**Classification Of Breast Cancer**

**Project-Based Internship 2022 Report**

Submitted to

**CS Department ,ABESEC**

**Duration : Six Weeks**

**By**

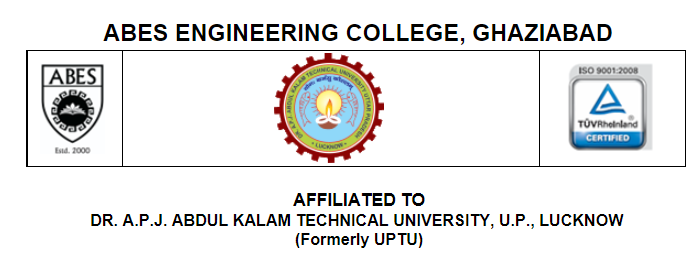
**Aditya Kumar Ojha: 1900320150003**

**Sakshi Singh: 1900320150043**

**Under the guidance of:-**

**1) MR. ASHWIN PERTI**

**2) MS. SAPNA JAIN**

****

**TABLE OF CONTENTS**

Page No.

STUDENTS DECLARATION ii

CERTIFICATE iii

ACKNOWLEDGEMENTS iv

ABSTRACT v

LIST OF TABLES vii

LIST OF FIGURES viii

**CHAPTER-1 INTRODUCTION 9**

1.1 Problem Definition 9

1.2 Motivation

1.3 Objective of the Project

1.4 Scope of the Project

1.5 Need of Work 9

**CHAPTER 2 RELATED WORK 10**

**CHAPTER 3 PROPOSED METHODOLOGY 11**

3.1 Dataset Description 11

3.2 Methods

3.3 Hardware / Software Requirements 11

3.3 Our Methodology 11

**CHAPTER 4 EXPERIMENT AND RESULT ANALYSIS 17**

**CHAPTER 5 CONCLUSION**

**5.1 Discussion**

**5.2 Future Work**

**REFERENCES**

**STUDENT’S DECLARATION**

 We hereby declare that the work being presented in this report entitled **“Classification Of Breast Cancer”** is an authentic record of my / our own work carried out under the supervision of Mr Gopal Gupta/ Mr Ashwin Perti/ Ms Sapna Jain.

The matter embodied in this report has not been submitted by me / us for the award of any other degree.

**Date:** **Signature of students**

## CERTIFICATE

This is to certify that Project Report entitled “Classification Of Breast Cancer” which is submitted by Aditya Kumar Ojha and Sakshi Singh in partial fulfillment of the requirement for the summer internship of Data Analysis and Machine Learning using Python in Department of Computer Science of ABES Engineering College, is a record of the candidate own work carried out by him under my/our supervision.

**Signature of HOD  Supervisor 1: Mr Ashwin Perti**

**(Dr. Pankaj Sharma)**

**(Computer Science)  Supervisor 2: Sapna Jain**

**Date.: -**

ACKNOWLEDGEMENT

*It gives us a great sense of pleasure to present the report of the Project Based Internship 2022 undertaken during B-tech 3rd year. We owe special debt of gratitude to Gopal Sir/Ashwin Perti/Sapna Jain facilitators for HET Data Science Training for their constant support and guidance throughout the course of our work. Their constant motivation have been a source of inspiration for us. It is only their cognizant efforts that our endeavors have seen light of the day.*

*We also do not like to miss the opportunity to acknowledge the motivation of Computer Science Department to provide us the opportunity to undergo this training.*

*Signature:*

*Name : Aditya Kumar Ojha*

*Roll No : 1900320150003*

*Date :*

*Signature:*

*Name : Sakshi Singh*

*Roll No : 1900320150043*

*Date :*

**ABSTRACT**

Each year number of deaths is increasing extremely because of breast cancer. It is the most frequent type of all cancers and the major cause of death in women worldwide. Any development for prediction and diagnosis of cancer disease is capital important for a healthy life.

Consequently, high accuracy in cancer prediction is important to update the treatment aspect and the survivability standard of patients. Machine learning techniques can bring a large contribute on the process of prediction and early diagnosis of breast cancer, became a research hotspot and has been proved as a strong technique.

In this study, we’ll apply five machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5) and K-Nearest Neighbours (KNN) on the Breast Cancer Wisconsin Diagnostic dataset, after obtaining the results, a performance evaluation and comparison will be carried out between these different classifiers.

The main objective of this research paper is to predict and diagnosis breast cancer, using machine-learning algorithms, and find out the most effective whit respect to confusion matrix, accuracy and precision. All the work is done in the Jupyter environment based on python programming language and Scikit-learn library.

**LIST OF FIGURES**

Description of fig pgno.

1. Dataset description showing

no. of rows of data and how 11

much files are there in range.

1. Displaying our methodology. 13
2. Displaying the conversion of 15

null values.

1. Displaying our processed data. 15
2. Showing the accuracy of our 15

Machine.

**CHAPTER 1**

**INTRODUCTION**

* 1. **Problem Definition**

The goal of this study is to see which features are most useful in predicting whether a cancer is malignant or benign, as well as to look for general trends that can help with model selection and hyper parameter selection. The objective is to determine if the cancer is benign or malignant. We did this by fitting a function that can predict the discrete class of new input using data science classification algorithms.

* 1. **Motivation**

Breast cancer is one of the most commonly diagnosed malignancies in both men and women around the world. When cells in the breast begin to develop abnormally, breast cancer develops. Every year, a large number of people die as a result of this malignancy. It is the most common type of cancer and the leading cause of death in women around the world.

Classification and data mining technologies are particularly effective approaches for classifying data nowadays. Especially in the medical industry, where machine learning is used to diagnose diseases and conduct analyses in order to make specific recommendations.

So, if we can classify which sort of cancer a patient has using Data Science, doctors will be able to deliver prompt treatment to patients and improve their chances of survival

.

* 1. **Objective of the Project:**

In this Data Science assignment, we will evaluate and classify Breast Cancer (to determine which category the breast cancer belongs to), as there are two types of breast cancer:

• Breast cancer of the malignant type

• Breast cancer of the benign type

So the major goal of this research is to construct a model that can correctly classify whether breast cancer is malignant or benign using a dataset.

.

* 1. **Scope of the Project:**

More research in this area is needed to improve the classification techniques' performance so that they can forecast on more variables. We're trying to figure out how to parametrize our categorization approaches so that we can get high accuracy. We're investigating a variety of datasets to see how data science may be utilized to better characterize Breast Cancer. We want tor educe mistake rates as much as possible while maintaining optimum accuracy.

* 1. **Need of Work**

The need of classification is to select the best treatment. The effectiveness of a specific treatment is demonstrated for a specific breast cancer (usually by randomized, controlled trials). That treatment may not be effective in a different breast cancer. Some breast cancers are aggressive and life-threatening, and must be treated with aggressive treatments that have major adverse effects. Other breast cancers are less aggressive and can be treated with less aggressive treatments, such as lumpectomy.

**CHAPTER 2**

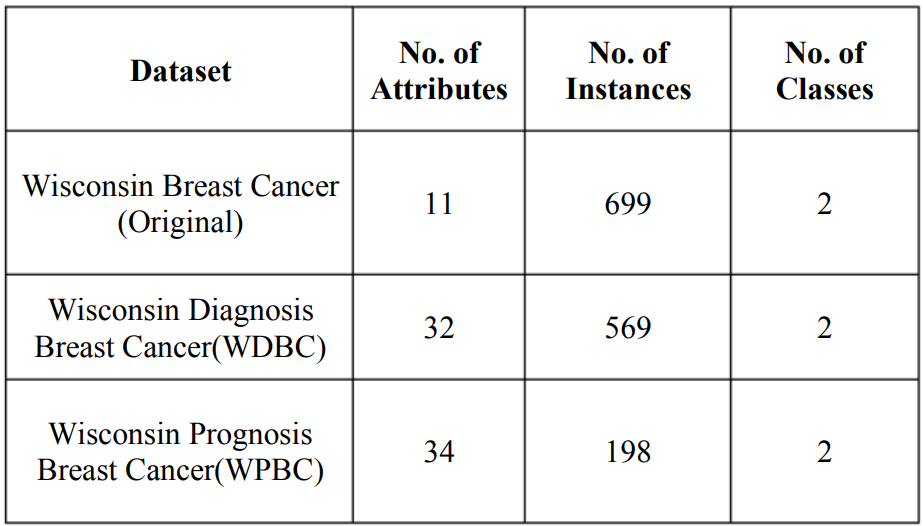
**RELATED WORK**

Twenty-four recent research articles have been reviewed to explore the computational methods to predict breast cancer. The summaries of them are presented below:

* Chaurasia et al. developed prediction models of benign and malignant breast cancer. Wisconsin breast cancer data set was used. The dataset contained 699 instances, two classes (malignant and benign), and nine integervalued clinical attributes such as uniformity of cell size. In addition, the researchers conducted sensitivity analysis and specificity analysis of the three algorithms to gain insight into the relative contribution of the independent variables to predict survival. The sensitivity results indicated that the prognosis factor Class was by far the most important predictor.
* Puspanjali Mohapatra et al. proposed the model based on the enhancement of histopathological images using deep learning. In this paper, many methods were used for feature extraction like PCA and LDA. The author also talked about machine learning techniques but there were large dataset so machine learning techniques don’t give better results. That’s why deep learning was used for that purpose.
* Tina Elizabeth Mathew et al. proposed the model based on Decision Tree classifier on breast cancer. The Wisconsin breast cancer dataset was implemented by decision tree. The paper also talked about naïve Bayes tree, rotation forest for classification. The research was done on the WEKA environment

**CHAPTER 3**

**PROPOSED METHODOLOGY**

**3.1 Dataset Description:** The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used , to distinguish malignant (cancerous) from benign (non-cancerous) samples. A brief description of these datasets is presented in the figure 1. Each dataset consists of some classification patterns or instances with a set of numerical features or attributes.

**3.2 Methods:** Breast cancer detection through an intelligent system is vital in the medical field. Various methods can be applied for classification of breast cancer such as**Neural Network, Support Vector Machine, KNN and decision tree.**

**3.3 Hardware/ Software Requirements**:

Hardware requirements:

• System : Pentium Dual Core.

• Hard Disk : 100 GB.

• Monitor : 14.3’’ LED

• Input Devices : Keyboard, Mouse

• Ram : 1 GB

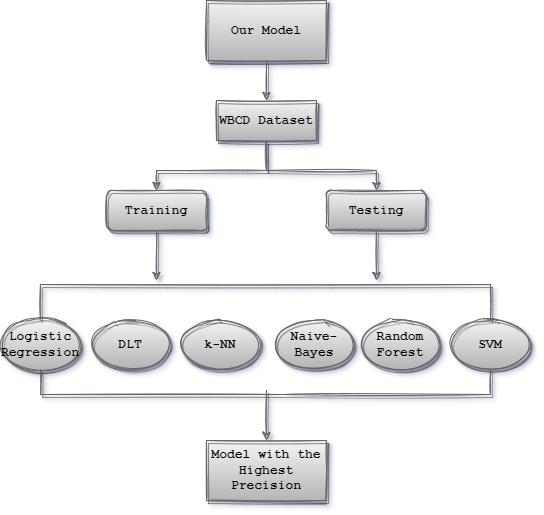
Software requirements:

• Operating system : Windows 7/8/10/11.

Coding Language : Python

• Database : MS Excel

**3.3 Our Methodology:**



**CHAPTER 4**

**EXPERIMENT AND RESULT ANALYSIS**

**Experiment:**

We used Jupyter Notebook for our project to run python. The following are the libraries used:

• numpy

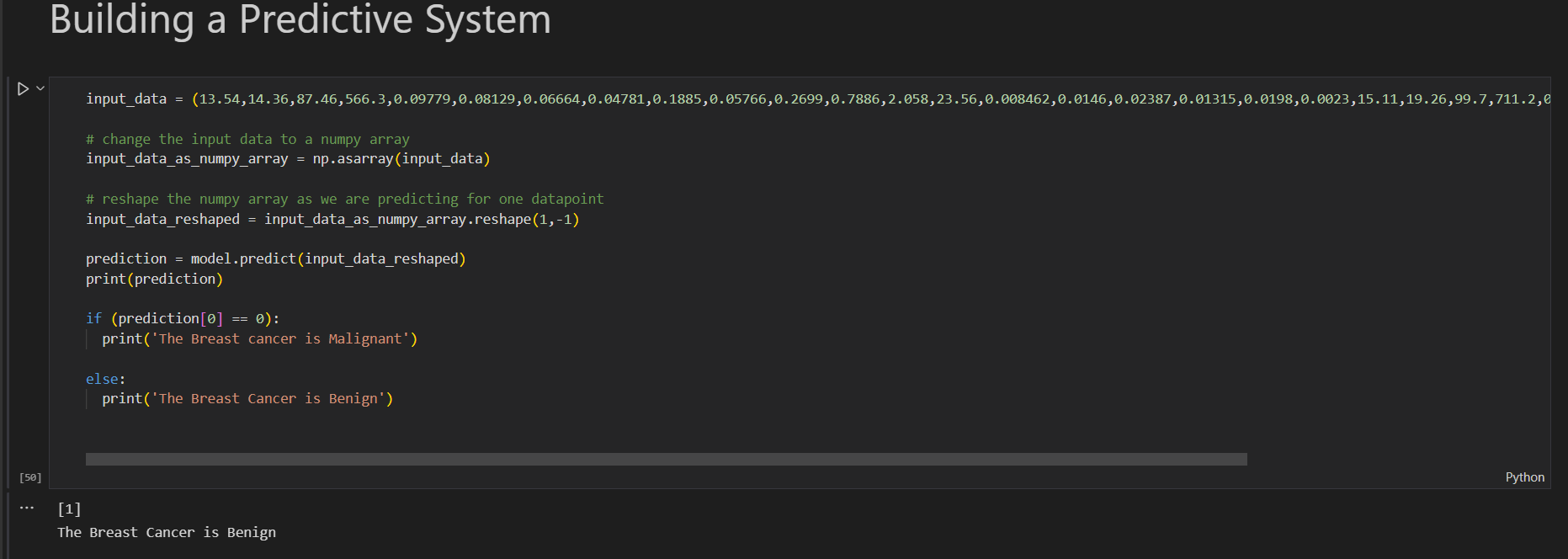
• pandas

• sklearn

We collected data from Kaggle in .csv format and then we processed it. After this we split the features and the target from each other. Further we split the data into training data and testing data. This is essential in to avoid overtraining because if we train the model 100% on the given data then it will learn the data and give correct output for the values in dataset but if we test on some other data, the prediction made may not be correct. Thus the training and testing data is divided in ratio 80:20 where 80 is training dataset and 20 is for testing. For training and building the model we made use of logistic regression and fitted the training data.

**Result Analysis:**

Now for checking whether the model we have created is making correct or incorrect prediction, we calculate accuracy. Accuracy is a measure by which we actually measure the correctness of our model. Higher the accuracy, better the predictions are made. We may have lower accuracy if training is not done properly. Therefore, if we get lower accuracy of model then we must train our machine learning model with more data so that it can learn properly

**Model Evaluation:**

**CHAPTER 5**

**CONCLUSION**

**Discussion:**

Breast cancer, if detected early, can save the lives of thousands of women and even men. These programs assist patients and clinicians in gathering as much information as possible in the real world. The data for the idea we offered was gathered through research on nine papers. We will be able to classify and forecast whether a cancer is benign or malignant using machine learning algorithms. Data Science can be utilised in medical research because they improve the system, minimise human errors, and reduce manual errors.

**Future Scope:**

Diagnosis of younger women is more difficult than that of older women owing to the higher breast tissue density of the former group. In previous models, the detection of one [cancer cell](https://www.sciencedirect.com/topics/medicine-and-dentistry/cancer-cell) was based on the performance of the neighbouring cancer cells. A cancerous cell could only be evaluated in relation to another cancerous cell. The use of [biosensors](https://www.sciencedirect.com/topics/engineering/biosensors) for cancer detection in the future shows great promise.

**REFERENCES**

* https://www.who.int/cancer/prevention/dia gnosisscreening/breast-cancer/en/
* https://towardsdatascience.com/building-a-simplemachine-learning-model-onbreast-cancer-data-eca4b3b99fa3
* OriginaldataSet:http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic% 29
* Confusion Matrix: https://tatwan.github.io/How-To-Plot-AConfusion-Matrix- In-Python
* https://seaborn.pydata.org/tutorial/axis\_gri ds.html