School of
Electrical and Electronics Engineering

Course Project Report

on

Prediction of breast cancer using machine learning algorithms

By:

1. **Rahul G Teli** USN: 01FE21BEE008
2. **Malhar Kulkarni** USN: 01FE21BEE016
3. **Pavan Kurtakoti** USN: 01FE21BEE007
4. **Chandrashekhar Angadi** USN: 01FE21BEE031

Semester: VI, 2023-2024

Under the Guidance of

Prof. Virupaxayya Hosallimath

K.L.E SOCIETY'S

KLE Technological University,
HUBBALLI-580031 2023-2024

SCHOOL OF ELECTRICAL AND ELECTRONICS ENGINEERING

CERTIFICATE

This is to certify that project entitled **""Prediction of breast cancer using machine learning algorithms"** is a bonafide work carried out by the student team of" **Rahul G Teli- 01FE21BEE008, Malhar Kulkarni- 01FE21BEE016, Pavan Kurtakoti- 01FE21BEE007, Chandrashekar R Angadi - 01FE21BEE031"**. The project report has been approved as it satisfies the requirements with respect to the course project work prescribed by the university curriculum for BE (VI Semester) in School of Electrical and Electronics Engineering of KLE Technological University for the academic year 2023-2024.

Prof. Virupaxayya Hosallimath
Guide

Ms. Kavita Chachadi
Guide

ABSTRACT

Cancer in whole have become a new normal in the 'disease' world and especially in this growing generation. Many are contributing to the risk phenomenon such as dietary conditions. Lifestyle too plays a major role here because many regret to do, eat or make something of a good will. Almost no one is surveying these factors and these have led to a rapid growth on this tally for he past 20 years, or more. In the varied population of the Americas and to a wider aspect, this has become an inevitable circumstance. In this case, female aging 40 and above are more prone to two inexorable circumstances being Urinary Tract Infections(UTIs) in one hand and Breast Cancer in the other. This has become a frequently researched and scary topic among not only the physicians and researchers but with the youth population too. Till day there is not even a slightest cure to the deadliest disease among them all.

As from the earlier times, Inhibiting is is better than cure. this still fits to these day among us all. There are many kinds of tests and therapies to treat almost every time of cancers but till this day, a cure is the biggest question. This has been the case since its inception. Awareness is being created in the form of printing warning signs on the front of cigarette packets, chewing gums etc., but they must be mandatorily imposed upon the people to create a widespread impact.

Here in this paper Detection of breast cancer is easily elaborated to ease up the process before going professionally to get a small view on the prediction of the disease. The need to detect this disease earlier has been of course a growing concern among the people of every nation.

This Breast Cancer Prediction system is mainly aimed at predicting the accuracy on how furious the cancer have spread or how not at all. This code describes if the patient have cancer or not at all using the given input, predicting the accuracy.

Keywords— Random Forest Classifier, KNearest Neighbor (KNN) XGBoost, Regression, Classification, Mining, Training, Testing

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS	EXPANSION
ML	Machine Learning
AI	Artificial Intelligence
SNA	Social Network Analysis
RFM	Recency, Frequency, and Monetary
GUI	Graphical User Interface
EDA	Exploratory Data Analysis
RMSE	Root Mean Square Error
MAPE	Mean Absolute Percentage Error
MAE	Mean Absolute Error

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION OF THE PROJECT

Breast cancer occupies the very top place among the growing concern of fatal-cancer worldwide, as the symptoms regarding the first stage is very hard to find making it very deadly. This can be only cured, as mentioned earlier should be diagnosed earlier so that the victim can survive and lead an healthy lifestyle. Before having a detailed medical examination, Machine learning algorithms can almost predict the cancer using the given datasets and the circumstances. This not only helped in predicting the required accuracy but have also evolved to be a step-ahead method in terms of evaluating the disease. Many lives have been saved using this so called computer generated response. A woman named Ashley Graham denoted that she had developed a cancer in one breast, which makes her a subject of vicinity to develop cancer in the other breast too. This makes her vulnerable and makes up to one of the risk factors involving breast cancer.

Family history impinges the risk of aggrevating breast cancer too. If or before a family member of the suspected patient had cancer, it is advisable to undergo a test to find out if they are showing any symptom or not.

An aging population varying between ages 50 and 80 too can have adverse effect on breast cancer. People ageing anywhere between 60-65 can develop any type of cancer given the circumstances. Especially in woman they are more vulnerable to breast cancer than any other counterparts. This is because they go through a lot of situations, namely pregnancy, menopause etc., this poses them at a major risk.

1.2 MACHINE LEARNING

Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as 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 in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

Despite the fact that the reasons mentioned are valid, we have added a dimension in the last decade where data is being utilized for predicting what could potentially happen in the future. Then comes Machine Learning which play a significant role in doing so. Machine learning is a subset/subfield of Artificial Intelligence. Generally, the main aim of Machine learning is to understand the structure of data and apply the best possible models that can be utilized or identify a hidden pattern. Developing a machine learning model is one of the key factors in predicting a future problem which again requires machine learning algorithms. There are numerous machine learning algorithms that have been developed and mature enough to solve various real-world business problems.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Using Machine learning, information is being turned into knowledge. In the last 5-6 decades, enormous data has been recorded or collected which will be of no use if we don’t utilize or analyze to find hidden patterns. In order to find useful and significant patterns with complex data, we have several Machine Learning

techniques available to ease our struggle for discovery. Subsequently, those identified hidden patterns and knowledge of the problem can be helpful to perform complex decision making and predict future occurrence.

1.2.1 History and relationships to other fields

The term machine learning was coined in 1959 by Arthur Samuel, an American IBMer and pioneer in the field of computer gaming and artificial intelligence. Also, the synonym self-teaching computers was used in this time period. A representative book of machine learning research during the 1960s was the Nilsson's book on Learning Machines, dealing mostly with machine learning for pattern classification. Interest related to pattern recognition continued into the 1970s, as described by Duda and Hart in 1973.In 1981 a report was given on using teaching strategies so that a neural network learns to recognize 40 characters (26 letters, 10 digits, and 4 special symbols) from a computer terminal.

Modern day machine learning has two objectives, one is to classify data based on models which have been developed, the other purpose is to make predictions for future outcomes based on these models. A hypothetical algorithm specific to classifying data may use computer vision of moles coupled with supervised learning in order to train it to classify the cancerous moles. Whereas, a machine learning algorithm for stock trading may inform the trader of future potential predictions.

1.3 MACHINE LEARNING APPROACHES

In machine learning, tasks square measure is typically classified into broad classes. These classes square measure supported however learning is received or however, feedback on the education is given to the system developed. Two of the foremost wide adopted machine learning strategies are square measure supervised learning that trains algorithms supported example input and output information that's tagged by humans, and unattended learning that provides the algorithmic program with no tagged information to permit it to search out structure at intervals its computer file.

Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system:

Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). As it navigates its problem space, the program is provided feedback that's analogous to rewards, which it tries to maximize.

1.3.1 Supervised Learning

In supervised learning, the pc is given example inputs that square measure labeled with their desired outputs. The aim of this technique is for the algorithmic program to be ready to “learn” by comparing its actual output with the “taught” outputs to search out errors, and modify the model consequently. Supervised learning thus uses patterns to predict label values on extra unlabeled information. For example, with supervised learning, an algorithm may be fed data with images of sharks labeled as fish and images of oceans labeled as water. By being trained on this data, the supervised learning algorithm should be able to later identify unlabeled shark images as fish and unlabeled ocean images as water.

A common use case of supervised learning is to use historical information to predict statistically probably future events. It's going to use historical stock exchange info to anticipate approaching fluctuations or be used to filter spam

emails. In supervised learning, labeled photos of dogs are often used as input files to classify unlabeled photos of dogs.

Types of supervised learning algorithms include active learning, classification and regression. 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. As an example, for a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email.

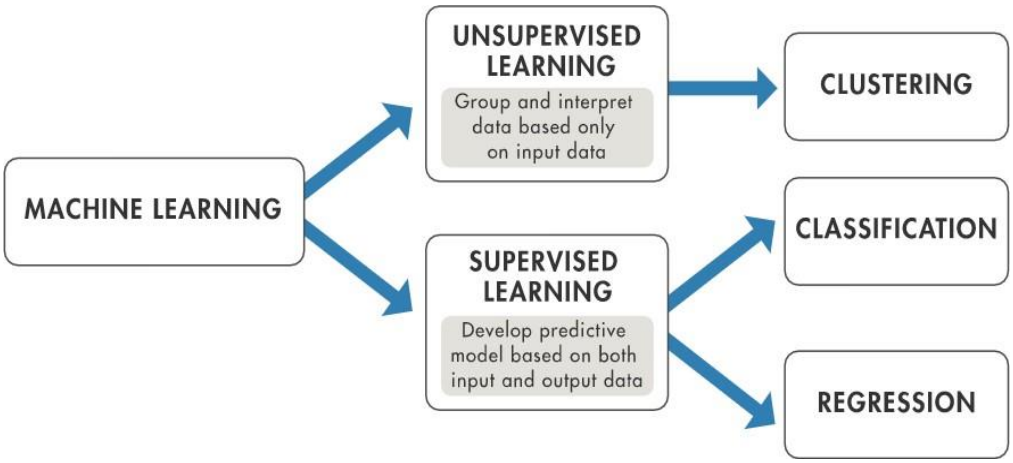
Similarity learning is an area of supervised machine learning closely related to regression and classification, but the goal is to learn from examples using a similarity function that measures how similar or related two objects are. It has applications in ranking, recommendation systems, visual identity tracking, face verification, and speaker verification.

1.3.2 Unsupervised Learning

In unsupervised learning, information is unlabeled, and the learning rule is left to seek out commonalities among its input file. The goal of unattended learning is also as easy as discovering hidden patterns at intervals in a dataset, however, it should even have a goal of feature learning, that permits the procedure machine to mechanically discover the representations that square measure required to classify data.

Unsupervised learning is usually used for transactional information. You will have an oversized dataset of consumers and their purchases, however, as a person, you'll probably not be able to add up what similar attributes will be drawn from client profiles and their styles of purchases.

With this information fed into the Associate in Nursing unattended learning rule, it should be determined that ladies of a definite age vary UN agency obtain unscented soaps square measure probably to be pregnant, and so a promoting campaign associated with physiological condition and baby will be merchandised



1.1 Machine Learning Classification

CLASSIFIERS IN MACHINE LEARNING

1. **Logistic Regression in Machine Learning**

   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 is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems**.**

   Advantages: Logistic regression is easier to implement, interpret, and very efficient to train. It makes no assumptions about distributions of classes in feature space. It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.

Disadvantages: If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting. The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.

2). Support Vector Machine:

Definition: Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Advantages: Effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient.

Disadvantages: The algorithm does not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

Examples of hyperplanes
H1 is not a good hyperplane as it doesn't separate the classes
H2 does but only with small margin
H3 separates them with maximum margin (distance)

Parameters of SVM

There are three main parameters which we could play with when constructing a SVM classifier:

- Type of kernel
- Gamma value
- C value

3). K-NEAREST NEIGHBOUR (KNN):

kNN classfied an object by a majority vote of the object's neighbours, in the space of input parameter. The object is assigned to the class which is most common among its **k (an integer specified by human) nearest neighbour**.

It is a **non-parametric, lazy algorithm**. It's non-parametric since it does not make any assumption on data distribution (the data does not have to be normally distributed). It is lazy since it does not really learn any model and make generalization of the data (It does not train some parameters of some function where input X gives output y).

So strictly speaking, this is not really a learning algorithm. It simply classifies objects based on **feature similarity** (feature = input variables).

Classification is computed from a simple majority vote of the k nearest neighbors of each point.

Advantages: This algorithm is simple to implement, robust to noisy training data, and effective if training data is large.

Disadvantages: Need to determine the value of K and the computation cost is high as it needs to computer the distance of each instance to all the training samples.

4). Random Forest

Random forest is an ensemble model that grows multiple tree and classify objects based on the "votes" of all the trees. i.e. An object is assigned to a class that has most votes from all the trees. By doing so, the problem with high bias (overfitting) could be alleviated.

Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive

accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

Pros of RF:

- It could handle large data set with high dimensionality, output **Importance of Variable**, useful to explore the data
- Could handle missing data while maintaining accuracy

Cons of RF:

- Could be a black box, users have little control on what the model does

Advantages: Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases.

Disadvantages: Slow real time prediction, difficult to implement,& complex algorithm.

CHAPTER 2

LITERATURE SURVEY

2.1 RELATED WORK

- Anika Singh from UEM, Kolkata in 2020 stated that Breast cancer can be predicted not only by using eager learners but also lazy learners. This however fails in grinding the maximum accuracy one can procure through the available algorithms. They stated that using lazy learners to point out just the cell growth can attain the required accuracy which should be around a whopping 88 percentage.
- Nitasha in 2019 stated too that lazy learners can get one the required accuracy by using the required accuracy, they not only trained the lazy learners but also the eager learners too which falls under the same category of around ninety percent.
- Furthermore Navya Sri in 2019 gave away a cross comparitive analysis that has been made between Bayesian Classifiers and Decision tree algorithm that predicts the accuracy of Bayesian with the usage of Waikato Environment for Knowledge Analysis with an outcome of 75.27% and Decision tree as 75.875%.
- A professor from Brown University in 2016 stated that deep belief networks (DBN) combined with Artificial Neural Networks (ANN) can yield a better accuracy pointing out the need of excessive training and testing to be more assured of the derieving outcome.
- Shannon from 2011 used an image scanning algorithm, which can be analytically trained to give a value nearer to 95 percent using his metaplasticity Artificial Neural Network technique which gave away a value of 99.26% which in terms of today is an overfit value. But it provided a major breakthrough in terms of the value reaching above 90 percent and consistently inspired researchers to train and test their data accordingly to raise their accuracy standards above 95 percent.
- Jiaxin Li in 2020 stated that whatever the algorithm is,

training and testing is to be done to acheive the one true accuracy, which acts as one of this article's principle inspirations.

- Nam Nhut Phan in 2021 said that using Convolutional Neural Network has proved to be split into 3 parts, giving equal importance to training (50%) and testing (43%) and the remaining for validation to acheive the results without any correlation or duplicate empty values.
- N Gupta in 2021 usedd an ensemble based training model to acheive the required results. This gave away a result of 96.77 without and gradient boosting algorithms, which is a major breakthrough.
- Md. Milon Islam staded in 2020 that using Artificial Neural Network gave away a result of 96.82 percent and an accuracy of 0.9777 by using Support Vector Machine. They researched particulary the previous breakthrough of using the comparitative analysis of using both the Artificial Neural Networks and Deep Belief Network into just using the ANNs to get an accuracy of above 95% previously researched by the scholar Shannon [2011].
- Chang Ming in 2019 used BCRAT and BOADICEA together with eight simulated datasets with the cancer carriers and their cancer free relatives and found at a shocking result of one of the cancer free patients giving away a positive accuracy of 97% making them prone to cancer relatively.

CHAPTER 3

METHODOLOGY

3.1 EXISTING SYSTEM

The existing model for the customer segmentation depicts that it is based on the K-means clustering algorithm which comes under centroid-based clustering. The suitable K value for the given dataset is selected appropriately which represents the predefined clusters. Raw and unlabeled data is taken as input which is further divided into clusters until the best clusters are found. Centroid based algorithm used in this model is efficient but sensitive to initial conditions and outliers

3.2 PROPOSED SYSTEM

The main proposal of this project is to get the maximum accuracy, that is being vakued at above 95% without using parameter tuning and overfitting. To do so, every

3.3 OBJECTIVE OF PROJECT

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarities among customers in each group. The main objective of segmenting customers is to decide how to relate to customers in each segment to maximize the value of each customer to the business

The emergence of many competitors and entrepreneurs has caused a lot of tension among competing businesses to find new buyers and keep the old ones. As a result of the predecessor, the need for exceptional customer service becomes appropriate regardless of the size of the business.Furthermore, the ability of any business to understand the needs of each of its customers will provide greater customer support in providing targeted customer services and developing customized customer service plans. This understanding is possible through structured customer service.

3.4 SOFTWARE AND HARDWARE REQUIREMENTS

3.4.1 Software Requirements:

- ✓ Python
- ✓ Anaconda
- ✓ Jupyter Notebook

3.4.2 Hardware Requirements:

- ✓ Processor: Intel Core i5
- ✓ RAM: 8GB
- ✓ OS: Windows

3.4.3 **Libraries:**

- ✓ **Tkinter**- Tkinter is a library of python used often by everyone. It is a library that is used to create GUI-based applications easily. It contains so many widgets like radio buttons, text files and so on. We have used this for creating an account registration screen, log in or register screen, prediction interface which is a GUI based application
- ✓ **Sklearn**- Scikit Learn also known as sklearn is an open-source library for python programming used for implementing machine learning algorithms. It features various classification, clustering, regression machine learning algorithms. In this, it is used for importing machine learning models, getting accuracy, get a confusion matrix.
- ✓ **Pandas-** Library of python which can be used easily. It gives speed results and is also easily understandable. It is a library that can be used without any cost. We have used it for data analysis and to read the dataset.
- ✓ **Matplotlib-** Library of python used for visualizing the data using graphs, scatterplots, and so on. Here, we have used it for data visualization.
- ✓ **Numpy-** Library of python used for arrays computation. It has so many

functions. We have used this module to change the 2-dimensional array into a contiguous flattened array by using the ravel function.

- ✓ **Pandas Profiling**-This is a library of python which can be used by anyone free of cost. It is used for data analysis. We have used this for getting the report of the dataset.

3.5. PROGRAMMING LANGUAGES

3.5.1 Python

Python is the best programing language fitted to Machine Learning. In step with studies and surveys, Python is the fifth most significant language yet because the preferred language for machine learning and information science.

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

Features of python

There are many features in Python, some of which are discussed below –

1. Easy to code: Python is a high-level programming language. Python is very easy to learn the language as compared to other languages like C, C#, Java script, Java, etc. It is very easy to code in python language and anybody can learn python basics in a few hours or days. It is also a developer-friendly language.
2. Free and Open Source: Python language is freely available at the official website and can download it easily.

3. Object-Oriented Language: One of the key features of python is Object-Oriented programming. Python supports object-oriented language and concepts of classes, objects encapsulation, etc.

4. High-Level Language: Python is a high-level language. When we write programs in python, we do not need to remember the system architecture, nor do we need to manage the memory.

5. Extensible feature: Python is an Extensible language. We can write us some Python code into C or C++ language and also, we can compile that code in C/C++ language.

6. Python is Portable language: Python language is also a portable language. For example, if we have python code for windows and if we want to run this code on other platforms such as Linux, Unix, and Mac then we do not need to change it, we can run this code on any platform.

7. Python is Integrated language: Python is also an Integrated language because we can easily integrated python with other languages like C, C++, etc.

8. Interpreted Language: Python is an Interpreted Language because Python code is executed line by line at a time. like other languages C, C++, Java, etc. there is no need to compile python code this makes it easier to debug our code. The source code of python is converted into an immediate form called bytecode

9. Large Standard Library: Python has a large standard library which provides a rich set of module and functions so you do not have to write your own code for every single thing. There are many libraries present in python for such as regular expressions, unit-testing, web browsers, etc.

10. Dynamically Typed Language: Python is a dynamically-typed language. That means the type (for example- int, double, long, etc.) for a variable is decided at run time not in advance because of this feature we don't need to specify the type of variable.

Advantages of python

- **Integration Feature:** Python integrates the Enterprise Application Integration that makes it easy to develop Web services by invoking COM or COBRA components. It has powerful control capabilities as it calls directly through C, C++ or Java via Python. Python also processes XML and other mark-up languages as it can run on all modern operating systems through same byte code.

- **Improved Programmer's Productivity:**The language has extensive support libraries and clean object-oriented designs that increase two to tenfold of programmer's productivity while using the languages like Java, VB, Perl, C, C++ and C#.

- **Productivity:** With its strong process integration features, unit testing framework and enhanced control capabilities contribute towards the increased speed for most applications and productivity of applications. It is a great option for building scalable multi-protocol network applications.

Disadvantages of Python

Python has varied advantageous features, and programmers prefer this language to other programming languages because it is easy to learn and code too. However, this language has still not made its place in some computing arenas that includes Enterprise Development Shops. Therefore, this language may not solve some of the enterprise solutions, and limitations include-

- Difficulty in Using Other Languages: The Python lovers become so accustomed to its features and its extensive libraries, so they face problem in learning or working on other programming languages. Python experts may see

the declaring of cast “values” or variable “types”, syntactic requirements of adding curly braces or semi colons as an onerous task.

- Weak in Mobile Computing: Python has made its presence on many desktop and server platforms, but it is seen as a weak language for mobile computing. This is the reason very few mobile applications are built in it like Carbon Nelle.
- Gets Slow in Speed: Python executes with the help of an interpreter instead of the compiler, which causes it to slow down because compilation and execution help it to work normally. On the other hand, it can be seen that it is fast for many web applications too.
- Run-time Errors: The Python language is dynamically typed so it has many design restrictions that are reported by some Python developers. It is even seen that it requires more testing time, and the errors show up when the applications are finally run.
- Underdeveloped Database Access Layers: As compared to the popular technologies like JDBC and ODBC, the Python’s database access layer is found to be bit underdeveloped and primitive. However, it cannot be applied in the enterprises that need smooth interaction of complex legacy data.

3.5.2 Domain

Machine learning could be a subfield of computer science (AI). The goal of machine learning typically is to know the structure information of knowledge of information and match that data into models which will be understood and used by folks. Although machine learning could be a field inside technology, it differs from ancient process approaches. In ancient computing, algorithms are sets of expressly programmed directions employed by computers to calculate or downside solve. Machine learning algorithms instead give computers to coach on knowledge inputs and use applied math analysis to output values that fall inside a particular vary. Thanks to this, machine learning facilitates computers in building models from sample knowledge to modify decision-making processes supported knowledge inputs.

3.6 SYSTEM ARCHITECTURE

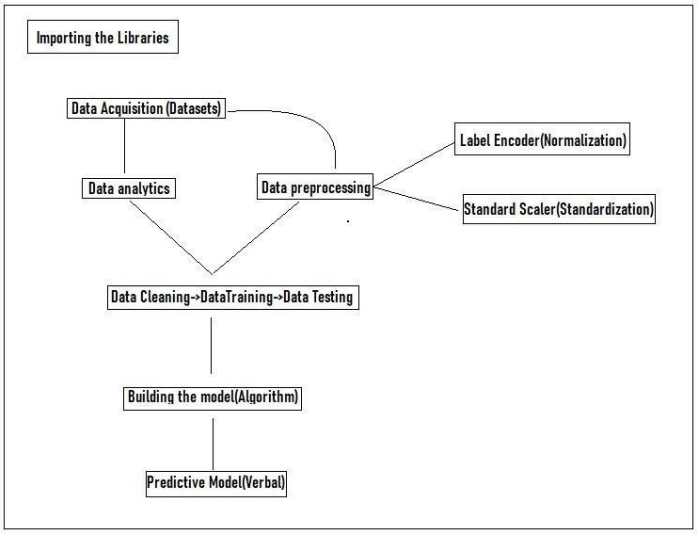


FIG 1: WORKING ARCHITECTURE OF THE MODULE

Data collection

Data used in this project is a set of product reviews collected from credit card transactions records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which I already know the target answer is called labelled data.

Data pre-processing

Pre-processing is the process of three important and common steps as follows:

- Formatting: It is the process of putting the data in a legitimate way that it would be suitable to work with. Format of the data files should be formatted according to the need. Most recommended format is .csv files.
- Cleaning: Data cleaning is a very important procedure in the path of data science as it constitutes the major part of the work. It includes removing missing data and complexity with naming category and so on. For most of the data scientists, Data Cleaning continues of 80% of work.
- Sampling: This is the technique of analysing the subsets from whole large datasets, which could provide a better result and help in understanding the

behaviour and pattern of data in an integrated way

Data visualization

Data Visualization is the method of representing the data in a graphical and pictorial way, data scientists depict a story by the results they derive from analysing and visualizing the data. The best tool used is Tableau which has many features to play around with data and fetch wonderful results.

Feature extraction

Feature extraction is the process of studying the behaviour and pattern of the analysed data and draw the features for further testing and training. Finally, my models are trained using the Classifier algorithm. I used to classify module on Natural Language Toolkit library on Python. I used the labelled dataset gathered. The rest of my labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Random forest. These algorithms are very popular in text classification tasks.

Evaluation model

Evaluation is an essential part of the model development process. It helps to find the best model that represents our data and how well the selected model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can effortlessly generate overoptimistically and over fitted models. To avoid overfitting, evaluation methods such as hold out and cross-validations are used to test to evaluate model performance. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is well-defined as the proportion of precise predictions for the test data. It can be calculated easily by mathematical calculation i.e. dividing the number of correct predictions by the number of total predictions.

3.7 ALGORITHMS USED

3.7.1 Logistic Regression

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**.**

3.7.2 Decision Tree

A decision tree is one of the simplest yet highly effective classification and prediction visual tools used for decision making. It takes a root problem or situation and explores all the possible scenarios related to it on the basis of numerous decisions. Since decision trees are highly resourceful, they play a crucial role in different sectors. From programming to business analysis, decision tree examples are everywhere. If you also want to learn what a decision tree is and how to create one, then you are in the right place. Let's begin and uncover every essential detail about decision tree diagrams.

3.7.3 Random Forest

A random forest (RF) is an ensemble classifier and consisting of many DTs similar to the way a forest is a collection of many trees. DTs that are grown very deep often cause overfitting of the training data, resulting a high variation in classification outcome for a small change in the input data. They are very sensitive to their training data, which makes them error-prone to the test dataset. The different DTs of an RF are trained using the different parts of the training dataset. To classify a new sample, the input vector of that sample is required to pass down with each DT of the forest. Each DT then considers a different part of that input vector and gives a classification outcome. The forest then chooses the classification of having the most 'votes' (for discrete classification outcome) or the average of all trees in the forest (for numeric classification outcome). Since the RF algorithm considers the outcomes from many different DTs, it can reduce the variance resulted from the consideration of a single DT for the same dataset**.**

Steps for Implementation:

- Initialise the classifier to be used.
- Train the classifier: All classifiers in scikit-learn uses a fit(X, y) method to fit the model(training) for the given train data X and train label y.
- Predict the target: Given an non-label observation X, the predict(X) returns the predicted label y.
- Evaluate the classifier model

3.8 MODULES

The project contains three parts:

- **Dataset Collection**- We had collected datasets from Kaggle notebooks. The dataset contains the symptoms and the corresponding disease. It contains 303 rows.
- **Train and test the model**- We had used three classification algorithms named Decision Tree, Logistic regression, and Random Forest to train the dataset. After training, we had tested the model and found the prediction of disease with maximum accuracy.
- **Hyperparameter tuning**-Hyperparameters cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.

Following are the steps to do this project (use Jupyter Notebook):

A) Collect the dataset.

B) Import the necessary libraries.

C) Visualize the dataset.

D) Train the dataset using LR, KNN, RF, SVM.

E) Test the model and find the accuracy score

F) Based on the scores predict which algorithm is best for prediction.

G) Build a deployment model using Azure, AWS or Heroku

H) Enter the values and predict the accuracy.

CHAPTER 4

RESULTS AND DISCUSSION

4.1. PERFORMANCE ANALYSIS

Coming to the performance it works in a time rate of 1 second per statement and code implied. Duplicated and similar lookalike data's can be removed efficiently too. The performance of a predictive model is calculated and compared by choosing the right metrics. So, it is very crucial to choose the right metrics for a particular predictive model in order to get an accurate outcome. It is very important to evaluate proper predictive models because various kinds of data sets are going to be used for the same predictive model.

Algorithm	Precision	Recall	F-measure	Accuracy
KNN	0.845	0.823	0.835	89.62%
Logistic Regression	0.857	0.882	0.869	92.25%

Random Forest	0.867	0.882	0.909	86.16%
SVM	0.837	0.911	0.873	88.25%

Fig 3: Accuracies Obtained

CHAPTER 5

CONCLUSION AND FUTURE ENHANCEMENTS

Nothing should go noticed. Symptoms should be checked upon before the arrival of the unnoticed demon. Prevention is, was and will always be better than the cure. Cancer are the most brutal thing a person will be experiencing in their lifetime, but if found beforehand. It can be handled and the respective person can see through their remission.

In this paper, we have researched the possible outcome of almost every Machine Learning algorithm and came to a discussion that whatever be the algorithm, a clear cut need of pre-processing, training and testing is needed to achieve the maximum accuracy in not just this Breast Cancer Module, but every module.

Using this bit of a code, one can easily detect the possibility of whether a person has Breast Cancer or not and can enquire the hospitals about further actions to be taken.

The subsequent results show us that by the usage of graphical representation and attribute filtering in successive levels increased the accuracy to almost a whopping 6% in our case.

The highest accuracy obtained here was almost 97% which has been achiever by using Random Forest Algorithm. Due to the proper cleaning mechanism, almost

every algorithm can reach up to a minimum of a 90 percent value and out of this Random Forest stands out.

Physical diagnosis has become a very well waged business nowadays. Even a slightest help from a machine can help one save heap loads of money for someone in any corner of the world. By this way Machine Learning provided a significant breakthrough not only in medical field but every other field too. Random Forest not only gives the perfect result but it stands out and stays stable throughout the code making it relevant to make it possible to use it for every other code too.

REFERENCES

[1] Shannon Doherty, Breast cancer analysis using lazy 2011learners https://www.webmd.com/breast-cancer/features/ shannen-doherty-breast-cancer

[2] M Navya Sri, ANIT, Analaysis of NNC and SVM for Machine Learning 2020

[3] N Gupta, Google Scholar, Prediction of Areolar cancer

[4] Jiaxin Li, Jilin University, 5year survival forpersonhaving-breast-cancer(2020).

[5] Mohammad Milan Islam, University of Waterloo, Prediction of residual diseases and breast cancer.2020 https:// link.springer.com/article/10.1007/s42979-020-00305-

[6] National Cancer Institute. Inflammatory breast cancer. http://www.cancer.gov/types/breast/ibc-fact-sheet, 2016.

[7] Chang Ming, BCRAT and BOADICEA comparison. Peking University,2019,Presonalized breast cancer risk prediction https://link.springer.com/article/10.1007/ s42979-020-00305-w

[8] Rouse HC, Ussher S, Kavanagh AM, Cawson JN. Examining invasive biopsy of ultrasound mammogram in breast cancer 2019.

[9] Nitasha, Punjab Technical Univerisity, 2019, Review on Prediction of breast cancer using data mining http://www.ijcstjournal.org/volume-7/issue-4/IJCST-V7I4P8.pdf

[10] Rucha Kanade, Xavier School of Engineering 2019, Breast cacner prediction using gradient boosters.

APPENDICES

A. SOURCE CODE

```
import numpy

import matplotlib.pyplot as plt

import pandas as pd

import seaborn as sns

df=pd.read_csv("data.csv")

df.head()

df.info()

df.isna().sum()

df.shape

df=df.dropna(axis=1)
```

```
df.shape

df.describe()

df['diagnosis'].value_counts()

sns.countplot(df['diagnosis'])

from sklearn.preprocessing import LabelEncoder

labelencoder_Y = LabelEncoder()

df.iloc[:,1]=labelencoder_Y.fit_transform(df.iloc[:,1].values)

df.iloc[:,1:32].corr()

plt.figure(figsize=(10,10))

sns.heatmap(df.iloc[:,1:10].corr(),annot=True,fmt=".0%")

X=df.iloc[:,2:31].values

Y=df.iloc[:,1].values

from sklearn.model_selection import train_test_split

X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.20,random_state=0)

from sklearn.preprocessing import StandardScaler

X_train=StandardScaler().fit_transform(X_train)

X_test=StandardScaler().fit_transform(X_test)

def models(X_train,Y_train):

    from sklearn.linear_model import LogisticRegression
```

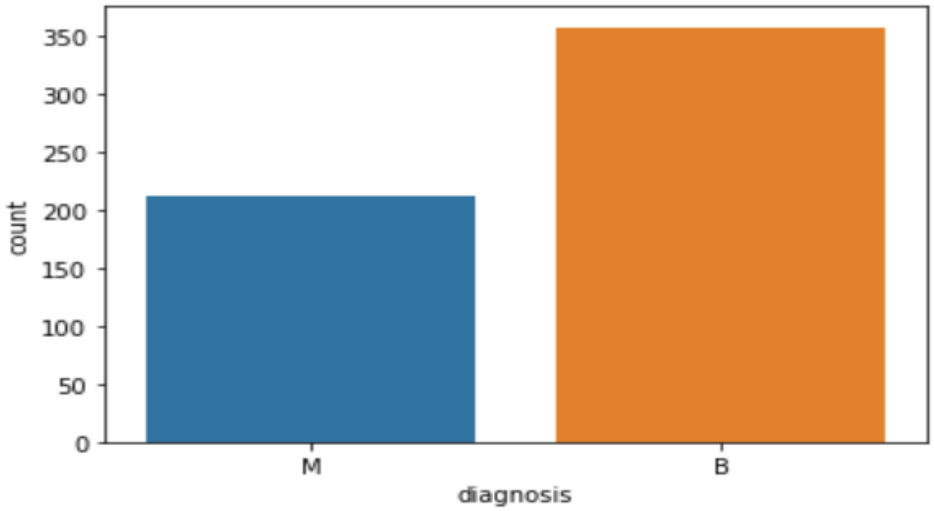
```
    log=LogisticRegression(random_state=0)
    log.fit(X_train,Y_train)



    from sklearn.tree import DecisionTreeClassifier
    tree=DecisionTreeClassifier(random_state=0,criterion='entropy')
    tree.fit(X_train,Y_train)



    from sklearn.ensemble import RandomForestClassifier

forest=RandomForestClassifier(random_state=0,criterion="entropy",n_estimators=10
)
    forest.fit(X_train,Y_train)


    print('[0]logistic regression accuracy:',log.score(X_train,Y_train))
    print('[1]Decision tree accuracy:',tree.score(X_train,Y_train))
    print('[2]Random forest accuracy:',forest.score(X_train,Y_train))


    return log,tree,forest
```

```
model=models(X_train,Y_train)

from sklearn.metrics import accuracy_score

from sklearn.metrics import classification_report

for i in range(len(model)):

    print("Model",i)

    print(classification_report(Y_test,model[i].predict(X_test)))

    print('Accuracy : ',accuracy_score(Y_test,model[i].predict(X_test)))
```

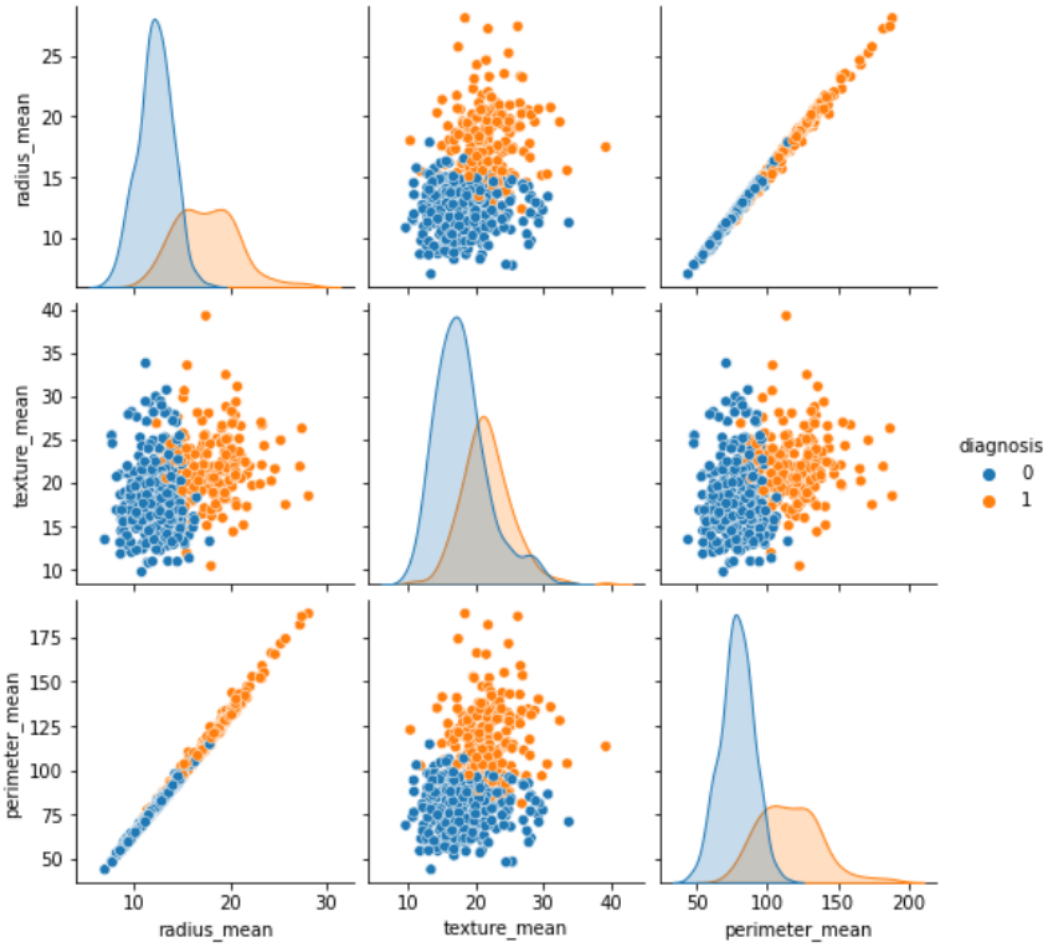
B. SCREENSHOTS

	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension	target
.	29.41	179.10	1819.0	0.14070	0.41860	0.6599	0.2542	0.2929	0.09873	0.0
.	26.40	166.10	2027.0	0.14100	0.21130	0.4107	0.2216	0.2060	0.07115	0.0
.	38.25	155.00	1731.0	0.11660	0.19220	0.3215	0.1628	0.2572	0.06637	0.0
.	34.12	126.70	1124.0	0.11390	0.30940	0.3403	0.1418	0.2218	0.07820	0.0
.	39.42	184.60	1821.0	0.16500	0.86810	0.9387	0.2650	0.4087	0.12400	0.0
.	30.37	59.16	268.6	0.08996	0.06444	0.0000	0.0000	0.2871	0.07039	1.0

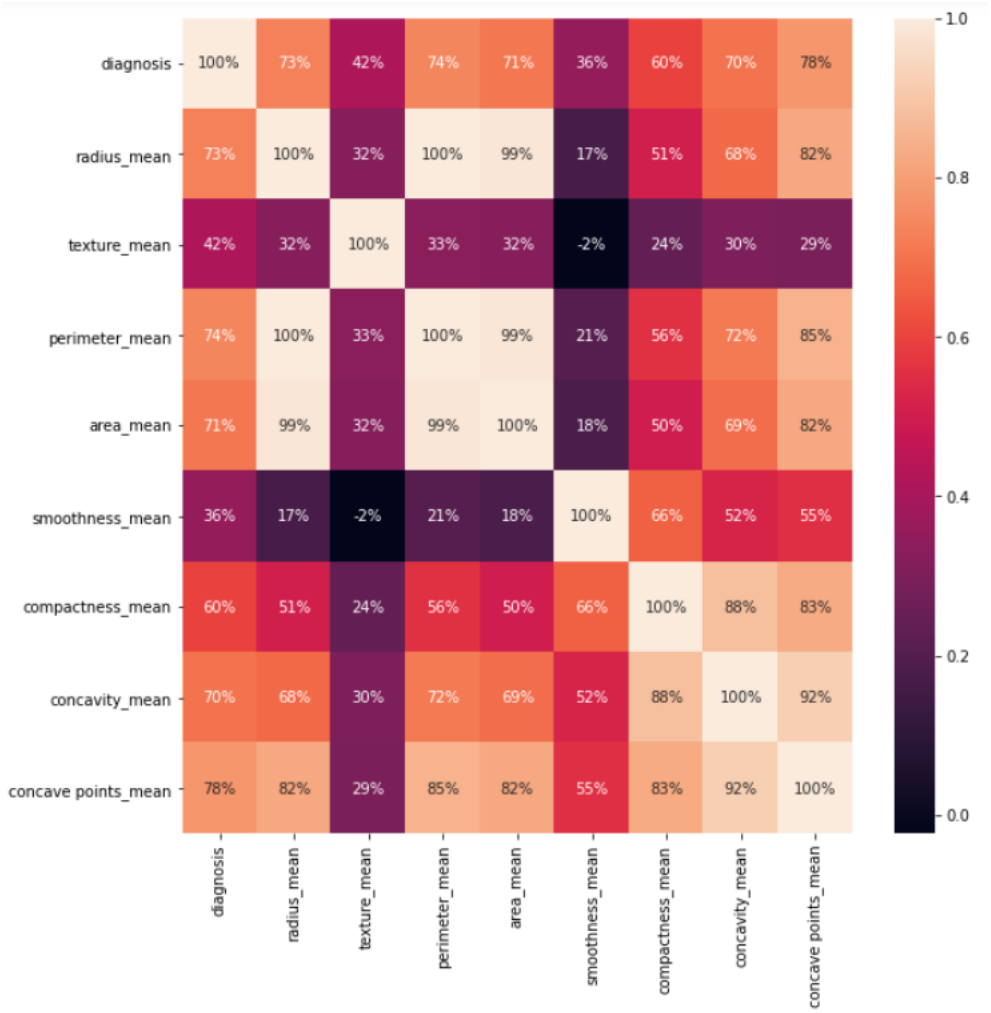
B-1: DATASET



B-2: COUNTPLOT



B-3: PAIRPLOT



B-4: HEAT MAP

```
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report

for i in range(len(model)):
    print("Model",i)
    print(classification_report(Y_test,model[i].predict(X_test)))
    print('Accuracy : ',accuracy_score(Y_test,model[i].predict(X_test)))
```

B-5: REPORT GENERATION

```
Model 0
              precision    recall  f1-score   support

           0       0.96      0.99      0.97        67
           1       0.98      0.94      0.96        47

    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114

Accuracy :  0.9649122807017544
```

```
Model 1
              precision    recall  f1-score   support

           0       0.94      0.96      0.95        67
           1       0.93      0.91      0.92        47

    accuracy                           0.94       114
   macro avg       0.94      0.94      0.94       114
weighted avg       0.94      0.94      0.94       114

Accuracy :  0.9385964912280702
```

```
Model 2
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        67
           1       1.00      0.94      0.97        47

    accuracy                           0.97       114
   macro avg       0.98      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114

Accuracy :  0.9736842105263158
```

B-6: DETAILED COMPARISION OF THE ACCURACIES