EXPLORING DATA SCIENCE FOR HIGHLIGHTING BREAST CANCER PREDICTION USING PYTHON

Ashmeet Sahni[1], Shikha Singh[2], Garima Srivastava[3]*

*[*1]Student, Department of Computer Science and Engineering, ASET, Amity University Lucknow Campus, (Uttar Pradesh), INDIA (Email id: ashmeetsahni91@gmail.com)*
[2&3]Assistant Professor, Department of Computer Science and Engineering, ASET, Amity University Lucknow Campus, (Uttar Pradesh), INDIA (Email id: ssingh8@lko.amity.edu, gsrivastava1@lko.amity.edu)

Abstract: This paper summarizes Data science which is an upcoming and advanced field in technology. We know that in today's world storing large amounts of data is not easy and hence we use data science for it. Data science is a multi- disciplinary field that makes use of algorithms, processes to extract information from structured and unstructured data. Breast cancer is a form of cancer that develops in the breast cells. Breast cancer is the second most common cancer diagnosed in women all over the world, after skin cancer. Breast cancer can strike both men and women, but it affects women much more frequently. Major support for breast cancer awareness and education has assisted in the advancement of breast cancer diagnosis and care. This paper focuses on the use of data science to predict Breast Cancer. Generally, we use two languages for coding in Data Science – Python and R. Authors in this paper have used Python for implementation of the work. The main focus of this paper will revolve around Data Science, Python and Breast Cancer Prediction. Data science has importance in various fields and has shown itself as well be it medical or any other it has been on the list. The main focus of the paper is to lay emphasis on breast cancer prediction and use of python and data science in it.

Keywords: Breast Cancer Prediction, Data , Data mining, Machine Learning Algorithms, Data Analytics, Python.

I. INTRODUCTION

In today's world we have a large amount of data to be stored and knowledge to be extracted; all this can be done through certain algorithms, processes or methods that are enabled by Data science. Data science is an upcoming technology that is catching people's attention. More and more people are now interested in this field. Data science works on DATA as the name suggests. Data is of 2 types- structured and unstructured. Data science has its roots in all fields. Data science is nothing but the same as Data Mining and Big Data. It uses the most efficient and powerful algorithms and programming systems to solve the problems. Data science is not only a particular group but it is like an umbrella containing many sub groups. Data analytics is a part of this wide umbrella. A person who is a professional in this field of data science is known as a Data Scientist [1]. A Data scientist is a person who works with big data sets that contain large amounts of data A Data Scientist is a person who is well versed with the data sets and how to extract large amounts of data. Since we need programming, we use python because python is very simple, straightforward syntax, case sensitive, uses variables without declaration hence it makes the user similar to programming faster.

The learning path of python starts with -PYTHON BASICS and end on PYTHON FOR IOT including GUI app with python, web app with python, python for data science, python for ai and machine learning, big data analysis with python. The report starts with data science with python that includes how a beginner starts it. So it starts with basics of python, packages, data science, that includes data visualization, data mining, regression analysis along with the practical examples and syntax of all and it will end with the project that I have focused on. Project is entitled "Breast Cancer Prediction". This project works on detection of the type of cancer and the data that the data set contains. It works on visualization of data and the plotting it by using the package Matplotlib. Further it focuses on all algorithms of machine learning to find out which algorithm is best for finding its accuracy.

II. LITERATURE REVIEW

Several studies have been conducted on the implementation of ML on Breast Cancer detection and diagnosis using different methods or combination of several algorithms to increase the accuracy. S. Gc et al. [1] worked on extracting features including variance, range, and compactness. They used SVM classification to evaluate the performance. Their findings showed the highest variance of 95%, range 94%, compactness 86%. According to their results, SVM can be considered as an appropriate method for Breast Cancer Detection. 24% 13% 10% 6% 6% 41% Breast Trachea, Bronchus, Lung Colorectum Ovary Cervix Uteri Other Early Detection of Breast Cancer Using Machine Learning Techniques.

Chunqiu Wang et al. [2] chose Microwave Tomography Imaging (MTI) to extract features and classify the images using ANN. Two different techniques were compared in this study, GMM and KNN. Their results showed that the sensitivity obtained by KNN is 87%, while for GMM is 67%. The accuracy was 85% for KNN and 75% for GMM. The result for Matthews Correlation Coefficient (MCC) was 67% and 48% for KNN and GMM, respectively. Finally, the specificity was 84% for KNN and 86% for GMM. According to their findings, Sensitivity, Accuracy, and MCC for KNN were better than GMM, but GMM was better in Specificity and Precision.

Chowdhary and Acharjya [3] focused on mammogram images as they are cheaper and more efficient in detection. However, since selecting and extracting features are important for improving performance, Fuzzy Histogram Hyperbolization(FHH) was chosen to increase the quality of images, Fuzzy C-mean for segmenting, and Gray level dependence model for extracting the features. Their method showed 94% accuracy for detecting malignant breast lesions.

In a study conducted by Amin Khan ghahi et al. [4], wireless cyber mammography images were explored. After selecting features and extracting them, the researcher has chosen two different ML techniques, SVM and GMM to check their accuracy. Their findings showed that SVM is more accurate if there is no noise or error, else GMM is better and safer.

Durai et al. [5] have selected Data Mining techniques for detecting diseases including breast cancer. They used LRC and compared it with four other techniques including BFI, ID3, J48, and SVM. The result shows that LRC is the most accurate one with 99.25% accuracy.

III. DATA SCIENCE

Data science as the name suggests it's something related to data. Data science is the study of information, where it is coming from, what it is representing and how it can be turned into something productive so that it can be used for business and strategies. Mining large amounts of structured and unstructured data can be important for companies for increasing efficiencies, recognizing new market opportunities, recognizing the competition. Data science is a concept that surrounds maths, statistics,

computer science and information science. The different components of data science start from Data mining, data analytics, big data, data science to machine learning [2]. Data science can be said as a tool to reduce costs of items. Data is taken out from different sectors, channels and platforms including cell phones, social media, healthcare surveys. Data scientists are a group of people in today's era.
"According to IBM, the demand for data scientists is expected to increase by 28% by 2020."
Imagine how much data scientists are needed. Data scientists can be described. Storyteller who describes the entire data to decision makers in a way that is easy to understand and can help in solving problems. Companies in today's era are completely relying on big data for their work. Companies like Netflix, Amazon extract large amounts of data to determine what and which type of products that needs to be delivered to its users. Netflix uses algorithms to create a personalized recommendation list for their users based on the videos their users view. Data science growth is seen rapidly and will continue to grow in future. Working on large amounts of data is not easy and it is done by the data scientists. Data science is not just a single process but is a life cycle that consists of various steps. This life cycle includes the following as written below:

1- Data collection.
2- Data preparation.
3- Data analysis.
4- Modelling.
5- Model evaluation.
6- Model development.
7- Business understanding.
And this cycle continues.

A. Data Science Methodology

As data analytics capabilities become more and more accessible and difficult, a data scientist needs a particular strategy or rules to be followed regardless of the technologies, large amounts of data or any sort of approaches that are involved. This methodology of data science is quite similar to methodologies that have been discovered before for data mining, but it lays important emphasises on the new practices that are practiced in data science such as the excess use of large amount of data proudly known as data sets, the conversion of text data into models for better visualisation and automation of various processes.
A Methodology is a type of plan or strategy that helps in guiding the process or algorithms or activities within a particular area or domain. Methodology is not dependent on a particular technology or any tool, nor is it a particular set of techniques or recipes that need to be followed. Rather it helps in providing our data scientists a type of framework that they can work on [5]. It helps them to provide them with a layout on which they can work and understand what work need to be done first asin which method, processes and heuristics can be used to obtain the result which can be beneficial for our company. This methodology consists of in total 10 steps or stages that forms a continuous cycle for using data forming data sets and then uncover the insights. Each stage plays a vital role in the context of the overall process.

1--FROM BUSINESS UNDERSTANDING TO ANALYTIC APPROACH AND FROM DATA REQUIREMENTS TO DATA COLLECTION

A project starts with understanding and strategies. The business sponsor who needs a solution plays the most important part in this stage by defining the level of problem and deciding the objective accordingly. The first step lays emphasis on a successful solving of the problem. To solve this problem, we need to follow an approach known as analytic approach that is needed to solve the problem. This stage consists if machine learning techniques so that the company can decide the most suitable one for the outcomes. When we choose a particular approach then we require data, which is our next step known as data requirement. Data is required to be used in certain format or data content and representation that is guided by some knowledge. Since now we know the type of data, we need then we focus on the most important part that is data collection, we find out the data for formation of data sets[10].

2-FROM DATA UNDERSTANDING TO DATA PREPARATION AND FROM MODELLING TO EVALUATION

After the collection of data our major goal is to understand the data. Data scientists use descriptive statistics and various techniques to understand the content data needs. [3]Additional data collection may be required afterwards to fill the gap. Since now we have collected the entire data and understood it so now, we move to the next step that is DATA PREPARATION. Data preparation includes cleaning the data, eliminating duplicates, removing null values and combining data from multiple sources. Data preparation is the most time-consuming step because preparing the data is difficult. Sometimes some steps are common in all datasets but sometimes it is different. After that step that comes is modelling. Modelling is the process of representing the data sets through models or graphs for the user to understand it and make it look good. With these predictive models, data scientists uses predictive texts to build a efficient model. For a particular technique a data scientist may use various algorithms with the model to find the best one. Then we move on to evaluation, this is the step of finding or evaluating the models by seeing the result they are giving.

3-FROM DEPLOYMENT TO FEEDBACK

Once a particular satisfactory model is prepared and is approved by the company then it is deployed into a particular environment. Data is deployed in a limited way until its performance is evaluated on a full basis. Deployment is simple as it involves deploying a model into an operational process that involves skills and technologies.[4] Then comes the last step that is feedback whenever we do some work or complete our project the most important part is feedback so as to know whether the model made is efficient or not. The feedback can be in the form of response rates, growth or the reactions of the customers. This cycle continues as this methodology is iterative in nature. As a data scientist is learning they frequently return to the previous stage when they commit a mistake and this enables them to grow in a much better way. Models are not made once and left; change are made in them for improving them. In this way the model and the work can be improved and used for providing continuous profit [7] to the organisation as long as solution is needed.

B. Data Mining

DATA MINING is the process of extraction of data using various packages. Mainly data mining is the process of extraction of structured data from raw data. It means analysing data patterns in large forms using software's. Python uses four packages. -
1. NUMPY. NUMPY is a package used for scientific computing using python. NumPy works on multidimensional arrays, lists and has functions to work on them. Its importing is done.
2. PANDAS. PANDAS is a package that is used for providing fast, data and helps in labelling the data. It helps in practical, real world data analysis in python.
3. SCIPY is open source software that works on mathematics calculations including pie. It is user friendly and works on NumPy arrays.
4. MATPLOTLIB it works on data visualization that is representing the data with the help of graphs, pie charts and is extension of NumPy mathematical library. It uses GUI toolkit [9].

C. Data Visualization

DATA VISUALIZATION as the name suggests data visualization is the visual representation of data. The data can be represented through graphs, pie chart, bars etc. companies usually prefer representing data visually through diagrams graphs because a data which is represented through images is easy to understand and visualize. In the world of BIG DATA, visualization enables to analyse large amount of information and make decision on its basis. Our eyes are drawn to colours and graphs [6]. Common types of data visualization is done through-
Charts, Tables, Graphs, Maps, Gantt chart, Bar chart, Pie chart, Line chart, Scatter plot.
All this is done using the package MATPLOTLIB. This package is first imported and then worked on using various functions.

D. Data Analysis

DATA ANALYSIS is a process of inspecting, cleansing and modelling data with the ultimate goal of discovering useful data. Data analysis has two prominent methods. They are Qualitative research and Quantitative research. Both methods have different techniques to work on. Interviews,

observations are type of qualitative research and experiments, surveys are type of quantitative research [8].
To analysis data have steps in mind

Step1- Define your problem.
Step 2- Set your priorities.
Step 3- Collect data set.
Step 4- Analyse your data
Step 5- Interpret the answer or result.

Regression analysis is preferred in python. It's the process of plotting lines by finding slope. Regression involves working on lines. First, we try to find the slope of line manually and then with the help of regression. Best fit line is found and plotted. Two types of data sets are being worked on – TRAINING SET and TESTING SET. Regression is also of two types:

1- Linear regression.
2- Logistic regression.

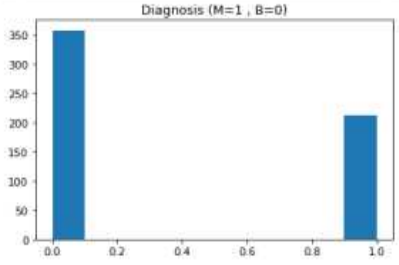
Linear regression means fitting a straight line according to the data. It is popularly known as "linear modelling". It makes model easy to understand and allows predictions for the data set. A data set is taken, then will plot regression line for it and show data visualisation. After that will use sklearn and import linear regression, in same way import logistic regression too can be done.

IV. BREAST CANCER PREDICTION

Breast cancer prediction is a project that works on prediction of cancer cells. Whether the tumour is malignant or benign. This prediction will be done on WISCONSIN BREAST CANCER DIAGNOSTIC DATA SET. This data set is of 2016. There are mainly two attributes
1) ID number
2) Diagnosis (M or D)

The nucleus is tested to determine the type of tumours; hence ten real valued features are examined and their data set is prepared for each cell nucleus
A. Radius
B. Texture (standard deviation of the grey scale values)
C. Perimeter
D. Area
E. Smoothness (difference in length of radius) F. Compactness(perimeter^2/area-1.0)
G. Concavity (concave portions)
H. Concave points (number of concave portions) I. Symmetry
J. Fractal dimension ("coastline approx.-1" We will find out the mean, standard and worst error of each and hence we will get 30 information. We will analyse the data and compute the result accordingly.
We will plot the bar graph using the information.

```
In [18]: df.describe()
         plt.hist(df['diagnosis'])
         plt.title('Diagnosis (M=1 , B=0)')
         plt.show()
```




```
In [19]: features_mean=list(df.columns[1:11])
         dfM=df[df['diagnosis'] ==1]
         dfB=df[df['diagnosis'] ==0]

In [20]: plt.rcParams.update({'font.size': 8})
         fig, axes = plt.subplots(nrows=5, ncols=2, figsize=(8,10))
         axes = axes.ravel()
         for idx,ax in enumerate(axes):
             ax.figure
```

Figure 1: Plotting graph

Plotting each separately we will get such graphs and then we can perform various machine learning algorithm to predict which one is better.So as to perform all the algorithms on the data set. Then all the algorithms are performed so as to know about its accuracy.

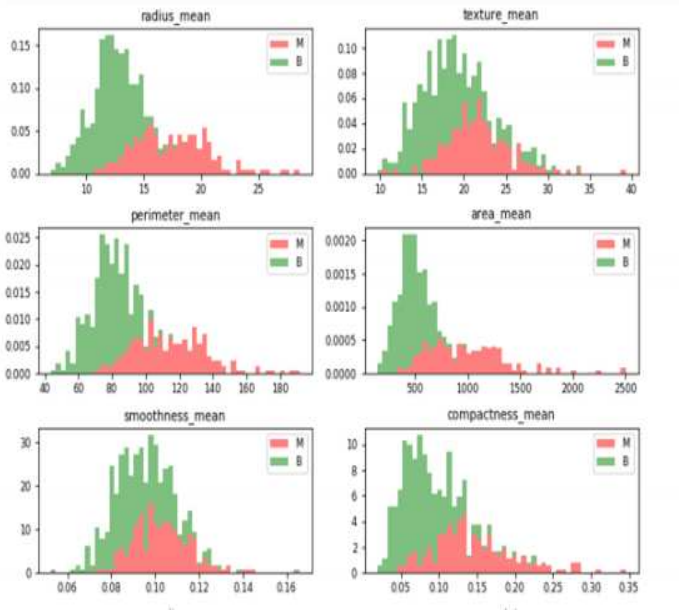


Figure 2: Comparing the Algorithms

V. CONCLUSION

The paper revolves around Data Science using Python. Data science is nothing but an explanation of how we can do mathematics, algorithms and programming together. It's a field that is improving day by day and has a lot demand in market. Data scientist are the highly paid employs in today's world. Data science is endless because it has it route in various other fields like machine learning, artificial intelligence, Internet of things. Presently authors have worked on data science using python and have worked on Breast Cancer Prediction. This technology has been remarkably helpful in making predictions on this deadly disease, which presently has affected a large section of our society and specifically women.

VI. FUTURE SCOPE

The future scope of this is its use in developing more projects as well as the project can be used in determining the factors that lead to breast cancer and preventing it to happen. The most important scope is determining how many people get affected by it and the cells that can cause such disease. It can be widely used by biotechnology department to work on the cells.

REFERENCES

1. ,M Sharma, SP Pradhyumna, S Goyal, K Singh Data Analytics and Management, 229-250 "Machine Learning and Evolutionary Algorithms for the Diagnosis and Detection of Alzheimer's Disease"
2. Authors Ayush Chauhan, Deepali Kamthania Publication date 2021 Book "Data Analytics and Management Pages" 31-39 Publisher Springer, Singapore ANN Model for Forest Cover Classification
3. Frank, Eibe; Hall, Mark A. (30 January 2011). Data Mining: Practical Machine Learning Tools and Techniques (3 ed.). Elsevier. ISBN 978-0-12-374856-0.
4. Chambers, John M. (1 December 1993). "Greater or lesser statistics: a choice for future research". Statistics and Computing. 3 (4): 182–184. doi:10.1007/BF00141776. ISSN 0960-3174.

5. Machine Learning Forensics for Law Enforcement, Security, and Intelligence. Boca Raton, FL: CRC Press (Taylor & Francis Group). ISBN 978-1- 4398-6069-4.
6. Manuela Aparicio and Carlos J. Costa (November 2014). "Data visualization". Communication Design Quarterly Review. 3 (1): 7–11. doi:10.1145/2721882.2721883.
7. Tufte, Edward (1983). The Visual Display of Quantitative Information. Cheshire, Connecticut: Graphics Press. ISBN 0-9613921-4-2.
8. Whitehouse, D. (9 August 2000). "Ice Age star map discovered". BBC News. Retrieved 20 January 2018. [9]Xia, B. S., & Gong, P. (2015). Review of business intelligence through data analysis. Benchmarking, 21(2), 300-311. doi:10.1108/BIJ-08-2012-0050
9. Aarons, D. (2009). Report finds states on course to build pupil-data systems. Education Week, 29(13), 6.1.