all the five types of cancers analysed in this model.

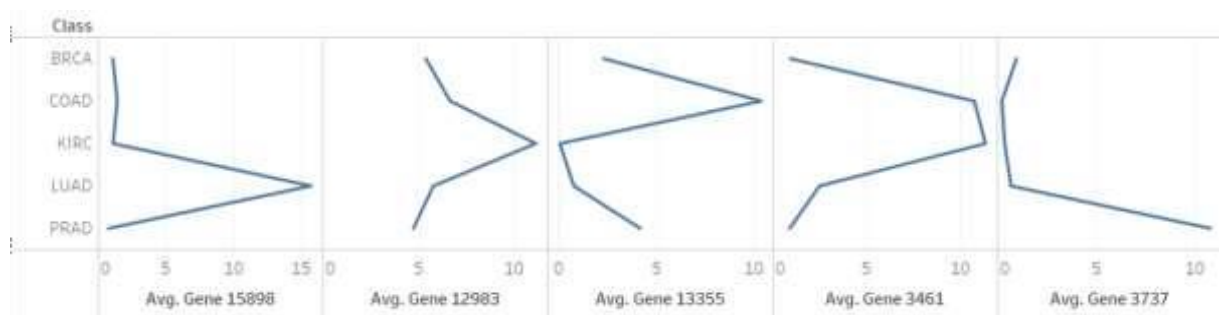


Fig 4.12 Line Graph - Averaged important features vs classes

The above graph has been created using the Data Visualization tool **Tableau**. The Desktop version of Tableau has been used to create this graph. The data was uploaded to the software and manually plotted using the outputs of important features provided by the model. The graph shows the

abnormalities of specific genes in specific types of cancer. Their expression levels fall under a certain limit for almost 4 out of 5 of the cancers but tend to deviate in their trend for one or two specific types of cancers. This may be used as an underlying proof to using patterns of gene expression values for cancer diagnosis

The proposed system uses the Jupyter Notebook IDE used in the Anaconda Environment for the building of the model itself and hence may require certain software requirements. The requirements for hardware is almost non-existent due to the fact that Jupyter Notebook runs on the internet. Hence, **Network Stability** plays a major role in the use of Jupyter Notebook.

4.3 SYSTEM REQUIREMENTS

4.3.1 System Requirements for Anaconda Free Distribution

- License: Free use and redistribution under the terms of the End User License Agreement.
- Operating system: Windows 7 or newer, 64-bit macOS 10.10+, or Linux, including Ubuntu, RedHat, CentOS 6+, and others.

If your operating system is older than what is currently supported, you can find older versions of the Anaconda installers in our archive that might work for you. Check our FAQ for version recommendations.

- System architecture: Windows- 64-bit x86, 32-bit x86; MacOS- 64-bit x86; Linux- 64-bit x86, 64-bit Power8/Power9.
- Minimum 5 GB disk space to download and install.

4.3.2 System Requirements for Tableau Desktop

Windows:

- Microsoft Windows 7 or newer (32-bit and 64-bit)
- Microsoft Server 2008 R2 or newer
- Intel Pentium 4 or AMD Opteron Processor or newer

- 2 GB memory
- 1.5 GB minimum free disk space

Mac:

- iMac/MacBook computers 2009 or newer
- macOS High Sierra 10.13, macOS Mojave 10.14 and macOS Catalina 10.15
- 1.5 GB minimum free disk space

Virtual Environments:

- Citrix environments, Microsoft Hyper-V, Parallels, VMware, Microsoft Azure and Amazon EC2.
- All of Tableau's products operate in virtualized environments when they are configured with the proper underlying Windows operating system and minimum hardware requirements.

CHAPTER – 5

SUMMARY AND FUTURE SCOPE

5.1 SUMMARY:

This project was created to further create new opportunities for diagnosing cancer in early stages or even before its occurrence. The concept of genetic mutations and increased gene expression values to diagnose or detect is a long thought of topic. But due to the difficulties faced for collecting proper and reliable data to process and produce optimal results have been a major drawback for the research. The processing behind extracting data requires actual genetic information from humans who have suffered or are suffering from cancer and advanced techniques such as Next Generation Gene Sequencing using modern tools such as BioPython.

The difficulties faced has caused this efficient idea to be only researched upon in limited terms. The success of this research, included that reliable data is sought, can produce revolutionary results and may change the course of how cancer is being detected or diagnosed in the future

5.2 FUTURE SCOPE:

I am thankful for this opportunity to work on this project. This project is one of my long thought research ideas which may turn to be useful. The future of this project has several opportunities.

This project, by my plan, is to be extended for further research with respect to my goal of establishing newer, more reliable and flexible methods of detecting cancer using Gene Expression patterns. Gene Expression patterns are a very reliable source of information as that is what forms human characteristics themselves. This inter-disciplinary approach may yield a great result if treaded rightly.

In other ways, this project can be used in conjunction with other cancer detection methods to confirm of the type of cancer, more specifically the area where the primary tumor is present, without heavy efforts and a loss of hefty amount of time as time is of essence when cancer is considered.

References

1. Joseph M. De Guaia, Madhavi Devaraj and Larry A. Vea, "Cancer Classification of Gene Expression Data using Machine Learning Models", 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), 18512214 (2018) - <https://ieeexplore.ieee.org/document/8666435>
2. TaeJin Ahn, Taewan Goo, Chan-hee Lee, SungMin Kim, Kyullhee Han, Sangick Park, TAesung Park, "Deep Learning-based Identification of Cancer or Normal Tissue using Gene Expression Data", 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 18411837 (2018). - <https://ieeexplore.ieee.org/document/8621108>
3. Qun-Xiong ZHu, Yuan Fan, Yan-Lin He and Yuan Xu, "Effective Cancer Classification based on Gene Expression Data using Multidimensional Mutual Information and ELM", 2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS), 18200778 (2018). - <https://ieeexplore.ieee.org/document/8515927>
4. Behrouz Shamsaei and Cuilan Gao, "Comparison of some machine learning and statistical algorithms for classification and prediction of human cancer type", 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics, 978-1-5090-2455-1 (2016). - <https://ieeexplore.ieee.org/document/7455893>
5. Chen Shen and Zhi-Ping Liu, "Identifying module biomarkers of hepatocellular carcinoma from gene expression data", 2017 Chinese Automation Congress (CAC), 978-1-53856-3524-7 (2017). - <https://ieeexplore.ieee.org/document/8243741>

6. Saima Rathore, Mutawarra Hussain and Asifullah Khan, "GECC: Gene Expression Based Ensemble Classification of Colon samples", IEEE/ACM Transactions on Computational Biology and Bioinformatics (Volume: 11, Issue:6), 1557-9964 (2014). - <https://ieeexplore.ieee.org/document/6872581>
7. Vitor Teixeira, Rui Camacho and Pedro G. Ferreira, "Learning influential genes on cancer gene expression data with stacked denoising autoencoders", 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 976-1-5050-3050-7 (2017) - <https://ieeexplore.ieee.org/document/8217828>
8. Muxuan Liang, Zhizhong Li, Ting Chen and Jianyang Zeng, "Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach", IEEE/ACM Transactions on Computational Biology and Bioinformatics (Volume:12, Issue: 4), 1557-9964 (2015). - <https://ieeexplore.ieee.org/document/8217828>
9. Yuhang Wang and F. Makedon, "Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data", 2004 IEEE Computational Systems Bioinformatics Conference (CBS), 0-7695-2194-0 (2004). - <https://ieeexplore.ieee.org/document/1332474>
10. Ujjwal Maulik, Anirban Mukhopadhyay and Debasis Chakraborty, "Gene-Expression-Based Cancer Subtypes Prediction Through Feature Selection and Transductive SVM", IEEE Transactions on Biomedical Engineering (Volume: 60, Issue: 4), 1558-2531 (2013). - <https://ieeexplore.ieee.org/document/6334430>
11. Ambrosio Hernandez, Farin Smith, QinDing Wang, Xiaofu Wang and B. Mark Evers, "Assessment of differential Gene Expression Patterns in Human Colon Cancers", Annals of Surgery, PMC1421190 (2000). - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1421190/>

12. Sara Alghunaim and Heyam H. Al-Baity, "On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context", IEEE Access (Volume: 7), 2169-3536 (2019). - <https://ieeexplore.ieee.org/document/8755975>

13. Jayadeep Pati, "Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach", IEEE Access (Volume: 7), 2169-3536 (2019). - <https://ieeexplore.ieee.org/document/8584430>

14. Jing Xu, Peng Wu, Yuehui Chen, Qinfang Meng, Hussain Dawood and Muhammed Murtaza Khan, "A Novel Deep Flexible Neural Forest Model for Classification of Cancer Subtypes Based on Gene Expression Data", IEEE Access (Volume: 7) (2019). - <https://ieeexplore.ieee.org/document/8640828>

15. Konstantina Kourou, Themis P.Exarchos, Konstantinos P.Exarchos, Michalis V. Karamouzis, Dimitros I.Fotiadis, "Machine learning applications in cancer prognosis and prediction", Computational and Structural Biotechnology Journal, Volume: 13, 2015, Pages 8 – 17. - <https://doi.org/10.1016/j.csbj.2014.11.005>

16. "Machine learning in cancer diagnostics", EBioMedicine, Volume 45, PI-2 (2019), PMC6642356. - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6642356/>

17. David G. P. van IJzendoorn, Karoly Szuhai, Inge H. Briaire-de Bruijn, Marie Kostine, Marieke L.Kuijjer, Judith V. M. G. Bovee, "Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas", PLOS Computational Biology, Version 2 (2019). - <https://doi.org/10.1371/journal.pcbi.1006826>

18. Tan AC, Gilbert D., "Ensemble machine learning on gene expression data for cancer classification", PMID: 15130820 (2003) - <https://www.ncbi.nlm.nih.gov/pubmed/15130820>

19. Sara Tarek, Reda Abd Elwahab, Mahmoud Shoman, "Gene expression based cancer classification", Egyptian Informatics Journal, Volume: 18, Pages 151-159 (2017) - <https://doi.org/10.1016/j.eij.2016.12.001>
20. Machine Learning Is the Future of Cancer Prediction, Towards Data Science - <https://towardsdatascience.com/machine-learning-is-the-future-of-cancer-prediction-e4d28e7e6dfa>
21. Data set - <https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq>
22. Anaconda Free Distribution - <https://www.anaconda.com/>
23. Jupyter Notebook - <https://jupyter.org/>
24. Data Science - https://en.wikipedia.org/wiki/Data_science
25. Machine Learning - https://en.wikipedia.org/wiki/Machine_learning
26. Scikit-learn - <https://scikit-learn.org/stable/>
27. Pandas - <https://pandas.pydata.org/>
28. Matplotlib - <https://matplotlib.org/>
29. Seaborn - <https://seaborn.pydata.org/>
30. Random Forest - https://en.wikipedia.org/wiki/Random_forest

APPENDIX

SOURCE CODE

#Packages required

```
import pandas as pd

import numpy as np

from sklearn.model_selection import train_test_split, RandomizedSearchCV

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification_report, r2_score

from sklearn.tree import export_graphviz

import pydot

import matplotlib.pyplot as plt

import scikitplot as skplt

import seaborn as sns

from numpy import inf

from sklearn.metrics import accuracy_score
```

#Import dataset from local folder

```
data= pd.read_csv("C:/Users/USER/Desktop/Final Project/Data Sets/TCGA-
PANCAN-HiSeq-801x20531/data.csv")
```

#Example Visualizations of the dataset

##Count Plot of Classes

```
ax=sns.countplot(x="Class", data=data, facecolor=(0,0,0,0), linewidth=5,
edgecolor=sns.color_palette("Blues_d",2))
```

##Normal Plot (Scatter and Bar)

```
sns.pairplot(data.loc[:, data.columns != "Class"].iloc[:,1:4])

g1 = sns.PairGrid(data.loc[:, data.columns != "Class"].iloc[:,1:5])

g1.map_diag(sns.kdeplot)

g1.map_offdiag(sns.kdeplot, n_levels=6)

g2 = sns.jointplot(x="gene_1", y="gene_2", data=data, kind="kde", color="k")

g2.plot_joint(plt.scatter, c="w", s=30, linewidth=1, marker=".")

g2.ax_joint.collections[0].set_alpha(0)
```

#Separating features and labels

```
y= data.Class

label=set(y)

x=data.loc[:, data.columns!="Class"]

labels=np.array(data["Class"])

labels=pd.get_dummies(labels)

features= np.array(x)

features_list=list(x.columns)
```

#Creating Training, Validation and Testing Sets

```
train_val_features, testing_features, train_val_labels, testing_labels =
train_test_split(features, labels, test_size = 0.2, random_state = 3)

testing_train_features,val_features,testing_train_labels,val_labels=
train_test_split(train_val_features, train_val_labels, test_size = 0.2, random_state
= 3)
```

#Creating dummy Random Forest model to find best parameters

```
rf_dummy=RandomForestClassifier()
```

#Finding best parameters using RandomizedSearchCV

##Parameter choices

```
n_estimators = [int(x) for x in np.linspace(start = 100, stop = 500, num = 10)]

max_features = ["auto", "sqrt"]

max_depth = [int(x) for x in np.linspace(100, 500, num = 11)]

max_depth.append(None)

min_samples_split = [2, 5, 10]

min_samples_leaf = [1, 2, 4]

bootstrap = [True]

criterion=["gini","entropy"]

oob_score=[True,False]

warm_start=[True,False]

class_weight=[None,"balanced"]

random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap,
               'criterion': criterion,
               'oob_score': oob_score,
               'warm_start': warm_start,
               'class_weight': class_weight
              }
```

##Testing for best parameters

```
rf_random = RandomizedSearchCV(estimator = rf_dummy, param_distributions =  
random_grid, n_iter = 100, cv = 3, verbose=2, random_state=3, n_jobs = -1)
```

##Fitting the training set

```
rf_random = RandomizedSearchCV(estimator = rf_dummy, param_distributions =  
random_grid, n_iter = 100, cv = 3, verbose=2, random_state=3, n_jobs = -1)
```

```
rf_random.fit(testing_train_features,testing_train_labels)
```

#Real Random Forest model with the best parameters applied

```
rf = RandomForestClassifier(n_estimators=144, min_samples_split=5,  
min_samples_leaf=1, max_features="auto", max_depth=380, bootstrap=True,  
oob_score=True, verbose=1, warm_start=True, random_state=3, criterion="gini",  
class_weight="balanced", n_jobs=-1)
```

##Fitting the training Set

```
rf= RandomForestClassifier(n_estimators=277, min_samples_split=5,  
min_samples_leaf=1, max_features="auto", max_depth=420, bootstrap=True,  
oob_score=True, verbose=2, warm_start=True, random_state=3,  
criterion="entropy", class_weight="balanced", n_jobs=-1)
```

```
rf.fit(testing_train_features, testing_train_labels)
```

```
rf_predict = rf.predict(val_features)
```

##Testing the model

```
test_score=rf.score(testing_features, testing_labels)
```

```
print(test_score)
```

##Out Of the Bag Score of the model

```
oob_score=rf.oob_score_
```

```
print(oob_score)
```

##Top 10 important features

```
skplt.estimators.plot_feature_importances(rf,  
feature_names=list(data.columns[1:]), max_num_features=10, figsize=(12,12))
```

##R-2 Score

```
r_score=r2_score(val_labels,rf_predict, multioutput="variance_weighted")  
  
print(r_score)
```

##Confusion matrix map

```
skplt.metrics.plot_confusion_matrix(val_labels.values.argmax(axis=1),  
rf_predict.argmax(axis=1),figsize=(12,12))
```

##Classification Report consisting of Precision, Recall and F1-score

```
class_rep=classification_report(val_labels, rf_predict,target_names=label)  
  
print(class_rep)
```

#Extracting a tree for visualization

```
tree=rf.estimators_[10]
```

```
export_graphviz(tree,out_file="tree_sample.dot",feature_names=features_list,  
rounded=True, precision=1)
```

```
(graph, )=pydot.graph_from_dot_file("tree_sample.dot")
```

```
graph.write_png('tree_sample10.png')
```

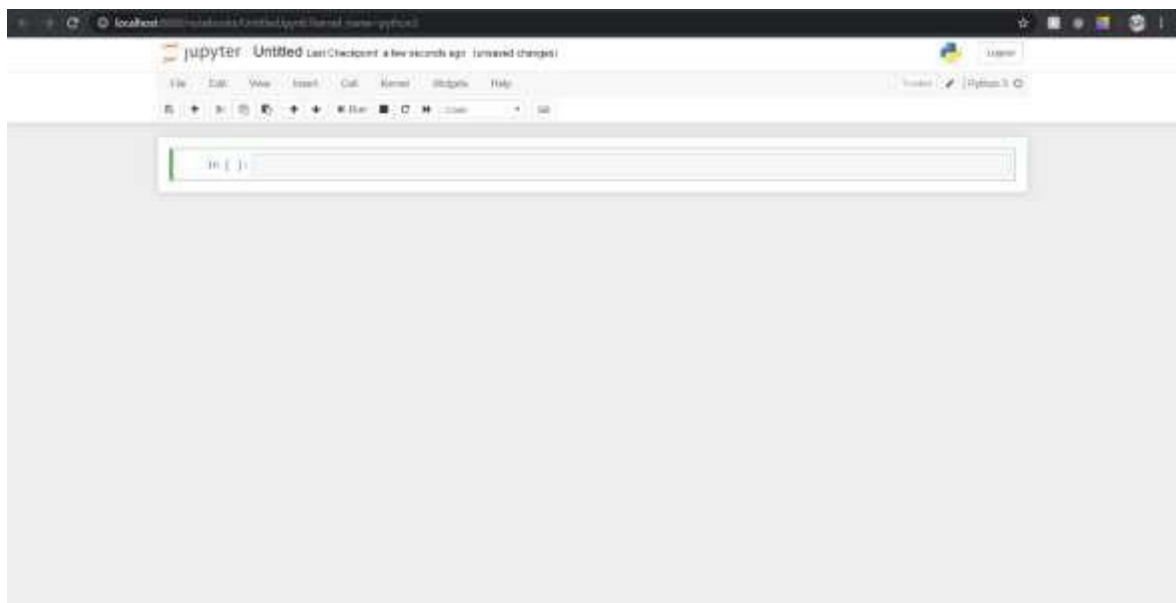
#Accuracy Score of the model

```
accuracy_score(val_labels,rf_predict,normalize=True)
```

SCREENSHOTS



Jupyter Notebook Home Page



Jupyter Notebook Environment

The image shows a Microsoft Excel spreadsheet with a large dataset. The columns are labeled 'gene_1' through 'gene_27'. The rows contain numerical data, likely representing gene expression levels or similar biological data. The spreadsheet is displayed in a standard Excel interface with a ribbon at the top and a formula bar.

Dataset Preview

The image shows the Tableau Home Page. It features a sidebar with navigation options like 'Connect', 'Open', 'Discover', and 'Sample Workbooks'. The main area displays a grid of sample workbooks, including 'Supersize', 'Regional', and 'World Indicators'. The 'Connect' sidebar lists various data sources such as Microsoft Excel, Text File, JSON, Microsoft Access, PostgreSQL, and others. The 'Discover' section provides information about Tableau's capabilities and offers a 'Get Tableau Free' button.

Tableau Home Page