**VISVESVARAYA TECHNOLOGICAL UNIVERSITY**

**JNANASANGAMA, BELAGAVI - 590018**

[](https://www.google.co.in/imgres?imgurl=http://i727.photobucket.com/albums/ww276/sankethkumar/vtu_logo.jpg&imgrefurl=http://photobucket.com/images/vtu%20logo&docid=7vJL4VgBOFzETM&tbnid=dQewAEkp0o9foM:&vet=10ahUKEwi50anvqdXTAhWINY8KHXN8CooQMwgqKAEwAQ..i&w=269&h=320&bih=805&biw=1600&q=vtu%20symbol&ved=0ahUKEwi50anvqdXTAhWINY8KHXN8CooQMwgqKAEwAQ&iact=mrc&uact=8)

An Internship Report

**MACHINE LEARNING**

Submitted in partial fulfillment of the requirement for the award of the

Degree of Bachelor of Engineering

in

Computer Science & Engineering

Submitted by

**RAKSHITHA S**

**1AT16CS081**

**Internship Carried Out**

**TechCiti Technologies Private Limited**

**Internal Guide External Guide**

**Mr. Vijay Swaroop A Mr. HARISH L**

**Associate Professor Software Associate Trainee**

**Atria Institute of Technology**

A close up of a logo

Description automatically generated

**ATRIA INSTITUTE OF TECHNOLOGY**

**Department of Computer Science & Engineering**

**Anandanagar, Bangalore-560024**

**2019-20**

A close up of a logo

Description automatically generated

**ATRIA INSTITUTE OF TECHNOLOGY**

**ANANDNAGAR, BENGALURU – 560 024**

**CERTIFICATE**

Certified that this internship work on “**CLASSIFICATION OF CANCEROUS STAGES PROFOLES USING MACHIINE LEARNING**” presented by **RAKSHITHA S (1AT16CS081)** of Atria Institute of Technology, Bangalore in partial fulfillment for the award of Bachelor of Engineering in Computer science & Engineering of Visvesvaraya Technological University, Belagavi during 2019-20. It is certified that all corrections/ suggestions indicated during internal assessment have been incorporated in the report. The internship report has been approved and it satisfies the academic requirements with respect to the internship report as prescribed for the said Degree.

Signature of Guide Signature of Coordinator

Mr. Vijay Swaroop Mr. Vijay Swaroop

Signature of HOD Signature of Principal

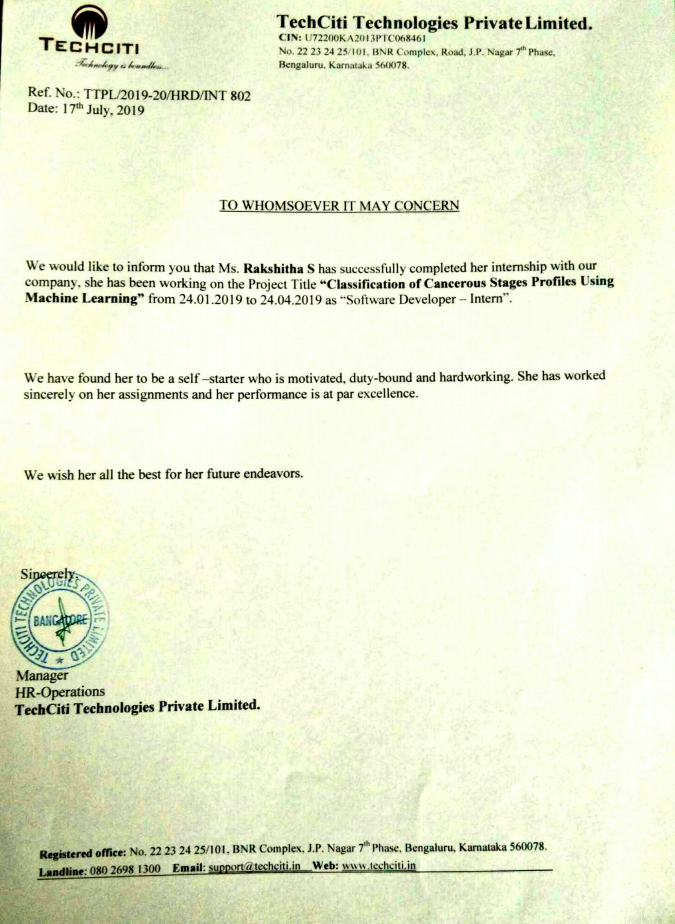
Dr. Aishwarya P Dr. K. V. Narayanaswamy

**External-Viva**

(Name of Internal /External Examiner with Signature & Date)

Internal Examiner:

External Examiner:

****

**DECLARATION**

I RAKSHITHA S (1AT16CS081**),** the student of Eighth Semester, Department of Computer Science & Engineering, Atria Institute of Technology, hereby declare that this internship work on **“CLASSIFICATION OF CANCEROUS STAGES PROFOLES USING MACHIINE LEARNING”** has been carried out by us, under the guidance of Mr. Vijay Swaroop Assoc. Professor, Dept, of CS&E. This internship work is submitted to Visvesvaraya Technological University in partial fulfillment of the requirement for the award of degree of **Bachelor of Engineering in Computer Science & Engineering** for the academic year 2019-2020.

Place: Bangalore \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: Signature of Student

**ACKNOWLEDGMENT**

The foundation for any successful venture is laid out not just by the individual accomplishing the task, but also by several other people who believe that the individual can excel and put in their every bit in every endeavor he/she embarks on, at every stage in life. And the success is derived when opportunity meets preparation, also supported by a well-coordinated approach and attitude.

I would like to express my sincere gratitude to the respected principal **Dr. K.V. Narayanaswamy,** for providing a congenial environment to work in. We also like to express our sincere gratitude to **Dr. Aishwarya.P,** Head of Department, Computer Science, for her continuous support and encouragement.

I am indeed indebted to **Mr. Vijay Swaroop,** ourcoordinator andguide, for his continued support, advice and valuable inputs during the course of this internship work.

Last, but not the least I would like to thank my family who have acted as a beacon of light throughout our life.

My sincere gratitude goes out to all my comrades and well-wishers who have supported me through all the ventures.

**EXECUTIVE SUMMARY**

Machine learning is a technology that has grown to prominence over the past ten years (as at this time of writing) and is fast paving the way for the “Age of Automation”. In a nutshell, machine learning primarily deals with prediction and classification. Machine learning algorithms build a [mathematical model](https://en.wikipedia.org/wiki/Mathematical_model) based on sample data, known as "[training data](https://en.wikipedia.org/wiki/Training_data)", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as [email filtering](https://en.wikipedia.org/wiki/Email_filtering) and [computer vision](https://en.wikipedia.org/wiki/Computer_vision), where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks. Machine learning is closely related to [computational statistics](https://en.wikipedia.org/wiki/Computational_statistics), which focuses on making predictions using computers. In its application across business problems, machine learning is also referred to as [predictive analytics](https://en.wikipedia.org/wiki/Predictive_analytics).

There are wide various approaches of ML involved such as Supervised ML, Unsupervised ML, Semi-supervised learning, Reinforcement learning, Self learning, Feature learning, Sparse – dictionary learning, Anomaly detection, Robot leaning, Association rules etc. But here the approach used is supervised machine learning and the algorithms come under it.

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as [training data](https://en.wikipedia.org/wiki/Training_data), and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an [array](https://en.wikipedia.org/wiki/Array_data_structure) or vector, sometimes called a feature vector, and the training data is represented by a [matrix](https://en.wikipedia.org/wiki/Matrix_(mathematics)). An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task. Types of supervised learning algorithms include [Active learning](https://en.wikipedia.org/wiki/Active_learning_(machine_learning)), [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis).  Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range.

There are a variety of options available for cancer treatment. The type of treatment recommended for an individual is influenced by various factors such as cancer-type, severity of cancer (stage) that is either MALIGN or BENIGN and recommended therapies required. To study anti-cancer therapies we need to understand cancerous profiles. These profiles carry information which can reveal the underlying factors responsible for cancer growth. Hence there is a need to analyse cancer data for predicting optimal optimal-treatment options and detect survival period and death period of a patient. Therefore analysis of profiles can help predicting and discover potential therapies and survival. In this paper the main aim is to provide ML based classification technique for cancerous profiles.

Scikit-learn comes with a few small standard datasets that do not require downloading any file from any external website. The dataset that we will be using for our machine learning problem is the Breast cancer wisconsin (diagnostic) dataset. The dataset includes several data about the breast cancer tumors along with the classifications labels, viz., malignant or benign.

Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms.

**TABLE OF CONTENTS**

**EXECUTIVE SUMMARY** **A-B**

LIST OF FIGURES E

LIST OF TABLES F

**CHAPTER 1. ABOUT THE ORGANIZATION 1-5**

1.1 INTRODUCTION 1

1.2 BACKGROUND OF INTERNSHIP 1

1.3 OBJECTIVES OF INTERNSHIP 1

1.4 BACKGROUND OF THE ORGANIZATION 1

1.5 OBJECTIVES AND SERVICES 2

1.6 COMPANY SIZE AND SHARE CAPITAL 4

1.7 ORGANIZATIONAL CULTURE 4

1.8 MILESTONES 5

**CHAPTER 2. ABOUT THE DEPARTMENT 5-9**

2.1 TECHCITI DEPARTMENT 5

2.1.1 MACHINE LEARNING DEPT 6

2.1.2 WEB DEVELOPMENT DEPT 7

2.1.3 PYHTON WITH DJANGO DEPT 7

2.1.4 JAVA WITH SPRING DEPT 7

2.1.5 HUMAN RESOURCE DEPARTMENT 8

2.1.6 FINANCE DEPARTMENT 9

**CHAPTER 3. TASK PERFORMED 10-23**

3.1 RECORD OF CHANGES 10

3.2 ANALYZING THE PLAN 11

**CHAPTER 4. REFLECTION NOTES (SPECIFIC OUTCOMES) 23-29**

4.1 LEARNING IT AREA OF INTEREST 23

4.2 WORKING WITH EXPERIENCED INDIVIUALS 23

4.3 BENEFITS FROM INTERNSHIP 24

4.4 KNOWLEDGE AND SKILLS GAINED 25

4.5 SNAPSHOTS 25

**LIST OF REFERENCES 30**

**LIST OF FIGURES**

FIGURE 3.2.8.1 SCATTER PLOT 19

FIGURE 3.2.8.2 PAIR PLOT 19

FIGURE 3.8.2.3 BOX PLOT 20

FIGURE 3.8.2.4 VIOLIN PLOT 20

FIGURE 3.8.2.5 DISTRIBUTION PLOT 21

FIGURE 3.8.2.6 BAR CHART 21

FIGURE 3.8.2.7 LINE CHART 22

FIGURE 4.5.1 DATASET COLLECTION 25

FIGURE 4.5.2 PREDICTING EITHER MALIGN / BENIGN 25

FIGURE 4.5.3 THERAPIES THROUGH MATRIX 26

FIGURE 4.5.4 PLOT OF DATASET 26

FIGURE 4.5.5 PLOT OF CLINICAL VARIABLES & PROGNOSIS 27

FIGURE 4.5.6 SURVIVAL PLOT 27

FIGURE 4.5.7 PLOT OF CLINICAL VARIABLES 27

FIGURE 4.5.8 CLINICAL THERAPIES PLOT 28

FIGURE 4.5.9 PLOT IN HISTOPATHOLOGIICAL CLASS 28

FIGURE 4.5.10 SEABORN PLOT INVARIANT REALTIONSHIP 29

**LIST OF TABLES**

TABLE 1.6.1 SHARE CAPITAL & NO. OF EMPLOYEES 4

TABLE 1.6.2 LISTING & ANNUAL COMPLIANCE DETAILS 4

TABLE 3.1 RECORD OF CHANGES 10

**CHAPTER 1**

**ABOUT THE ORGANISATION**

**1.1 Introduction**

This internship report is about the skills attained, lessons learnt, and challenges, relatedness of theory covered in class and recommendations during my internship at TECHCITI from 24th January to 24th April 2019. The report also represents my experiences and benefits of the internship.

**1.2 Background of the Internship**

Internship is a field-based practical training experience that prepares trainees for the tasks they are expected to perform on completion of their programs. The main intention is to produce practically oriented graduates that meet the required job-related competences of their future employers

**1.3 Objectives of the internship**

* To enable students, get hands-on/real life experience they are expected to work in when they graduate.
* To provide an opportunity for students to apply the principles and techniques theoretically learnt into real-life problem solving situations.
* To provide an opportunity for students and academic staff to interact with the stakeholders and potential employers and thus appreciate field situations that will also generate information for curricula review and improvement.
* To develop student understanding of work ethics, employment demands, responsibilities and opportunities.
  1. **Background of the Organization**

**TechCiti** recognizes the importance of implementing the right solution for any business. It offers a wide range of services to build a solution that is right for your business needs.Every business, no matter the size, needs advice and support. They have a several years of technical experience and have accumulated a wealth of IT infrastructure knowledge. Their free consultation service helps you to establish your requirements. It will be with you every step of the way, from product selection through to configuration and installation.To stay relevant to our customers, it is important that they continuously demonstrate our ‘Value Creating Ability’ by:-

* " Helping our customers choose the right technology "
* " Bring competitive commercials to the discussion table "
* " Remaining predictable in our execution "
* " Backing it with excellent post-sales services "

Since established in 2013, TechCiti has become a pioneer in providing distinguished end-to-end IT infrastructure solutions to its customers through our business functions maximizing customer engagement with personalized services. They also believe that today more than ever, businesses are dependent on technology solutions.

They have successfully established in our business functions across 12 + major cities across PAN India with over 56 + satisfied corporate customers.

**1.5 Objectives and Services**

**TECHCITI** has been recognized as an industry leader in the development and implementation of websites, software solutions that consistently exceed client expectations.

**1.5.1. SERVERS AND STORAGE**

* Servers and storages includes SERVERS of RACK, TOWER & BLADE
* Storage area network is also one of feature
* Data attached storage
* Network attached storage

**1.5.2. IP (INTERNET PROTOCOL) INFRASTRUCTURE WIRED & WIRELESS**

* Routers, switches & POE solutions
* Fiber optics and Ethernet solutions
* Network security and solutions
* Wireless infrastructure solutions

**1.5.3. UNIFIED COMMUNICATIONS AND COLLABORATION**

* small meeting / board room solutions
* training room solutions
* digital signage & led display solutions

**1.5.4. LAPTOP, DESKTOPS & PERIPHERALS**

* LENOVO – LENOVO T SERIES, LENOVO W SERIES, LENOVO X SERIES, AIOS
* HP – HP PROBOOK, HP PAVILION, HP ELITEBOOK, AIOS
* DELL – DELL VOSTRO, DELL INSOPIRON, DELL LATTITUDE, DELL OPTIPLEX, AIOS
* APPLE MAC BOOK PRO, APPLE MACBOOK AIR, APPLE CCARE PROTECTION PLAN, APPLE ADAPTERS & CONVERTORS

**1.5.5. SOFTWARE BUSINESS APPLICATIONS**

* END POINT PROTECTION AND DLP
* APPLICATION SOFTWARES

**1.5.6. SERVICES**

* End to End IT Infrastructure Solutions

 This means that the provider of an application program, software and system will supply all the software as well as hardware requirements of the customer such that no other vendor is involved to meet the needs.

* Enterprise Software Solutions

Enterprise solutions are designed to integrate multiple facets of a company's business through the interchange of information from various business process areas and related databases.

* Web Application Development

Web application development is the creation of application programs that reside on remote servers and are delivered to the user's device over the Internet. ... Front-end development for web applications is accomplished through client-side programming. Client refers to a computer application such as a web browser.

**1.6 COMPANY SIZE and SHARE CAPITAL**

#### **TABLE 1.6.1 Share Capital & Number of Employees**

|  |  |
| --- | --- |
| Authorised Capital | ₹100,000 |
| Paid up capital | ₹100,000 |
| Number of Employees | 60+ |

#### **TABLE 1.6.2 Listing and Annual Compliance Details**

| Listing status | Unlisted |
| --- | --- |
| Date of Last Annual General Meeting | 29 September 2019 |
| Date of Latest Balance Sheet | 31 March 2019 |

**1.7 Organizational Culture**

**1.7.1 Values**

**Techciti** focuses on emerging corporates/institutions, small & medium scale companies to provide various technological solutions at their start-up stages helping our clients to save costs and succeed in their IT initiatives. Also, the company understands the huge requirement of post- sales services, which is promptly & adequate delivered to our clients. Our strong management team, robust and automated order processing systems, professional business development team supported by a dedicated execution team makes Techciti a Solution provider of Choice to corporates/institutions.

**1.7.2. Life at TechCiti**

"To be a pioneer in IT solutions and to become the most trusted service provider" the defining characteristic of our corporate culture. It is based on our core values, i.e. integrity, responsibility, passion, which provides the foundation for our success. Together with these core values and our effective business strategy, it also defines our business goals and specifies a clear direction for our development."

**1.8 Milestones**

05th April 2013 TechCiti was established.

02- July-2015- Took an edge over software technologies

15-April-2016-worked on creating top notch application softwares and became recognised as top Startups

**CHAPTER 2**

**ABOUT THE DEPARTMENT**

**2.1 TechCiti Department**

## **2.1.1. Good Visualizations for Machine Learning:** We are a highly visual species. Visuals speak to us better than numbers and charts. Data visualisation is an easy and concise manner to convey ideas in a universal manner. Having the knowledge of a scipy and scikit of tools will enable us to present our ideas in the best way possible and help us be the most effective.

**A. Supervised Learning**

This is a particular department where, **Supervised learning** algorithms are applied to real world problems and outcome is seen with a various visualisations. Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as [training data](https://en.wikipedia.org/wiki/Training_data), and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an [array](https://en.wikipedia.org/wiki/Array_data_structure) or vector, An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task.

**B. Unsupervised Learning**

**Unsupervised learning** algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms, therefore, learn from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. A central application of unsupervised learning is in the field of [density estimation](https://en.wikipedia.org/wiki/Density_estimation) in [statistics](https://en.wikipedia.org/wiki/Statistics), such as finding the [probability density function](https://en.wikipedia.org/wiki/Probability_density_function). Though unsupervised learning encompasses other domains involving summarizing and explaining data features.

**C. Semi – Supervised learning**

**Semi-supervised learning** falls between [unsupervised learning](https://en.wikipedia.org/wiki/Unsupervised_learning) (without any labelled training data) and [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) (with completely labelled training data). Some of the training examples are missing training labels, yet many machine-learning researchers have found that un-labelled data, when used in conjunction with a small amount of labelled data, can produce a considerable improvement in learning accuracy.

**2.1.2. Web development**

**Web development**is the work involved in developing a [website](https://en.wikipedia.org/wiki/Website) for the [Internet](https://en.wikipedia.org/wiki/Internet) ([World Wide Web](https://en.wikipedia.org/wiki/World_Wide_Web)) or an [intranet](https://en.wikipedia.org/wiki/Intranet) (a private network). Web development can range from developing a simple single [static page](https://en.wikipedia.org/wiki/Static_web_page) of [plain text](https://en.wikipedia.org/wiki/Plain_text) to complex web-based [internet applications](https://en.wikipedia.org/wiki/Internet_application) (web apps), [electronic businesses](https://en.wikipedia.org/wiki/Electronic_business), and [social network services](https://en.wikipedia.org/wiki/Social_network_service). A more comprehensive list of tasks to which web development commonly refers, may include [web engineering](https://en.wikipedia.org/wiki/Web_engineering), [web design](https://en.wikipedia.org/wiki/Web_design), [web content development](https://en.wikipedia.org/wiki/Web_content_development), client liaison, [client-side](https://en.wikipedia.org/wiki/Client-side_scripting)/[server-side](https://en.wikipedia.org/wiki/Server-side_scripting) [scripting](https://en.wikipedia.org/wiki/Computer_programming), [web server](https://en.wikipedia.org/wiki/Web_server) and [network security](https://en.wikipedia.org/wiki/Network_security) configuration, and [e-commerce](https://en.wikipedia.org/wiki/E-commerce) development.

**2.1.3. Python with Django team**

**Django** is a high-level **Python** Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel.

**2.1.4. Java with Spring team**

The **Spring** Framework is an application framework and inversion of control container for the Java platform. The framework's core features can be used by any Java application, but there are extensions for building web applications on top of the Java EE (Enterprise Edition) platform.

**2.1.5. Human Resource Department**

**Human resources** are the people who make up the workforce of an organisation, business sector, or economy. "Human Resource" is sometimes used synonymously with "human resources", although human capital typically refers to a narrower effect (i.e., the knowledge the individuals embody and economic growth). Likewise, other terms sometimes used include manpower, talent, labour, personnel, or simply people.

A human-resources department (HR department) of an organization performs human resource management, overseeing various aspects of employment, such as compliance with labour law and employment standards, administration of employee benefits and some aspects of recruitment.

A human resources manager has several functions in a company

* Determine needs of the staff.
* Determine to use temporary staff or hire employees to fill these needs.
* Recruit and train the best employees.
* Supervise the work.
* Manage employee relations, unions and collective bargaining.
* Prepare employee records and personal policies.
* Ensure high performance.
* Manage employee payroll, benefits and compensation.
* Ensure equal opportunities.
* Deal with discrimination.
* Deal with performance issues.
* Ensure that human resources practices conform to various regulations.
* Push the employees’ motivation.
* Mediate disputes internally.
* Upgrade learning knowledge of employees
* Disseminate information in the organization so as to benefit its growth

**2.1.6. Finance Department**

**Finance** is a field that is concerned with the allocation (investment) of assets and liabilities over space and time, often under conditions of risk or uncertainty. Finance can also be defined as the art of money management. Participants in the market aim to price assets based on their risk level, fundamental value, and their expected rate of return. Finance can be split into three sub-categories: public finance, corporate finance and personal finance.

Matters in personal finance revolve around:

* Protection against unforeseen personal events, as well as events in the wider economies
* Transference of family wealth across generations (bequests and inheritance)
* Effects of tax policies (tax subsidies or penalties) management of personal finances
* Effects of credit on individual financial standing
* Development of a savings plan or financing for large purchases (auto, education, home)
* Planning a secure financial future in an environment of economic instability
* Pursuing a checking and/or a savings account.

**CHAPTER 3**

**TASKS PERFORMED**

**3.1 RECORD OF CHANGES**

**TABLE 3.1 RECORDS OF CHANGES**

|  |  |  |
| --- | --- | --- |
| **Sl. No** | **Date** | **Task Performed** |
| **01** | **1– 02 –2019** | **Analyzing the plan** |
| **02** | **4–02–2019** | **Learning PYHTON and concepts from scratch** |
| **03** | **11–02–2019** | **Learning about Machine Learning** |
| **04** | **18–02–2019** | **Installation of python and ML software and executions** |
| **05** | **25–02–2019** | **ML in Supervised learning** |
| **06** | **5–03–2019** | **ML with Python** |
| **07** | **11–03–2019** | **Working both ML & python by considering few basic examples:** |
| **08** | **18–03–2019** | **Applying supervised algorithm called Regression and Classification algorithms on Cancer dataset.** |
| **09** | **16–04–2019** | **Visualization in ML with various plots** |

**3.2 ANALYZING THE PLAN**

Analysing the purpose of the internship and the requirements of the internship. Requirements includes, about the basics of python with respective software installation, and also working on a particular application of Machine Learning with 100% outcome with references.

**3.2.1. Learning PYHTON and concepts from scratch**

**Python** is a general purpose, dynamic, [high-level](https://www.javatpoint.com/classification-of-programming-languages), and interpreted programming language. It supports Object Oriented programming approach to develop applications. It is simple and easy to learn and provides lots of high-level data structures. Python is easy to learn yet powerful and versatile scripting language, which makes it attractive for Application Development. Python's syntax and dynamic typing with its interpreted nature make it an ideal language for scripting and rapid application development. Python supports multiple programming pattern*,* including object-oriented, imperative, and functional or procedural programming styles. Python is not intended to work in a particular area, such as web programming. That is why it is known as multipurpose programming language because it can be used with web, enterprise, 3D CAD, etc. Python makes the development and debugging fast because there is no compilation step included in Python development, and edit-test-debug cycle is very fast. Basics of python includes Python control statements, Python strings and its built in functions, Python function programs, Python programs with conditions and loops, Python Function programs, Python List, Python method, Python strings, Python tuples, Python dictionary, native data types and programs, Python set methods, Python fies I/O, Python modules, Python exceptions, Python date, Python OOPs, all functions of Python, Python programming examples.

**3.2.2. Learning about Machine Learning**

**Machine Learning** is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans:The ability to learn*.* Machine learning is actively being used today, perhaps in many more places than one would expect.

How ML works? A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately. Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output

Features of ML:

* Machine learning uses data to detect various patterns in a given dataset.
* It can learn from past data and improve automatically.
* It is a data-driven technology.
* Machine learning is much similar to data mining as it also deals with the huge amount of the data.

ML history, Ml applications in each domains was analysed, Life cycle of ML, Data processing was all learnt.

**3.2.3. Installation of python and ML software and executions**

* Download the **Python 3** Installer. Open a browser window and navigate to the Download page for Windows at python.org. ...
* Run the Installer. Once you have chosen and downloaded an installer, simply run it by double-clicking on the downloaded file.

Installing **PyCharm** **JET brains**

* Open the downloaded file. Before installing, there should be an option that says "Add Python 3.7 in PATH". Make sure to check this box. Then, continue installing normally.
* Open up the command prompt.
* Type py and press enter. You should see something that looks like this:

**Testing PyCharm**: PyCharm models a program as a 'project', which consists of one or more Python files, as well as any additional resources like images or text files.

All these helped to set up a path for execution.

**3.2.4. ML in Supervised learning**

Supervised learning is a machine learning method in which models are trained using labeled data. In supervised learning, models need to find the mapping function to map the input variable (X) with the output variable (Y).

Supervised Machine learning

Supervised learning needs supervision to train the model, which is similar to as a student learns things in the presence of a teacher. Supervised learning can be used for two types of problems: **Classification** and **Regression**

**Regression Types:**

* **Regression and analysis:**

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.

Types includes., **Linear Regression, Logistic Regression, Polynomial Regression, Support Vector Regression, Decision Tree Regression, Random Forest Regression, Ridge Regression,** **Lasso Regression.** Each regression was learnt by step by step procedure called algorithms.

* **Linear regression:**

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis.

* **Backward elimination**

Backward elimination is a feature selection technique while building a machine learning model. It is used to remove those features that do not have a significant effect on the dependent variable or prediction of output. There are various ways to build a model in Machine Learning, which are: **All-in Backward Elimination, Forward Selection, Bidirectional Elimination, Score Comparison.**

* **Polynomial regression:**

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.

**Classification Types:**

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, **Yes or No, 0 or 1, Spam or Not Spam, cat or dog,** etc. Classes can be called as targets/labels or categories.

* **Logistic regression:**

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.

* **KNN algorithm:**

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

* **Support Vector Machine:**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

* **Naïve Bayes classifier:**

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

**Other algorithms:**

* **Decision Tree Classifier:**

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

* **Random forest classifier**:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

**3.2.5.** **ML with Python**

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data. In this article, we’ll see basics of Machine Learning, and implementation of a simple machine learning algorithm using python.

Setting up the environment: Python community has developed many modules to help programmers implement machine learning. In this article, we will be using numpy, scipy and scikit-learn modules. We can install them using cmd command:

pip install numpy scipy scikit-learn

**3.2.6. Working both ML & python by considering few basic examples:**

Machine learning involves a computer to be trained using a given data set, and use this training to predict the properties of a given new data. For example, we can train a computer by feeding it 1000 images of cats and 1000 more images which are not of a cat, and tell each time to the computer whether a picture is cat or not. Then if we show the computer a new image, then from the above training, the computer should be able to tell whether this new image is a cat or not. The process of training and prediction involves the use of specialized algorithms. We feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. One such algorithm is [K-Nearest-Neighbor](https://www.geeksforgeeks.org/k-nearest-neighbours/) classification (KNN classification). It takes a test data, and finds k nearest data values to this data from test data set.

The steps include Training the dataset, Testing the dataset, Predict the accuracy. Similarly these particular example is worked with both regression and classification algorithms to obtain different accuracy and predict the outcome.

**3.2.7. Working on a project called Classification of Cancerous stages Profiles Using Machine Learning:**

Here the first and the foremost step involved is collection of dataset of patients with cancer. The breast cancer data was collected by the Netherlands Cancer Institute (NKI) and published on Nature 415 by van’t Veer et al. in 2002. It included clinical features but also gene expression levels; these represent how active genes and those that might contribute to cancer by being over-active or under-active. A cleaned version of this dataset was found on dataworld.com which was imported into a Pandas data-frame. It included expression levels of the 1554 most variable genes and 17 clinical features for 272 patients. A data dictionary was constructed by researching measures for breast cancer studies and by using the readme file associated with the original uncleaned dataset.

Data consists of Patient id, age, name, event-death, survival, time-recurrence, chemotherapy, hormonal therapy, amputation, hist-type, diameter of tumour, posnodes, grade, angionv, lymphinfil, barcode, gene id. Based on these parameters the patient was analysed the type of cancer called MALIGN or BENIGN that is represented as 1 or 0. Based on the accuracy obtained from each ML supervised algorithms the treatment of the therapy is involved that is amputation, chemo therapy or hormonal therapy.

**3.2.7. Applying supervised algorithm called Regression and Classification algorithms on Cancer dataset.**

* **Relationship between clinical variables and prognosis:**

Clinical features vs. event death The average patient was 44 years old and had grade 2 histopathological type 1 cancer with a tumour diameter of 23mm. 28% of the patients died; there was a positive correlation between the event of death and variables known to be associated with poor prognosis (tumour diameter, number of positive nodes, tumour grade), but a negative correlation between age and event death, which was unexpected given that cancer is an age-associated disease.

* **Time related features v/s Eventdeath:**

There were strong negative correlations between event death and a) the time between first treatment and follow-up and b) the time between first treatment and recurrence; the longer someone had survived or been disease free, the less likely they were to die. There is some argument to removing these variables for two reasons. Firstly because there is a lack of patients who – at the time of the most recent assessment – had survived breast cancer that had onset less than 5 years ago, causing any model to be heavily weighted to predict death for newly diagnosed patients. Secondly, the interval before recurrence would not be a relevant variable for these newly diagnosed patients with no prior history. These variables would therefore have to be removed and the model re-learned for use in a practical real-world scenario.

* **Categorical variables v/s Event-death:**

In general, less severe classes of angioinvasion and lymphocytic infiltration negatively correlated with event death whereas more severe classes positively correlated, but these relationships were weak and nearly always statistically insignificant. There was no correlation between each kind of treatment and outcome. The lack of strong correlations between clinical variables and the event of death are concurrent with the observation that breast cancer patients with the same stage of disease can have markedly different outcomes, highlighting the need to incorporate genetic or environmental variables for better prognosis.

* Therapy’s underwent by patients are: Chemotherapy, Hormonal therapy, Amputation These are the 3 treatments should be given to patients when based on malignant and benign type of cancers.
* Relations between clinical variables and each other based on the data.,

There were a number of interesting relationships between clinical variables; when comparing the three options of treatment, chemotherapy correlated strongly with the number of positive nodes, amputation correlated with tumour diameter and hormonal therapy correlated with age. It is logical to suggest that this might reflect the appropriateness of each treatment i.e., doctors choose to administer chemotherapy when the cancer is widespread but choose to amputate the breast when a tumour is large.

**3.2.8. Visualization in ML with various plots**

**i) Scatter plot:**

 one of the most commonly used plot for simple data visualization is scatter plots. This plot gives us a representation of where each points in the entire dataset are present with respect to any 2/3 features (Columns).

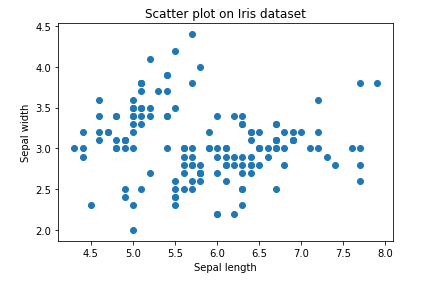


FIGURE 3.2.8.1.SCATTER PLOT

**ii) Pair plot**

Pair plot will create us a (n x n) figure where the diagonal plots will be histogram plot of the feature corresponding to that row and rest of the plots are the combination of feature from each row in y axis and feature from each column in x axis.

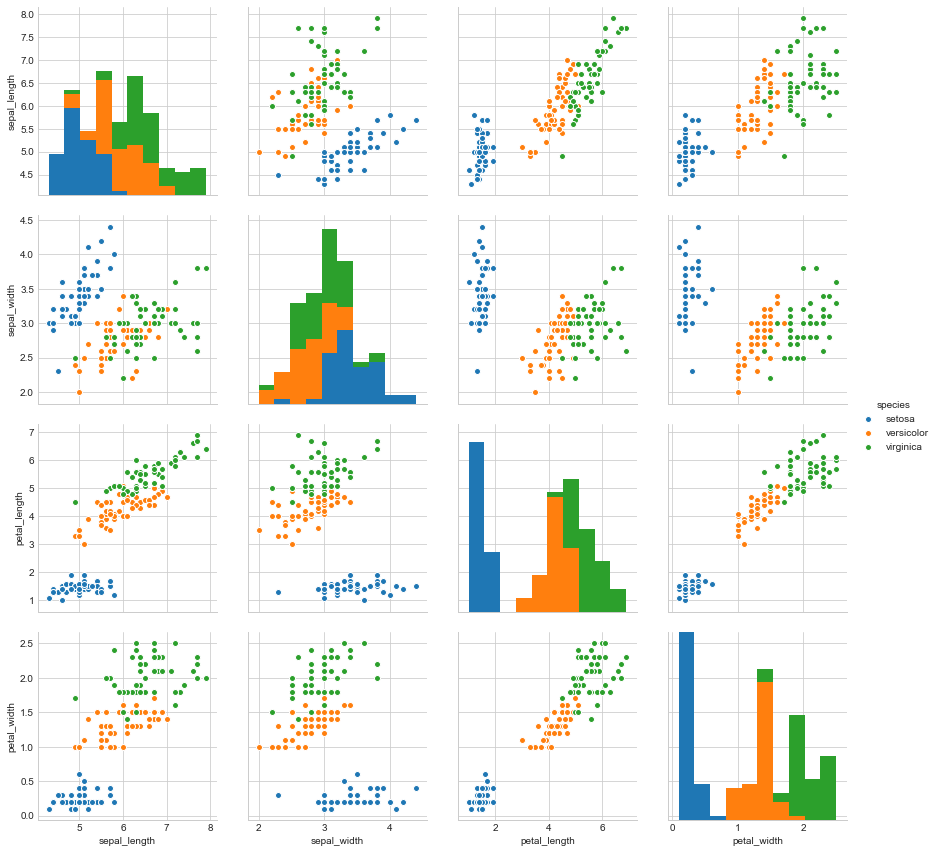


FIGURE 3.2.8.2 PAIR PLOT

**iii) Box plot**

This is the type of plot that can be used to obtain more of the statistical details about the data. The straight lines at the maximum and minimum are also called as **whiskers**. Points outside of whiskers will be inferred as an outliers. The box plot gives us a representation of 25th, 50th ,75th quartiles.

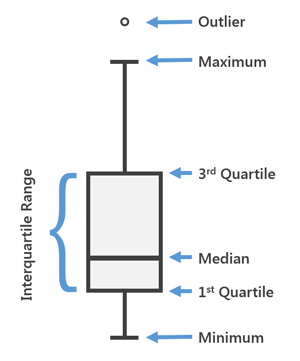


FIGURE 3.2.8.3 BOX PLOT

**iv) Violin plot**

The violin plots can be inferred as a combination of Box plot at the middle and distribution plots (Kernel Density Estimation ) on both side of the data. This can give us the details of distribution like whether the distribution is mutimodal, Skewness etc. It also give us the useful info like 95% confidence interval. The below image help us grasp some important parts from a violin plot.

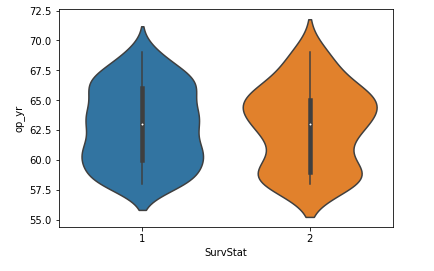


FIGURE 3.2.8.4 VIOLIN PLOT

**v) Distribution plot**

This is one of the best univariate plot to know about the distribution of data. When analyzing effect on dependent variable(output) with respective to a single feature(input), we use distribution plots a lot. It is also readily available in seaborn package.

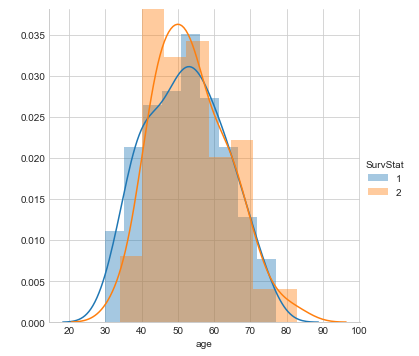


FIGURE 3.2.8.5 DISTRIBUTION PLOT

**vi) Bar chart**

This is one of the widely used plot, that we would have saw multiple times not just in data analysis, but wherever there is a trend analysis in many fields. Though it may seem simple it is powerful in analyzing data like sales figure every week, revenue from a product, Number of visitors to a site on each day of a week etc.

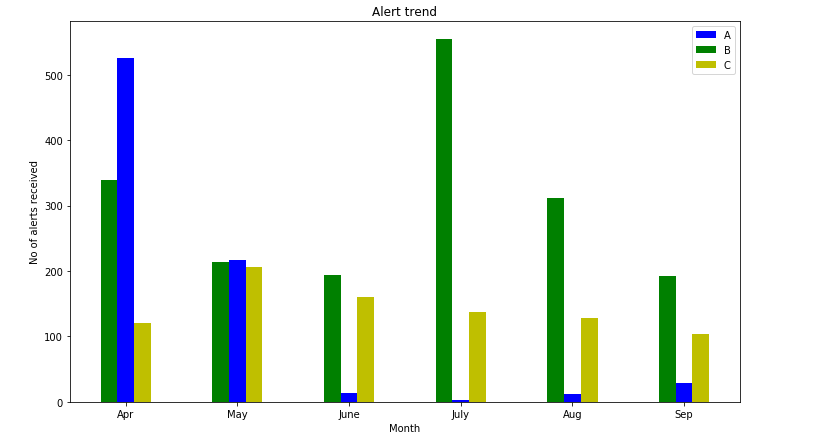


FIGURE 3.2.8.6 BAR CHART

**vii) Line plot**

This is the plot that you can see in nook and corners of any sort of analysis between 2 variables. The line plots are nothing but the values on series of datapoints will be connected with straight lines. The plot may seem very simple but it has more number of applications not only in machine learning but in many other areas.

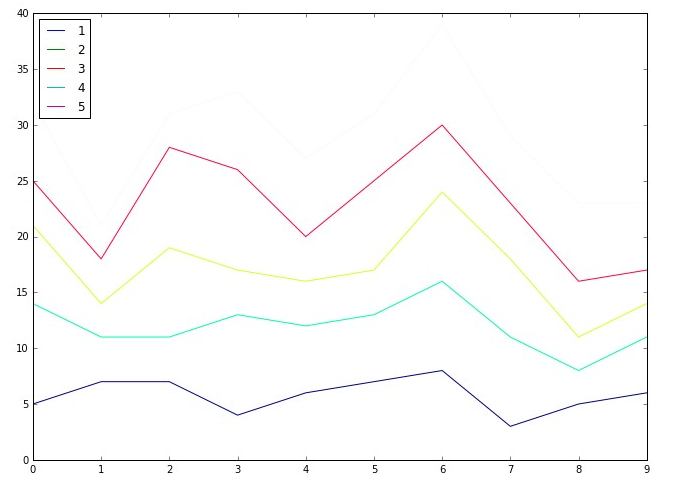


FIGURE 3.2.8.7 LINE PLOT

**CHAPTER 4**

**REFLECTION NOTES (SPECIFIC OUTCOMES)**

**4.1 Learning IT area of interest**

Internship offered me the opportunity to essentially customize my placement in order to gain valuable work experience in my areas of interest.

**4.2 Working with experienced and more skilled individuals**

Working with some of the best developers, designers and managers in information technology department, solving problems that you can never face in a college environment, bonding with other interns that make up a very diverse demographic.

**4.3 Benefits derived from internship**

The field attachment was of great importance, some of the benefits include

* Internship helped me understand work ethics, employment demands, responsibilities and opportunities.
* Field attachment provided career direction and confidence in my abilities by narrowing down the list of potential careers.
* My internship gave me the opportunity to try out computer related activities i.e. dismantling a computer which I had not previously considered and broadened my horizons through converting my academic knowledge into industry skills.
* It prepared me for the working environment.
* It enhanced my CV needed to negotiate future jobs
* Learnt Python, Machine Learning in Python by doing a project.

**4.4 Knowledge and skills gained**

**4.4.1. Machine Learning in Python**

This is one of the hot topic, because this field is used everywhere in all applications of day to day life. This particular domain was considered with a concept of Prediction of cancer by analyzing the data of cancer patient and also suggesting the therapy required to be implemented. This mainly involved collecting dataset from Wisconsin, and analyzing, applying supervised ML algorithms and predicting accuracy percentage, predicting either Malign or Benign, and suggesting therapy, visualization.

**4.4.2. Skills**

Computer competency, attention to detail, organization, problem solving, critical thinking, clear written and spoken communication, time management, close listening.

**4.4.3. Responsibilities**

Working towards achieving my individual goals, which helped the organization reach its objectives, Taking responsibility for your own professional and career development, Being open and accepting constructive feedback and take the initiative to improve, Giving others feedback, Completing any development plans assigned to me and applying the learning to improve my performance , Keeping record of my performance achievements, successes and challenges i.e. evaluation sheets in case of assignments , Completing my self-appraisal by the specified deadline.

During the internship period, I really enjoyed the experience of working at EBS including the comfortable working atmosphere, the technical guidance on the latest web technologies and the friendly relationship among the field staff.

**4.4.4. Outlines**

The project is related to Hierarchical employee system.

Main facilities available are:

* Data set collection
* Predicting cancer either Malign or Benign
* Suggesting therapies
* Applying all supervised algorithms
* Visualization

**4.5 SNAPSHOTS**

'radius\_1ean', 'texture\_1ean', 'peri1eter\_1ean', 'area\_1ean', 's1oothness\_1ean', 'co1pactness\_1ean', 'concavity\_1ean', 'concave points\_1ean', 'sy11etry\_1ean', 'fractal\_di1ension\_1ean', 'radius\_se', 'texture\_se', 'peri1eter\_se', 'area\_se', 's1oothness\_se', 'co1pactness\_se', 'concavity\_se' ,'concave points\_se', 'sy11etry\_se', 'fractal\_di1ension\_se', 'radius\_worst', 'texture\_worst', 'peri1eter\_worst', 'area\_worst', 's1oothness\_worst', 'co1pactness\_worst', 'concavity\_worst', 'concave points\_worst', 'sy11etry\_worst' ,'fractal\_di1ension\_worst'

FIGURE 4.5.1 DATASET

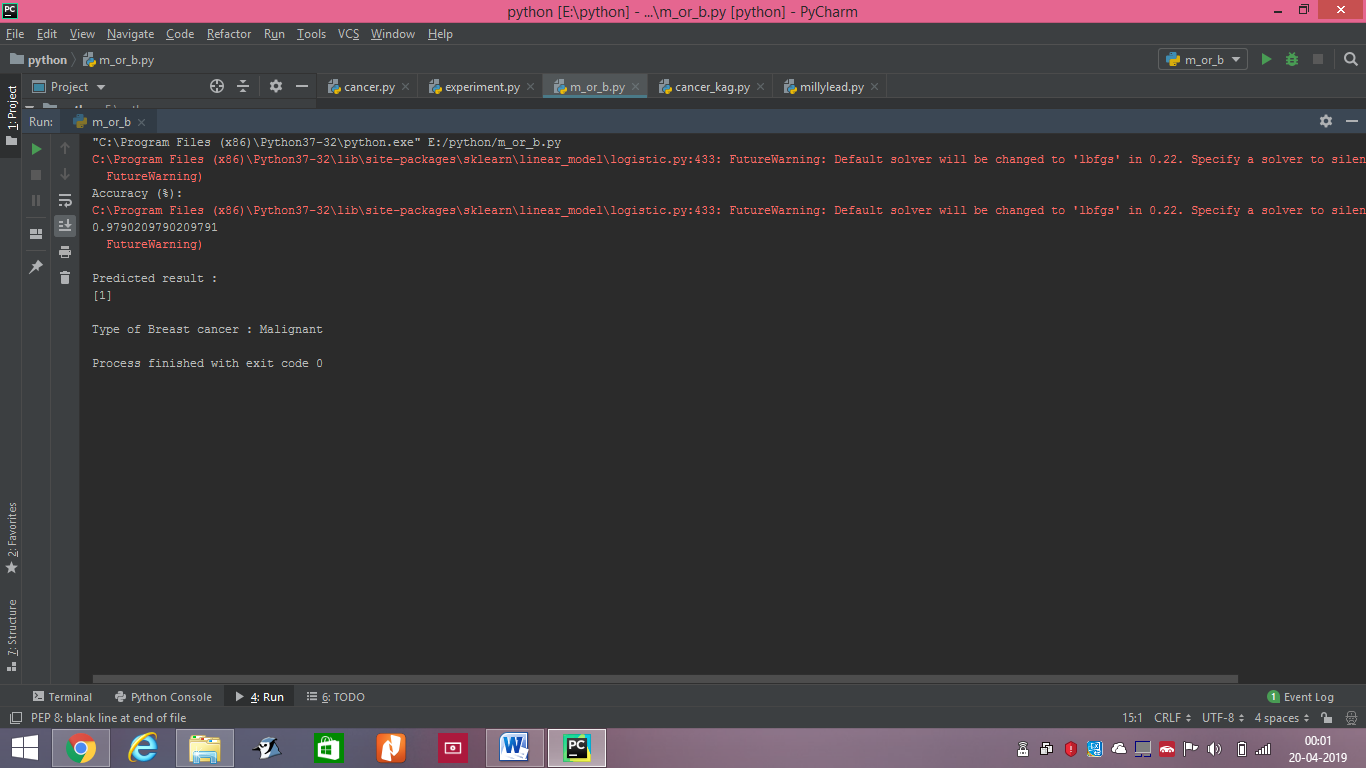


FIGURE 4.5.2 PREDICTING EITHER M / B

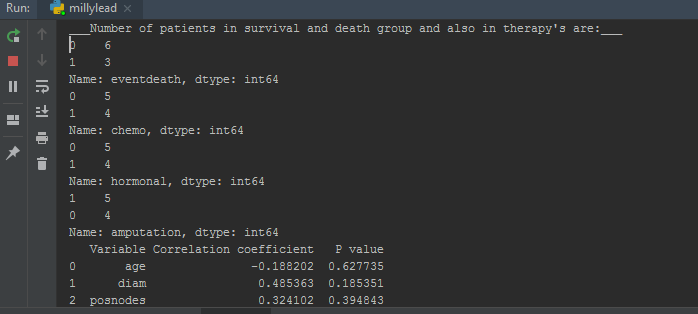


FIGURE 4.5.3 THERAPIES THROUGH MATRIX

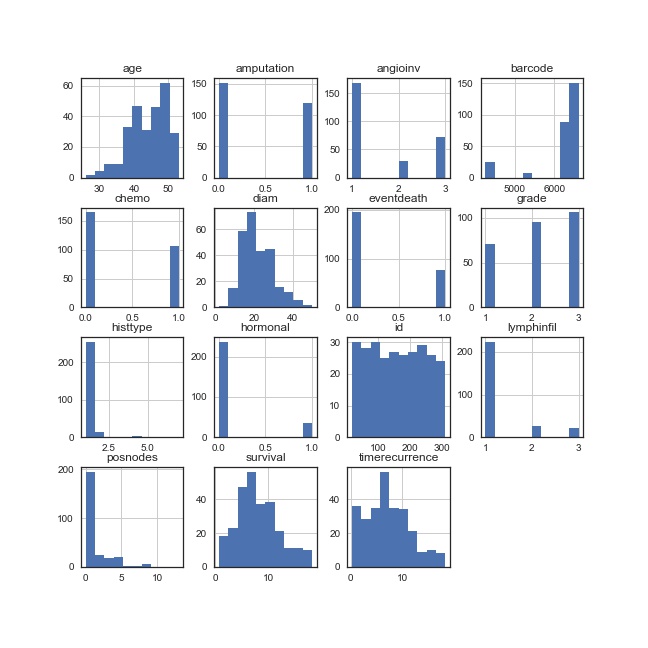
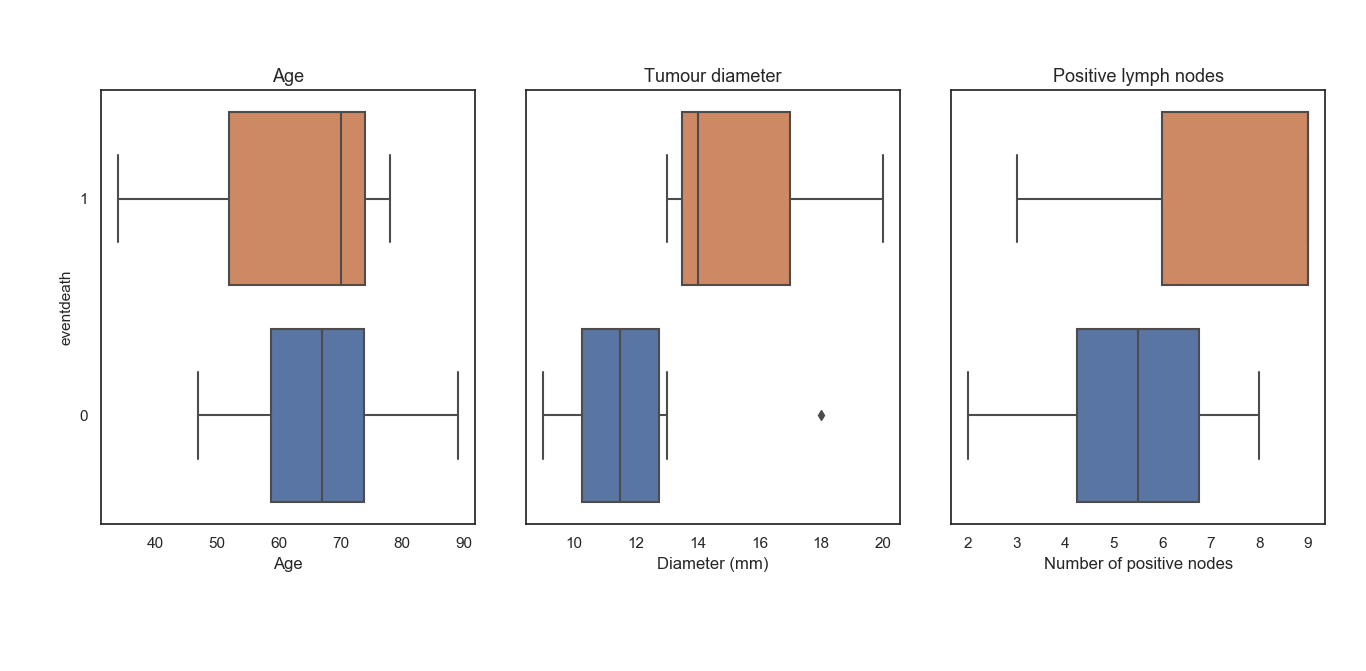


FIGURE 4.5.4 PLOT OF DATASET

 FIGURE 4.5.5 PLOT OF CLINICAL VARIABLES & PROGNOSIS

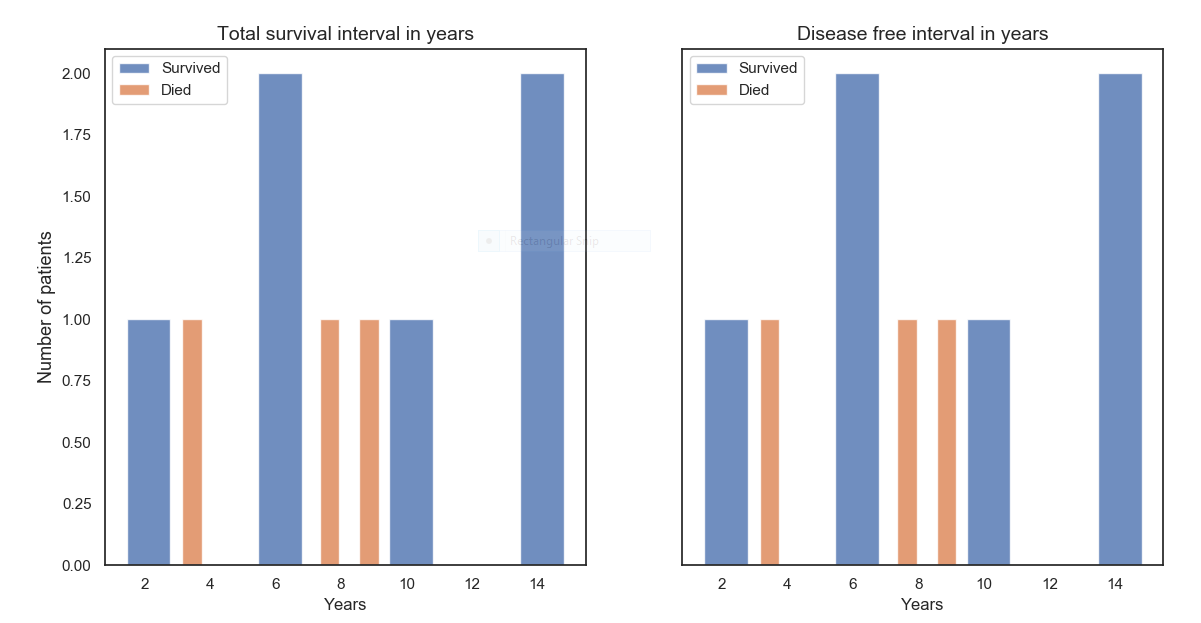


FIGURE 4.5.6 SURVIVAL PLOT

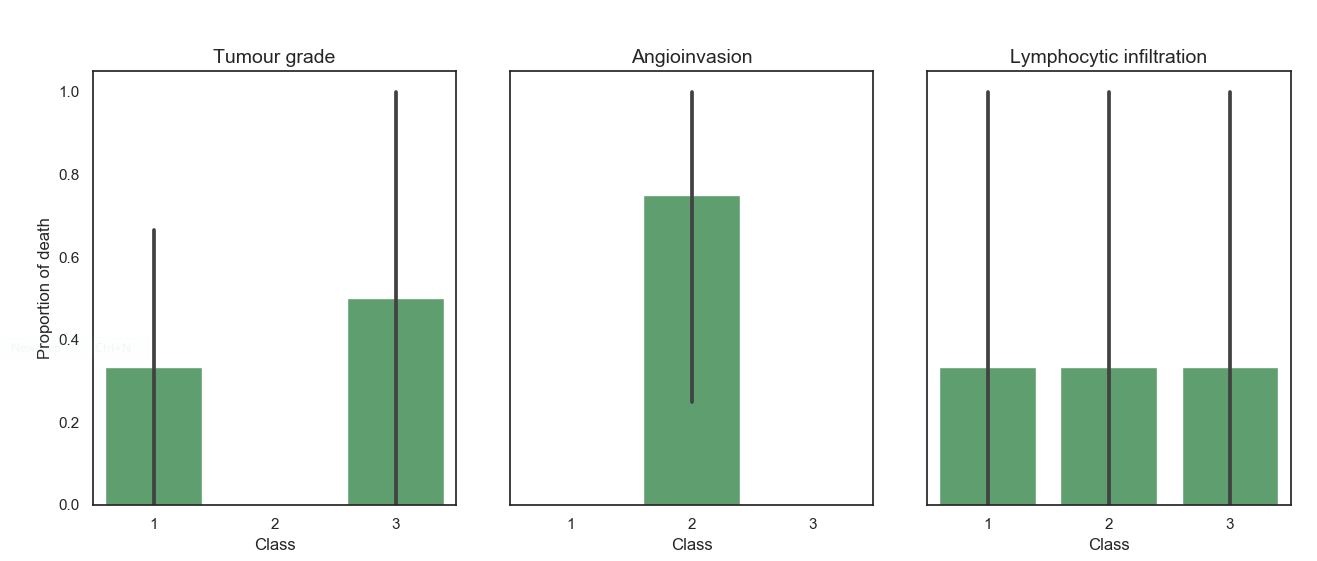
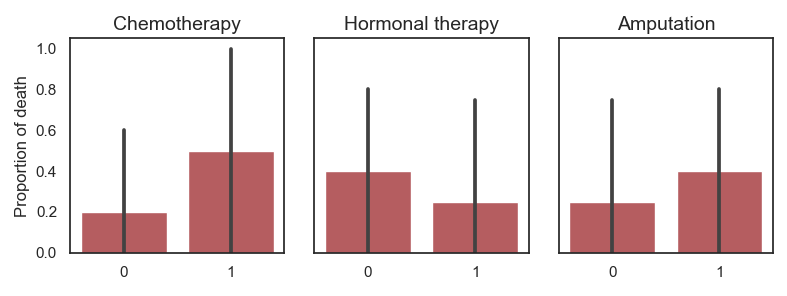


FIGURE 4.5.7 PLOT OF CLINICAL VARIABLES

 FIGURE 4.5.8 CLINICAL THERAPIES PLOT

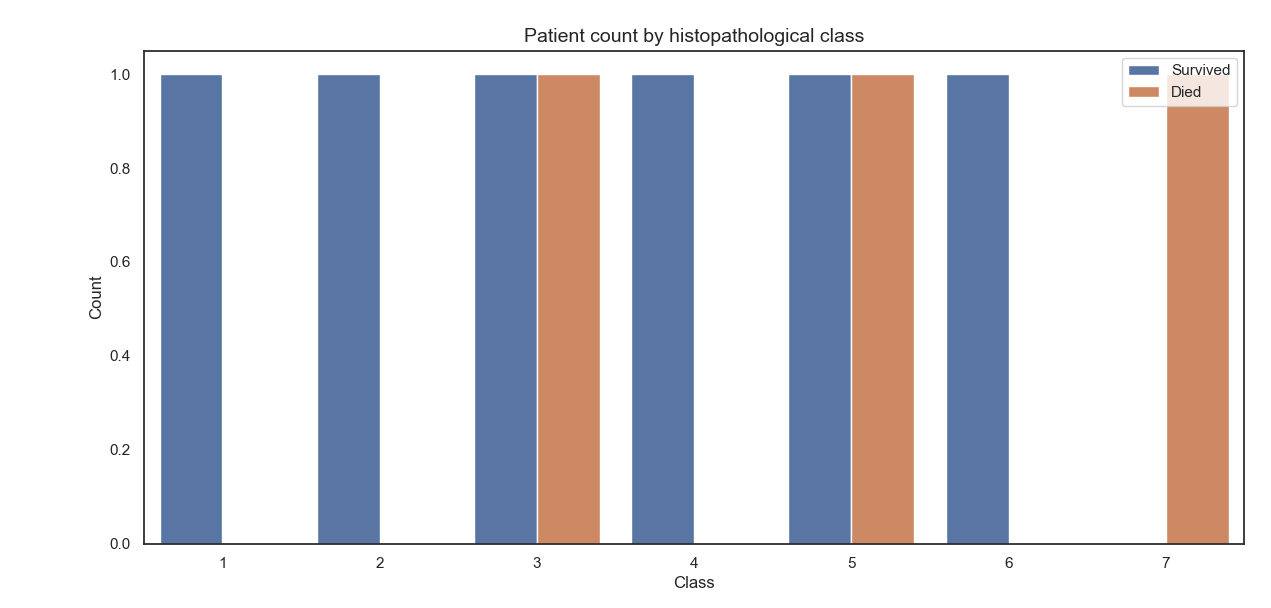
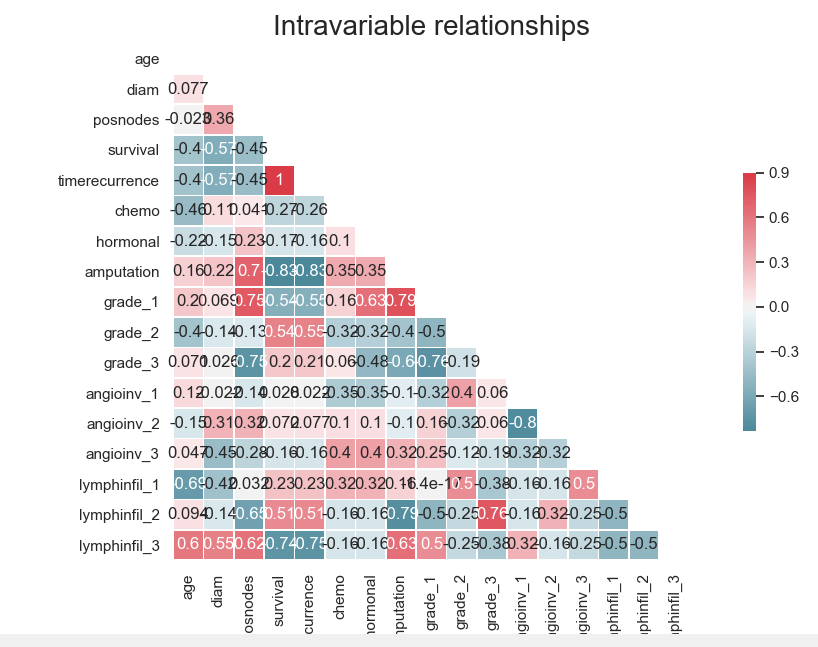


FIGURE 4.5.9 PLOT ON HISTOPATHOLOGICAL CLASS

 FIGURE 4.5.10 SEABORN PLOT ON INVARIANT RELATIONSHIP

**LIST OF REFERENCES**

1. [www.techciti.in](http://www.techciti.in)
2. <https://www.javatpoint.com/python-tutorial>
3. <https://www.javatpoint.com/machine-learning>
4. <https://www.geeksforgeeks.org/introduction-machine-learning-using-python/>
5. <https://www.google.com/search?q=various+plots+in+ML+using+matplotlib&oq=various+plots+in+ML+using+matplotlib&aqs=chrome..69i57.15766j0j7&sourceid=chrome&ie=UTF-8>
6. <https://towardsdatascience.com/data-visualization-for-machine-learning-and-data-science-a45178970be7>