TERM PAPER

On

‘Exploratory Data Analysis on Prostate Cancer’

Submitted to the Amity University Uttar Pradesh In partial fulfilment of requirements for the award of the Degree of

Bachelor of Technology in Bioinformatics

By

Keshav Mittal
Enrollment No: A0504222057

Under the Supervision of:

Internal Faculty Coordinator
(IFC)

Dr. Abhishek Sengupta

Amity Institute of Biotechnology
Amity University Uttar Pradesh
Sector 125, Noida – 201303 (India)

AMITY UNIVERSITY

UTTAR PRADESH

AMITY INSTITUTE OF BIOTECHNOLOGY

PLAGIARISM CERTIFICATE

This is to certify that the thesis entitled Exploratory Data Analysis on Prostate Cancer submitted by Keshav Mittal for the partial fulfilment of the degree of Bachelor of Technology in Bioinformatics has been checked by Turnitin software for plagiarism. The thesis has 12 percent plagiarism.

Signature of the IFC

Signature of the NTCC Coordinator

AMITY UNIVERSITY
UTTAR PRADESH

AMITY INSTITUTE OF BIOTECHNOLOGY

DECLARATION

I Keshav Mittal, student of BTech. Bioinformatics hereby declare that the TERM PAPER titled "Exploratory Data Analysis on Prostate Cancer using Python" which is submitted by me to Amity Institute of Biotechnology,Amity University Uttar Pradesh, Noida, for partial fulfilment for the award of the degree of BTech in Bioinformatics, has not been previously submitted for the award ofany degree, diploma or other similar title or recognition.

Noida

Date: 7th July 2023

Name and Signature of Student
Keshav Mittal

AMITY UNIVERSITY
UTTAR PRADESH

AMITY INSTITUTE OF BIOTECHNOLOGY

CERTIFICATE

On the basis of declaration submitted by Keshav Mittal, student of BTech. Bioinformatics, I hereby certify that the TERM PAPER titled "Exploratory Data Analysis on Prostate Cancer using Python" which is submitted to AmityInstitute of Biotechnology, Amity University Uttar Pradesh, Noida, for partial fulfilment for the award of the degree of BTech in Bioinformatics, is a faithful record ofwork carried out by him/her under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree Diploma to this University or elsewhere.

Noida

Date: 7th July, 2023

(Internal Faculty Coordinator)
Name and Signature

AMITY UNIVERSITY

UTTAR PRADESH

AMITY INSTITUTE OF BIOTECHNOLOGY

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have contributed to the completion of this term paper. First and foremost, I express my deepest appreciation to my supervisor Dr. Abhishek Sengupta, Assistant Professor at Amity Institute of Biotechnology, who was my Internal Faculty Coordinator (IFC) and provided constant support throughout the duration of this report. Without his guidance, I would have not been able to complete the report in a timely manner. Their expertise, insightful feedback and constant motivation has played a significant role in refining my ideas and enhancing the quality of this report.

I would further like to thank the NTCC Committee for giving me such an interesting and dynamic topic and for providing me with the opportunity to increasemy knowledge and skills.

Lastly, I would like to thank my parents, my friends and my peers for their help and guidance and for providing me with the inspiration and confidence to finishthis report. Their support and belief in my abilities have been a constant source of motivation and strength to persevere through challenges & strive for excellence.

Table of Contents

List of Figures and Tables

Abstract

One of the greatest widespread cancers harming men around the globe is still prostate cancer. Patients benefit from early cancer detection during the course of treatment. This study aims to conduct Exploratory Data Analysis (EDA) on prostate cancer manipulating a diverse dataset to gain insights to its various aspects. The dataset comprises data that was gathered over several years from a sizable cohort of patients who were eventually diagnosed with prostate cancer. We evaluated the distribution and interrelationships of numerous variables, including radius, smoothness, symmetry, texture and other parameters through methodical data research and visualization tools. A thorough overview of the dataset is also provided through the EDA's visualization. It allows one to identify trends and outliers by using bar charts, histograms, and line graphs to show the distribution of categorical and continuous variables. Additionally, scatter plots are used to show the interdependencies and correlations between various variables, assisting in the discovery of potential risk factors and prognostic markers.

The effectiveness of various supervised machine learning algorithms (such as Naïve Bayes) for predicting prostate cancer is compared and discussed in this paper. The main goal is to evaluate the efficacy and capability of each approach in the fields of various aspects like accuracy to ascertain whether the data classification was accurate. The training and test data may have an impact on the methods' accuracy. The findings of this study advance knowledge of prostate cancer and could help clinicians choose the best course of treatment and management for their patients. In the end, the information discovered through this exploratory data analysis may help in improving patient outcomes and prostate cancer detection and in its treatment.

1. Introduction

Cancerous cells are aberrant cells that multiply more quickly than usual and don't want to die. Cancer can affect many different organs of our body, including prostate. Benign and malignant cancer origin are the two sorts. In benign growth, only the tissue expands; in malignant growth, if an organ is not subjected to an initial diagnosis-treatment approach, it becomes nonfunctional. Each organ is affected by cancer in a distinct way [1]. When these symptoms are taken into account, it is feasible to eliminate the infected organ, stop its spread, and save the person's life using a variety of treatment procedures. Because of this, early cancer diagnosis has become one of the most significant medical issues. Certain prostate cancer sufferers have a higher propensity to manifest a fatal trait than others. With the therapy of the primary tumor, there may be a healing scenario in some patients. During the course of the disease, accurate patient follow-up is the most crucial component. To create more efficient treatment paradigms, it is necessary to identify the patients who are at the greatest danger. For diagnosis, classification must be completed.

This study focused on prostate cancer, which causes one of the most cancer-related deaths in males and has symptoms that are comparable to those of benign development. Disease diagnosis is one of medicine's major obstacles to overcome. Machine learning techniques may be advantageous in solving issues where there are no established regulations and it is possible to forecast the factors influencing an event [2]. In order to predict prostate cancer, this study compares and uses a variety of supervised machine learning algorithms, including Support Vector Machines (SVM), k-Nearest Neighbor (KNN), Logistic Regression (LR), Naive Bayes, and Decision Tree Method Name. The algorithm used in this system will learn the correlations between variables and be able to identify prostate cancer. With the success of this study, a patient might steer clear of an unnecessary biopsy.

2. Analogous Findings

The important and basic tasks in machine learning and data mining is classification. Multiple research efforts have been conducted to classify prostate cancer using diverse medical data sets employing data mining and machine learning. On the Kaggle data set, Sri Venkatesh (2020) predicted prostate cancer using machine learning methods. Here's a short summary of several linked articles and research projects on prostate cancer that used exploratory data analysis (EDA); we were able to get a sense from these studies of the kinds of analyses and understandings that EDA in the context of prostate cancer research can provide:

1. Zhang et al.'s "Exploratory Data Analysis of Prostate Cancer Data" (2019): In order to pinpoint the variables linked to prostate cancer diagnosis, this work focuses on EDA of a sizable dataset on the disease. To investigate the connections between different patient characteristics and the existence of prostate cancer, it uses descriptive statistics, visualization techniques, and correlation analysis [3].
2. Reardon et al. (2017)'s "Exploratory Data Analysis of Prostate Cancer Mortality Rates": The mortality rates for prostate cancer in various US regions are examined by the authors using EDA. They evaluate spatial patterns and find probable risk factors linked to increased mortality rates utilizing maps, scatter plots, and other visualization approaches [4].
3. Yang et al. (2018), "Exploratory Analysis of Prostate Cancer Data Using Visualization Techniques" This study investigates a dataset on prostate cancer using EDA approaches like box plots, heatmaps, and correlation matrices. In order to provide insights for additional analysis and modelling, the authors seek to identify relevant variables and potential correlations among them [5].
4. By Rafiq and Khan (2020), "Exploratory Data Analysis of Prostate Cancer Using Machine Learning and Visualization": Data on prostate cancer is reviewed in this study using machine learning algorithms and EDA. To determine significant features and forecast prostate cancer outcomes, the authors investigate feature selection, clustering, and classification algorithms [6].

5. According to Gha semi et al. (2019), "A Comprehensive Exploratory Data Analysis of Prostate Cancer Data": The study performs a thorough EDA on data related to prostate cancer, concentrating on examining demographic, clinical, and pathological aspects connected to prostate cancer. To comprehend the connections between these elements and the development of the disease, the authors use a variety of statistical methods and visualizations [7].

It became apparent that performance might be enhanced by establishing variable selection algorithms, depending on the correlation between the variables.

3. Dataset and Methodology

Figure 1 shows the study's design flow diagram, which was used to predict prostate cancer using a variety of supervised machine learning approaches. The subsections below this diagram provide extensive details on each phase.

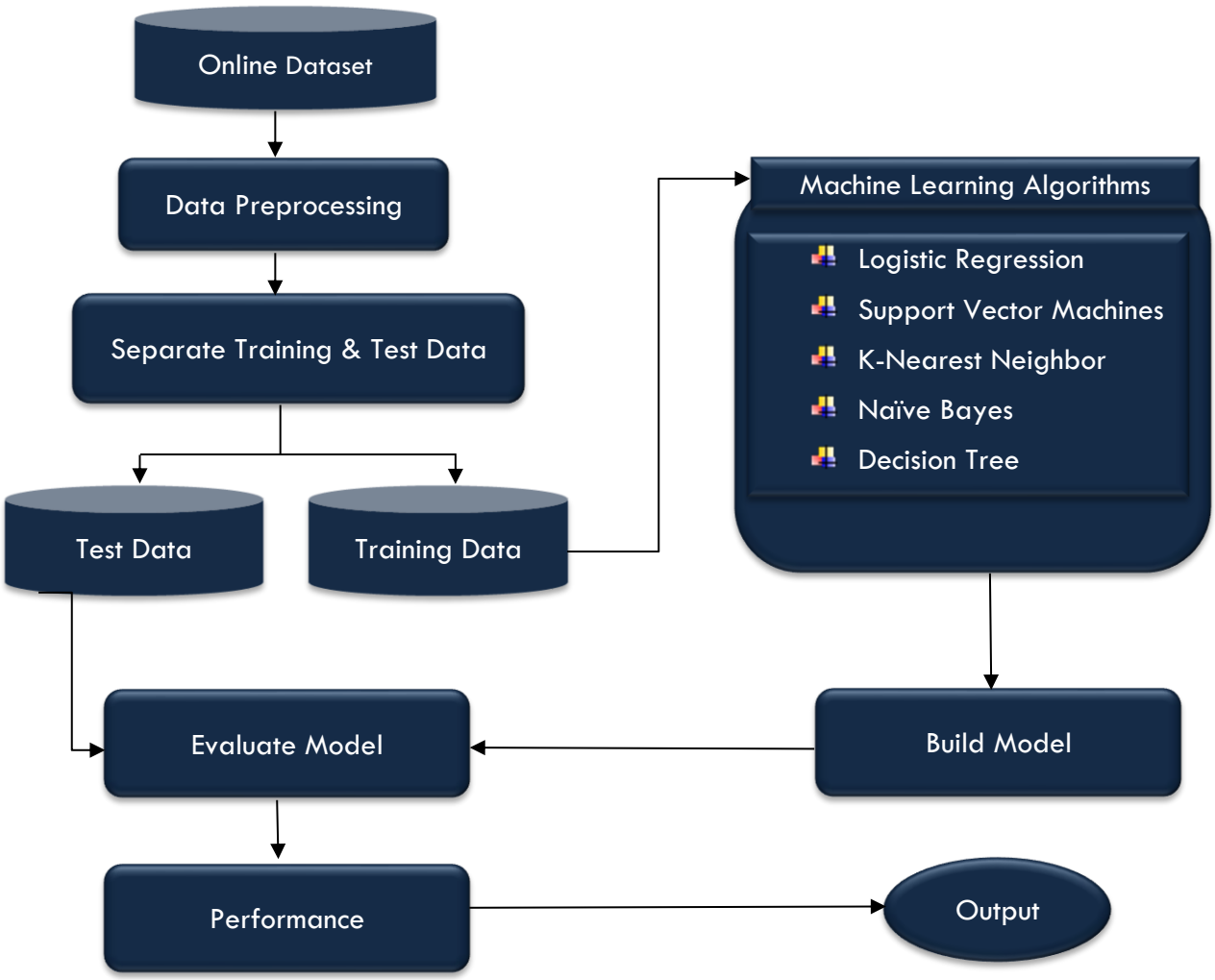


Figure 1. Workflow of supervised machine learning algorithms performance analysis

3.1. Dataset

As implied in the first phase of the design flow, a dataset on prostate cancer was utilized to analyze the effectiveness of the strategies tested in this study. The Kaggle platform allows access to the Prostate Cancer dataset (Sajid, 2018). 100 patient observations make up the dataset. The dataset consists of one dependent variable and eight independent variables. The variables used as predictors are as follows: 1-Radius, 2-Perimeter, 3-Texture, 4-Smoothness, 5- Area, 6- Compactness, 7-Symmetry, 8-Fractal dimension and 9-Diagnosis result (dependent) [8].

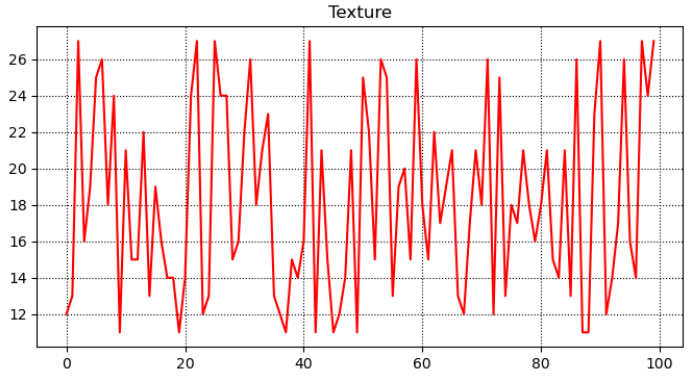
Table 1. Exploration of the dataset using statistics

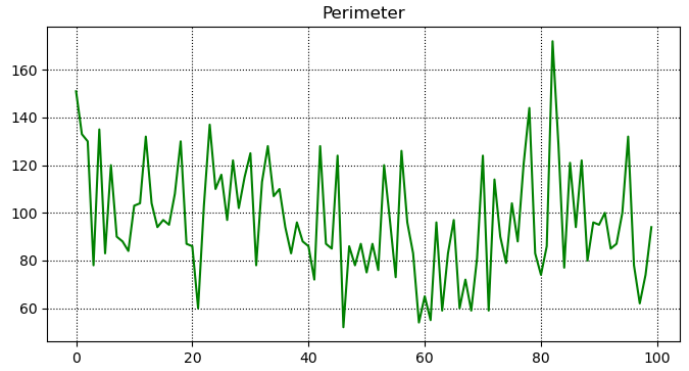
	Radius	Texture	Perimeter	Area	Smoothness	Compactness	Symmetry	Fractal Dimension
Count	100	100	100	100	100	100	100	100
Mean	16.85	18.23	96.78	702.88	0.102730	0.126700	0.193170	0.064690
Std	4.87	5.19	23.67	319.71	0.014642	0.061144	0.030785	0.008151
Min	9	11	52	202	0.07	0.038	0.135	0.053
Max	25	27	172	1878	0.143	0.345	0.304	0.097

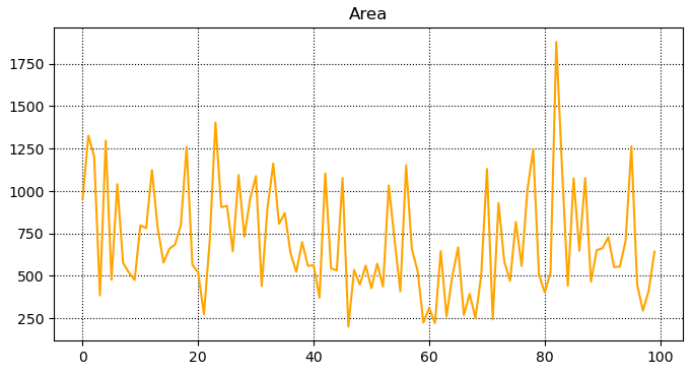
There are two possible values for the output response: "B" for benign tumors and "M" for malignant tumors. Table 1 presents a statistical analysis with comprehensive statistics on the data.

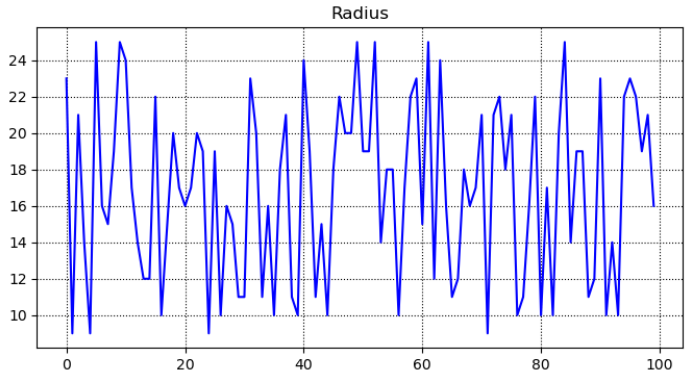
3.2. Preprocessing

A few attributes with high value and some with lesser value can be found together in the data set, as shown in Table 1. The data must be transformed in this situation since features with lower values will not have any impact compared to those with higher values. Since it will exceed the low-value features in weight calculations, the data (as shown in Figure 2), which is extremely varied in terms of size, will increase the variance and have a different impact on distance estimate


Texture
26
24
22
20
18
16
14
12
0
20
40
60
80
100


Perimeter
160
140
120
100
80
60
0
20
40
60
80
100


Area
1750
1500
1250
1000
750
500
250
0
20
40
60
80
100


Radius
24
22
20
18
16
14
12
10
0
20
40
60
80
100

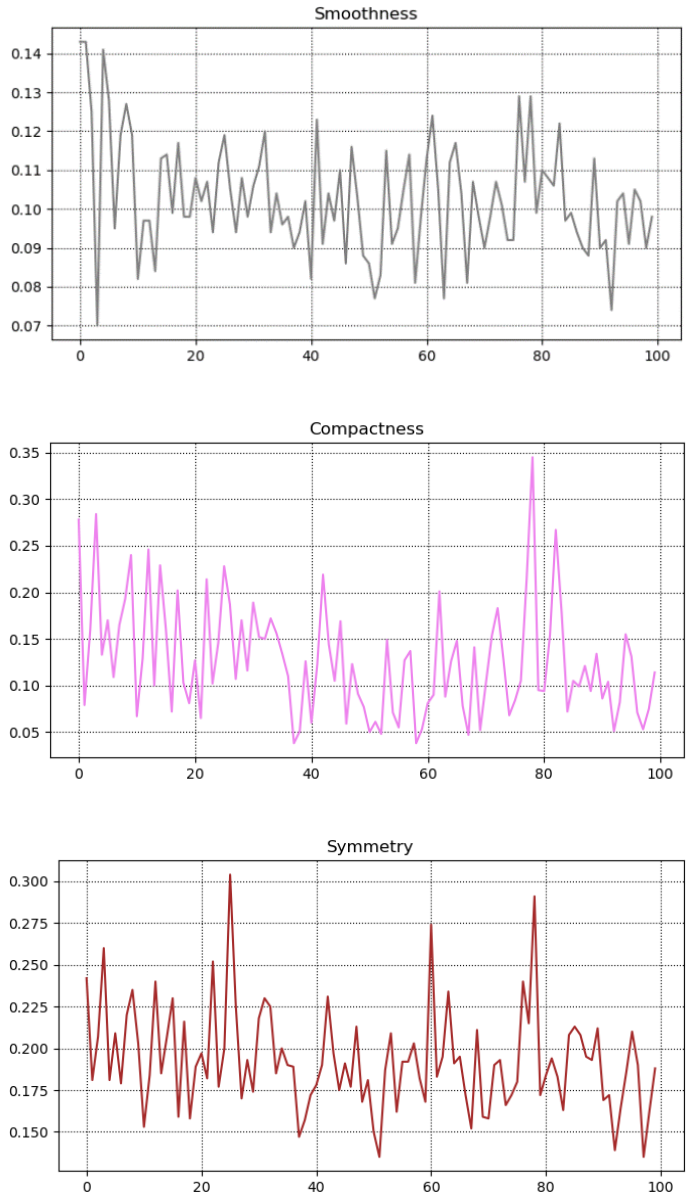


Figure 2. The medical traits and overall data flow

To reconcile the data, during the prep phase, the figures of the columns containing numbers in the data set must be changed to a common scale without influencing the variances in the range of values. Therefore, a scaling method is typically used to close the gaps between different dataset values. Similarly, it is not recommended to process non-numeric attributes while doing data transformations. Non-numerical data must be transformed into numerical data before data mining can be performed. The category field in the Diagnosis_Result column, which symbolizes the

dataset's dependent variable, has been changed to binary. To ensure that each variable equally contributes to the effectiveness of the algorithms employed in this study, these values must be scaled to the same standard. The dataset's data was scaled via feature standardization to assure standardization. Python's Standard Scaler method from the sklearn.preprocessing module was utilized for this.

3.3. Machine Learning Algorithms

In reference to this research project, a few well-known supervised machine learning methods were assessed to examine how well they performed on a dataset of prostate cancer cases. In order to do this, we used Naive Bayes (NB), Logistic Regression, K-Nearest Neighbor (K-NN), Support Vector Machines (SVM), and Decision Tree Method Name. These approaches are assessed in order to compare how well various machine learning strategies perform on the same dataset. These algorithms are favored because they are simple to use and capable of producing effective performance outcomes.

3.3.1. Logistic Regression

Finding the correlation between variables is done using regression analysis as an analysis technique. Regression analysis reveals the functional form of the relationship between the variables that are dependent and independent so that predictions can be made. Similar to other techniques for building statistical models, logistic regression analysis' foremost objective is to create a model. The objective is to construct a physiologically reasonable framework that can, with the fewest number of inputs, articulate an association between dependent and independent factors with the best fit (Yavuz & Ilengirolu, 2020).

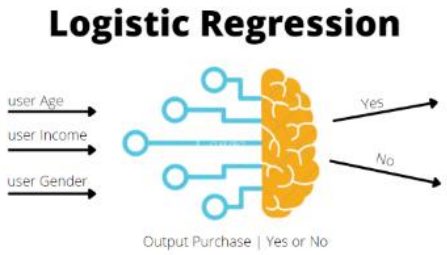


Figure 3. [9]

To make it easier to evaluate model parameters such fluctuations in log size, logistic regression is frequently employed in medical applications (Kurt et al., 2008). It is acceptable to interpret results as probabilities and to use variable selection techniques frequently used in commercial applications.

3.3.2. K-Nearest Neighbor (KNN)

The above algorithm is a grouping technique in which the group containing the sample data point and its nearest neighbor is selected centered on the value of k (Cover & Hart, 1967). An approach for supervised learning that resolves the grouping issue is K-NN. It is one of the classification methods that is used the most in the literature. Data whose class is known are used to classify data in the K-NN algorithm [10].

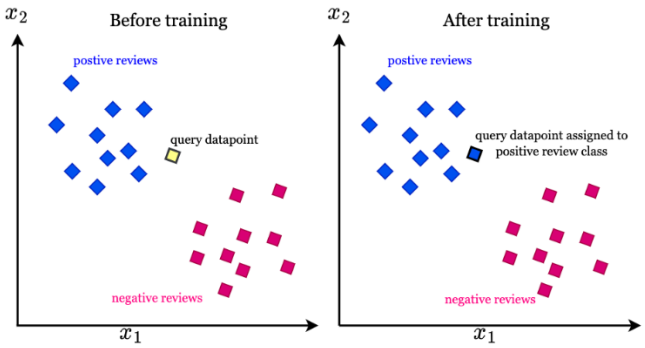


Figure 4. [11]

In spite of its basic nature, K-NN consistently outperforms more complex learning algorithms, sometimes even outperforming them. For lesser numbers

of classes, this strategy outperforms other machine learning approaches in terms of simplicity and effectiveness (Karakoyun & Hacbeyolu, 2014). The main benefit of this approach is that it enables successful research to be conducted on the categorization of numerous categorized data points. K-NN may be preferred for both classifying data and solving regression issues. It is employed when the classification procedure is applied based on the separations between observations and the independent variables are quantitative. Despite having a rather straightforward structure, it requires a lot of processing. The neighborhood value (k) that best fits the relevant dataset is initially determined in this study. For most data sets, the k value is set to be between 3 and 10. In order to do this, training data are used, and the best neighborhood value for the relevant dataset is found to be 5.

3.3.3. Decision Tree Method

Decision trees are a catching on machine learning method for applications involving classification and regression. They build a model that resembles an impact tree of decisions using labelled training data. The algorithm begins by choosing the most advantageous feature to divide the data according to a standard like information gain or Gini index. The values of the selected feature are then used to divide the data into subgroups. Recursively splitting each subset results in the formation of a binary tree structure. A stopping requirement, such as reaching a maximum tree depth or having every instance in a subset belong to the same class, must be satisfied before the splitting can end. Based on the dominant class or average value of the examples in that node, class labels or regression values are assigned to the leaf nodes of the tree. Pruning is an optional step that can be taken to prevent overfitting and remove nodes or branches that do not significantly improve prediction performance. In order to make predictions about future occurrences, we follow the decisions made by the tree from its root to its leaf node, and then we output the appropriate class label or regression value.

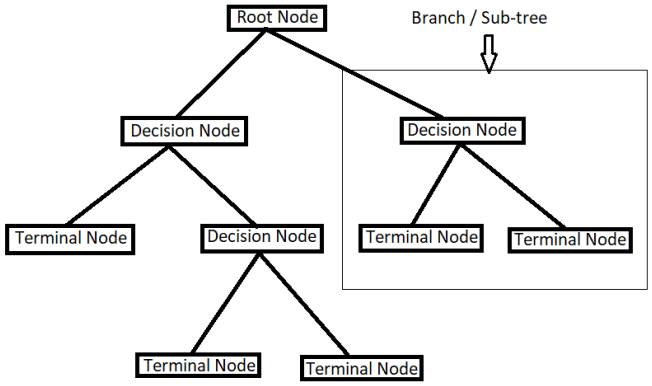


Figure 5. [12]

Decision trees are comprehensible, capable of handling both category and numerical data, and robust to outliers. However, they could overfit the training set and struggle with extrapolating to other situations. Pruning, ensemble methods (such as random forests), and boosting techniques (such as AdaBoost) can all be used to enhance decision tree performance and deal with these drawbacks.

3.3.4 Naïve Bayes

A fundamental statistical classifier is the Naive Bayesian algorithm. This classifier determines a probability set for a given dataset by counting combinations of values and frequency. In accordance to this algorithm, each variable is independent of the others. Furthermore, no interdependence between qualities is assumed. This conditional independence hypothesis is occasionally true in practical applications, making it naive. Despite this, the algorithm learns swiftly in a variety of controlled classification problems. It may be argued that this is generally superior than alternative techniques in various situations.

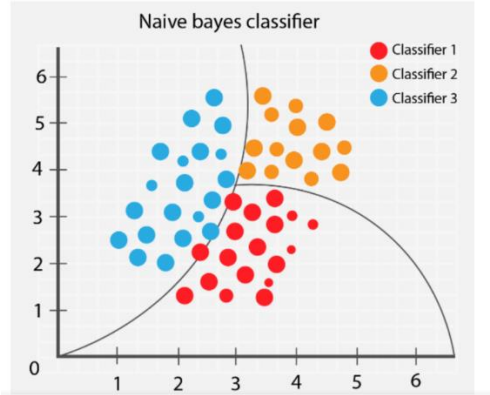


Figure 6. [13]

3.3.5 Support Vector Machines (SVM)

This is a well-known encoder that makes use of quadratic programming as its basis and has the ability to offer a high level of model complexity solution. Using a suitable kernel function, data is initially translated into the feature space in SVM. The hyper plane is used to classify things in that area. The foundation of SVMs is statistical learning theory. SVM can also be assessed in the category of feed-forward networks.

4. Research Findings and Analysis

In this part, the K-Nearest Neighbor (K-NN), Support Vector Machines (SVM), Decision Tree Method, Naive Bayes, and Logistic Regression classifiers are five supervised machine learning approaches that we compare and elaborate. To evaluate the effectiveness and accuracy of the classifier, the sets that were used for training and testing were arbitrarily split into 70 per cent training and 30 per cent test data based on the original data.

Depending on how the distribution of the test & training information is set up, different performance results can be obtained using various methodologies. We ran each algorithm more than 10 times at this point in order to get more consistent results, and we logged the top five outcomes for each algorithm. The classification led to the identification of "B" benign cells and "M" malignant cells (cancer).

Table 2. Different Results

Algorithm Name	Precision	Recall	Accuracy	F-1 Score
Logistic Reg.	0.83	0.9375	0.8	0.88
KNN	0.8	1.0	0.8	0.88
Decision Tree	0.8	0.75	0.65	0.77
Naïve Bayes	0.875	0.875	0.8	0.875
SVM	0.83	0.9375	0.8	0.88

For a dataset with two (binary) classes, we used a confusion matrix to calculate the Precision, Recall, Accuracy, and F-Measure of various machine learning techniques. Additionally, the F-measure combines recall and precision into a single metric. It occasionally might be more helpful than precision, particularly in classes with an unequal distribution of students.

We observed how the five classifiers performed in comparison, as given in Table 2. One of the most straightforward and heuristic metrics for measuring model accuracy is the confusion matrix. Figure 7. displays the machine learning techniques that were observed test complexity matrix. Values for TN, TP, FN, and FP are considered.

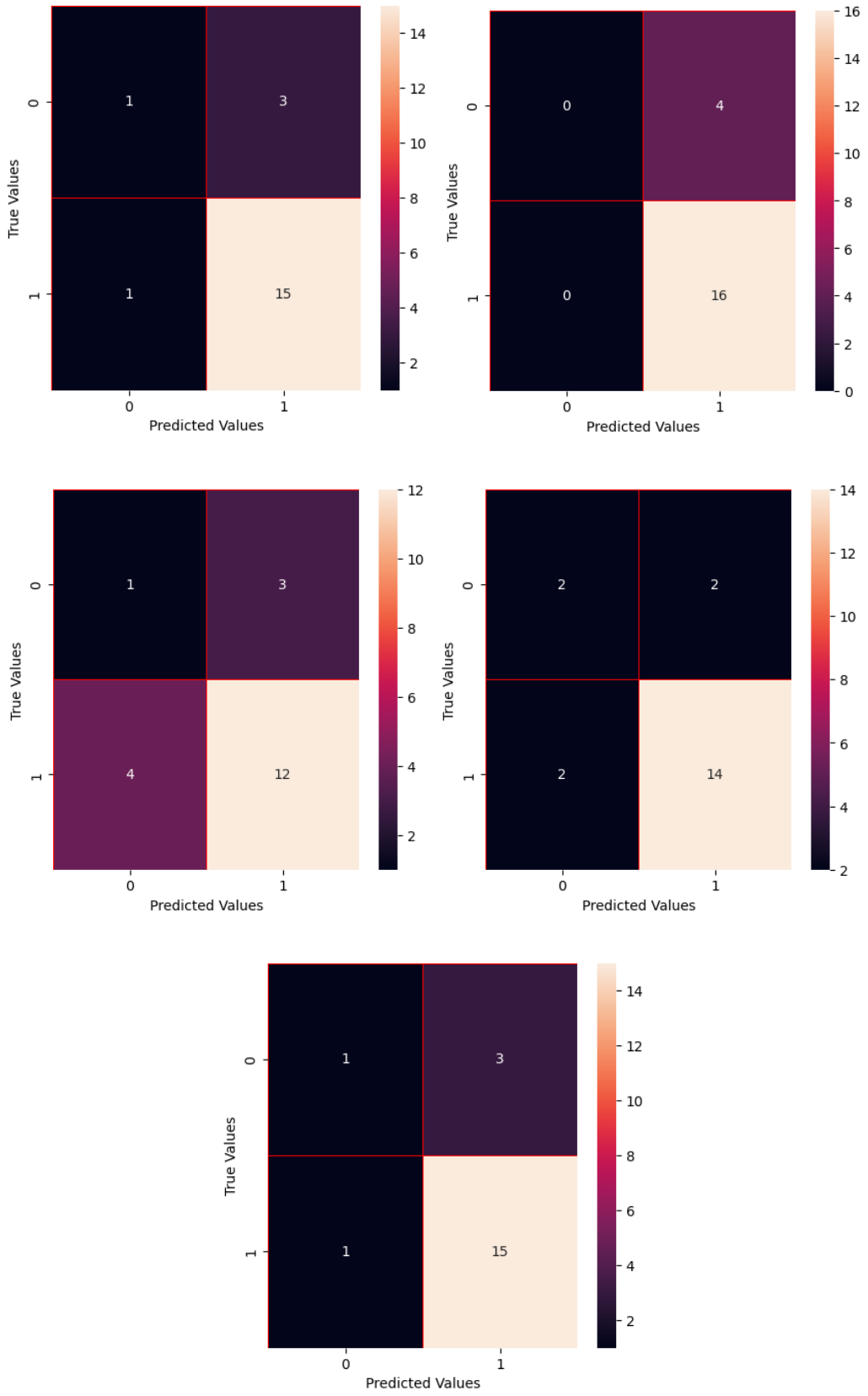


Figure 7. Obtained Confusion Matrix from various algorithms

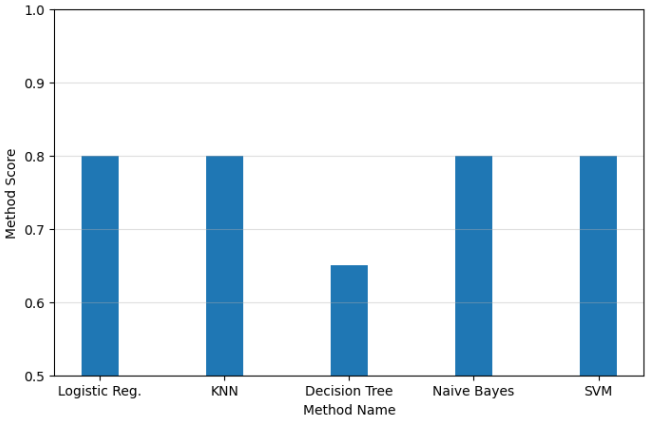


Figure 8. Final Accuracy Scores of different Algorithms [14]

Calculations are performed as under [15]:

- Precision = TP/(FP+TP)
- Recall = TP/(FN+TP)
- Accuracy = (TP+TN)/(TP+FN+TN+FP)
- F Score = 2* Precision Score * Recall Score/ (Precision Score + Recall Score)

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Figure 9. Confusion Matrix Model

5. Conclusion

The primary focus of this investigation was prostate cancer, which is one of the leading causes of cancer-related death in men and has symptoms that are similar to those of benign growths. Disease diagnosis is one of medicine's major challenges. Due to the absence of particular standards for evaluating prostate cancer indications and the inadequate predictive precision of the currently employed screening methods, this study is important. Where there are no clear-cut criteria and the elements influencing the occurrence can be predicted, we believe machine learning techniques utilizing Python can be useful in forecasting prostate cancer issues. In the context of this work, multiple supervised machine learning methods for prostate cancer prediction were assessed. Algorithms for machine learning were used for this. Using a dataset of 100 patients with prostate cancer from the open-source Kaggle platform, the efficacy of these algorithms is assessed. A few academic publications have also made use of this dataset, which is publicly available online as of 2018.

So, we may conclude that a computer that has been trained using machine learning techniques using patient data can be clinically valuable and highly accurate at predicting cancer. A needless biopsy of the patient can be avoided in this way.

6. References

[1] Geoffrey M Cooper, The Cell: A Molecular Approach, National Library of Medicine, [HTML Document] National Centre for Biotechnology Information, United States.

[2] Ashesh Anand (2021), Top 6 Machine Learning Techniques, [HTML Document] Analytic Steps, from World Wide Web.

[3] Zhang et al. (2019) Exploratory Data Analysis of Prostate Cancer Data [HTML Document].

[4] Rates Reardon et al. (2017), Exploratory Data Analysis of Prostate Cancer Mortality.

[5] Yang et al. (2018), Exploratory Analysis of Prostate Cancer Data Using Visualization Techniques.

[6] Rafiq and Khan (2020), Exploratory Data Analysis of Prostate Cancer Using Machine Learning and Visualization.

[7] Gha semi et al. (2019), A Comprehensive Exploratory Data Analysis of Prostate Cancer Data.

[8] Erdem, E. & Bozkurt, F. (2021). A Comparison of Various Supervised Machine Learning Techniques for Prostate Cancer Prediction. European Journal of Science and Technology, (21), 610-620

[9] Haq Nawaz (2022), Develop a Logistic Regression Machine Learning Model, Dev Genius, from World Wide Web.

[10] University of Regina, oURspace, from World Wide Web [HTML Document].

[11] Roshna S H, K-Nearest Neighbors algorithm, Intuitive Tutorials, from World Wide Web [HTML Document].

[12] Amrutha K, Jan (2022), Decision Tree Machine Learning Algorithm, Analytics Vidhya, from World Wide Web [HTML Document].

[13] Koushiki Dasgupta Chaudhuri (2022), Building Naïve Bayes classifier from scratch to perform sentiment analysis, Analytics Vidhya [HTML Document].

[14] Alihan Tabak (2019), Prostate Cancer Predictions with ML and DL Methods, Prostate Cancer, Kaggle.

[15] Harikrishan N B (2019), Binary classification metric, Analytics Vidhya, from World Wide Web [HTML Document].

AMITY UNIVERSITY
UTTAR PRADESH

AMITY INSTITUTE OF BIOTECHNOLOGY

Term Paper

Student Name	KESHAV MITTAL
Enrolment No	A0504222057
Programme	BTech (Bioinformatics)
Company's Name and Address	AMITY INSTITUTE OF BIOTECHNOLOGY Amity Institute of Biotechnology, Amity University, Noida, Uttar Pradesh, India 302017
Industry Guide	
Name	Dr Abhishek Sengupta
Designation	Assistant professor
Contact Number	
	Ph.(O) : 8800662904 (R) : 8800662904
	Mobile : 8800662904
	Fax : 01204392195
	E-mail : dpkatare@amity.edu

Project Information

1) Project Duration: **(40 Days)**

a) Date of Summer Internship commencement **(05/06/2023)**

a) Date of Summer Internship Completion **(14/07/2023)**

2) Topic

Exploratory Data Analysis on prostate cancer using python

3) Project Objective

The objective of exploratory data analysis (EDA) on prostate cancer using Python is to gain insights and understanding about the data related to prostate cancer. Further, using binary method, predict whether the following person is prone to the disease or not.

4) Methodology to be adopted

The methodology for conducting exploratory data analysis (EDA) on prostate cancer using Python typically involves; Data acquisition or importing of data from Kaggle in CSV form, Data loading and preprocessing, Data exploration and visualization, Correlation analysis, Feature selection, Hypothesis generation and testing, Documentation and reporting.

5) **Brief Summary of project (*to be duly certified by the industry guide*)**

The project aims to perform exploratory data analysis (EDA) on prostate cancer using Python. Through this project, we aim to uncover patterns, relationships, and anomalies in the dataset, identify potential risk factors or biomarkers, understand the impact of variables on prostate cancer outcomes, and generate research hypotheses for further investigation. The results of the EDA will provide valuable insights for subsequent analysis, modelling, or decision-making related to prostate cancer.

AMITY UNIVERSITY

UTTAR PRADESH

AMITY INSTITUTE OF BIOTECHNOLOGY

Term Paper

Weekly Progress Report (WPR1)

For the week commencing: 05/06/2023

Enrolment number: A0504222057

Student name: Keshav Mittal

Topic: EXPLORATORY DATA ANALYSIS ON PROSTATE CANCER USING PYTHON

Faculty guide name: DR. ABHISHEK SENGUPTA

Targets of the week:

1. Understanding of the topic by reading articles.
2. Understanding the basics of programming using python.
3. Forming a preliminary structure of the report.
4. Understating the use of data analytics in prostate cancer.

Achievements of the week:

1. The structure of the report is ongoing.
2. Basics of Exploratory Data Analysis have been understood.
3. Multiple articles related to the topic have been selected.

Plan for the next week:

1. Extracting the data from Kaggle.
2. To finalize the structure of the report.
3. To learn more about the various attributes related to prostate cancer.

AMITY UNIVERSITY
UTTAR PRADESH

AMITY INSTITUTE OF BIOTECHNOLOGY

Term Paper

Weekly Progress Report (WPR2)

For the week commencing: 12/06/2023

Enrolment number: A0504222057

Student name: Keshav Mittal

Topic: EXPLORATORY DATA ANALYSIS ON PROSTATE CANCER USING PYTHON

Faculty guide name: DR. ABHISHEK SENGUPTA

Targets of the week:

1. Extracting the data from Kaggle.
2. To finalize the structure of the report.
3. To learn more about the various attributes related to prostate cancer.

Achievements of the week:

1. The dataset from Kaggle has been extracted.
2. Different methods of machine learning for accuracy prediction have been studied.
3. The structure of the report has been finalized.

Plan for the next week:

1. To draft the introduction of the report.
2. Perform different algorithms for the machine learning.
3. Perform data analysis on the extracted dataset.

AMITY UNIVERSITY
UTTAR PRADESH

AMITY INSTITUTE OF BIOTECHNOLOGY

Term Paper

Weekly Progress Report (WPR3)

For the week commencing: 19/06/2023

Enrolment number: A0504222057

Student name: Keshav Mittal

Topic: EXPLORATORY DATA ANALYSIS ON PROSTATE CANCER USING PYTHON

Faculty guide name: DR. ABHISHEK SENGUPTA

Targets of the week:

1. To draft the introduction of the report.
2. Perform different algorithms for the machine learning.
3. Perform data analysis on the extracted dataset.

Achievements of the week:

1. Introduction of the report has been drafted.
2. Data analysis using algorithms is completed.
3. Plotting of graphs using information from the dataset has been achieved.

Plan for the next week:

1. To complete the major work of the report.
2. Form a rough first draft and get verified by the IFC.
3. To work upon the different figures required for the report.

AMITY UNIVERSITY
UTTAR PRADESH

AMITY INSTITUTE OF BIOTECHNOLOGY

Term Paper

Weekly Progress Report (WPR4)

For the week commencing: 26/06/2023

Enrolment number: A0504222057

Student name: Keshav Mittal

Topic: EXPLORATORY DATA ANALYSIS ON PROSTATE CANCER USING PYTHON

Faculty guide name: DR. ABHISHEK SENGUPTA

Targets of the week:

1. To complete the major work of the report.
2. Form a rough first draft and get verified by the IFC.
3. To work upon the different figures required for the report.

Achievements of the week:

1. The major bulk of the report has been prepared.
2. Different figures and graphs required for report have been prepared.
3. Self-check plagiarism is completed.

Plan for the next week:

1. To format the term paper and submit to IFC.

AMITY UNIVERSITY
UTTAR PRADESH

AMITY INSTITUTE OF BIOTECHNOLOGY

SYSTEMS BIOLOGY AND DATA ANALYTICS RESEARCH LAB

Certificate

This is to certify that the '**Term Paper**' work entitled **Exploratory Data Analysis on Prostate Cancer** has been carried out by Mr. Keshav Mittal, BTech in Bioinformatics, Amity Institute of Biotechnology, Noida as a part of NTCC work.

The work has been done under my supervision and guidance at Amity Institute of Biotechnology for a period of **5th June 2023** to **14th July 2023**

Date: **7th July 2023**
Place: **Noida**

Dr. Abhishek Sengupta
Assistant Professor
Systems Biology & Data Analytics Research Lab
Centre for Computational Biology & Bioinformatics
Amity Institute of Biotechnology
Amity University Uttar Pradesh, Sector 125
NOIDA, Uttar Pradesh, INDIA
Pin: 201313
Mobile: +91-8800662904
Email: asengupta@amity.edu, Web: www.amity.edu/aib