AN EFFICIENT PREDICTION FOR BREAST CANCER USING DATA MINING TECHNIQUES

A PROJECT REPORT

Submitted by

PREETHA . V (810015104067)

VINOTHA . K (810015104101)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

UNIVERSITY COLLEGE OF ENGINEERING – BIT CAMPUS,

TIRUCHIRAPPALLI

ANNA UNIVERSITY::CHENNAI 600 025

APRIL 2019

UNIVERSITY COLLEGE OF ENGINEERING,

BIT CAMPUS,

TIRUCHIRAPPALLI-620 024

BONAFIDE CERTIFICATE

Certified that this project report "**An Efficient Prediction For Breast Cancer Using Data Mining Techniques"** is the bonafide work of **"Ms. V.PREETHA (810015104067)** and **Ms. K. VINOTHA (810015104101)"** who carried out the project work under my supervision.

SIGNATURE

Mr. D. Venkatesan

HEAD OF THE DEPARTMENT

Assistant Professor

Computer Science &Engineering

University College of Engineering,

Anna University-BIT Campus,

Tiruchirappalli-620 024

SIGNATURE

Mr. L. Vanitha

SUPERVISOR

Teaching Fellow

Computer Science &Engineering

University College of Engineering,

Anna University-BIT Campus,

Tiruchirappalli-620 024

Submitted for the project Viva voce examination held on ……

Internal Examiner **External Examiner**

DECLARATION

We hereby declare the work entitled **"AN EFFICIENT PREDICTION FOR BREAST CANCER USING DATA MINING TECHNIQUES"** is submitted in partial fulfillment of the requirement for the award of the degree in B.E., Computer Science and Engineering, University College of Engineering(BIT Campus), Tiruchirappalli, is a record of our own work carried out by us during the academic year 2018-2019 under the supervision and guidance of Mrs. L. Vanitha, Teaching Fellow, Department of Computer Science and Engineering, University College of Engineering(BIT Campus), Tiruchirappalli. The extent and source of information are derived from the existing literature and have been indicated through the dissertation at the appropriate places. The matter embodied in this work is original and has not been submitted for the award of any degree, either in this or any other University.

SIGNATURE OF THE CANDIDATES

V. PREETHA (810015104067)

K. VINOTHA (810015104101)

I certify that the declaration made above by the candidate is true.

SIGNATURE OF THE GUIDE

Mr. L. VANITHA

Teaching Fellow,

Department of CSE,

University College of Engineering,

BIT Campus, Anna University,

Tiruchirappalli-620 024.

ACKNOWLEDGEMENT

I would like to convey my heartfelt thanks to our honorable Dean **Dr. T. SENTHILKUMAR,** Associate Professor for having provided me with all required facilities to complete my project without hurdles.

I would like to express my sincere thanks and deep sense of gratitude to guide **Mr. D. VENKATESAN,** Assistant Professor and Head, Department of Computer Science and Engineering, for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of this project work.

I would like to thank my project guide **Mrs. L. VANITHA,** Teaching Fellow, Department of Computer Science and Engineering, for his valuable guidance throughout the phase of the project. This is our responsibility to thank our project coordinator **Mr. C. SANKAR RAM,** Assistant Professor and **Mr. P. KARTHIKEYAN,** Assistant Professor, Department of Computer science and Engineering for his constant inspiration that he has all through the project period.

I would like to thank **Mr. C. SURESH KUMAR,** Teaching Fellow, Department of Computer Science and Engineering, for his encouragement for this work.

I extend my thanks to all other teaching and non-teaching staffs for their encouragement and support.

I thank my beloved parents and friends, for their full support in my career development of this project.

TABLE OF CONTENT

ABSTRACT

Breast cancer is a major threat for middle aged women throughout the world and currently this is the second most threatening cause of cancer death in women. Breast cancer has become the leading cause of death for women in the world. Every woman is at the risk of breast cancer. Breast cancer is an uncontrolled growth of breast cells. Recurrence of cancer is one of the biggest fears in the life of a cancer patient and thus one the issues that affect their quality of life. But early detection and prevention can significantly reduce the chances of death. Data mining algorithms can provide great assistance in prediction of early stage breast cancer. To compare the accuracy of few data mining algorithms in predicting breast cancer recurrence. This dataset contained total 35 attributes in which we applied Logistic regression, KNN, Random forest tree, Decision tree, Multilayer perceptron and Support Vector Machine (SVM) classification algorithms and calculated their prediction accuracy. An efficient feature selection algorithm helped us to improve the accuracy of each model by reducing some lower ranked attributes.

LIST OF FIGURES

LIST OF ABBREVATIONS

BC	Breast Cancer
KNN	k-nearest neighbor
SVM	Support vector machine
NB	Naïve Bayes
MLP	Multilayer perceptron
RFT	Random Forest Tree
DC	Decision Classifier
LR	Logistic Regression
RC	Recurrence
NRC	Non Recurrence
CSV	Comma Separated Values

CHAPTER 1

INTRODUCTION

1.1OVERVIEW OF DATA MINING

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data pattern analysis, typically deals with data that have already been collected for some purpose rather than the data mining analysis. This means that the objectives of data mining exercise play no role in the data collection strategy. The data sets examined in data mining are often large.

1.2 DATA MINING TERMINOLOGIES

- **Data:** Data are any facts, numbers, or text that can be processed by a computer.
- **Information:** The patterns, associations, or relationships among all this data can provide information.
- **Knowledge:** Information can be converted into knowledge about historical patterns and future trends.
- **Data Warehouses:** Data Warehouse is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.

- **Association Analysis:** Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data.
- **Data Mining:** It is the extraction of hidden predictive information from large databases. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

Data Mining Types

- **Predictive data mining:** It produces the model of the system described by the given data. It uses some variables or fields in the data set to predict unknown or future values of other variables of interest.
- **Descriptive Data Mining:** It produces new, non-trivial information based on the available data set. It focuses on finding patterns describing the data that can be interpreted by humans.

Data Mining Tasks

- Data processing [descriptive]
- Prediction [predictive]
- Regression [predictive]
- Clustering [descriptive]
- Classification [predictive]
- Link analysis/ associations [descriptive]
- Evolution and deviation analysis [predictive]

1.3 KNOWLEDGE DISCOVERY

Data mining has attracted a great attention in the information industry and in society as a whole in recent years, due to wide availability of huge amount of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for application ranging from market analysis, fraud detection, to production control, disaster management and science exploration. Data mining can be viewed as a result of the natural evolution of information technology. The database system industry has witnessed an evolutionary path in the development of various functionalities: data collection and database creation, database management (including data storage and retrieval, and database transaction processing and data analysis.

Knowledge discovery as a process consists of an iterative sequence of following steps:

- Data cleaning, that is, to remove noise and inconsistent data.
- Data integration, that is, where multiple data sources are combined.
- Data selection, that is, where data relevant to the analysis task are retrieved from the database.
- Data transformation, that is, where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- Data mining, that is, an essential process where intelligent methods are applied in order to extract the data patterns.
- Knowledge presentation, that is, where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

Three steps involved are:

1. Exploration
2. Pattern identification
3. Deployment

- **Exploration:** In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.
- **Pattern Identification:** Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.
- **Deployment:** Patterns are deployed for desired outcome.

Classification

Discovery of a predictive learning function that classifies a data item into One of several predefined classes. Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm.

In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models

- Classification by decision tree induction
- Bayesian Classification
- Back propagation
- Support Vector Machines (SVM)
- Classification Based on Associations
- K-Nearest Neighbor Classifies
- Case Based Reasoning
- Genetic Algorithm
- Rough Set Approach
- Fuzzy Set Approach

Clustering

A common descriptive task in which one seeks to identify a finite set of categories or cluster to describe the data. Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it

becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Types of clustering methods

- Hierarchical (divisive) Methods
- Partitioning Methods
- Density Based Methods
- Grid-Based Methods
- Model Based Algorithms

Categorization of Clustering Algorithms

Algorithms are key step for solving the techniques. In these clustering techniques, various algorithms are currently in the life, still lot more are evolving. But in general, the algorithm for clustering is neither straight nor canonical:

Hierarchical methods

- Agglomerative Algorithms
- Divisive Algorithms

Partitioning methods

- Relocation Algorithms
- Probabilistic Clustering
- K-Medoids Methods
- K-Means Methods

Density-based algorithms

- Density-based connectivity clustering
- Density functions clustering

Grid-based methods

- Methods based on co-occurrence of categorical data
- Constraint-based clustering
- Clustering algorithms used in machine learning
- Gradient descent and artificial neural networks
- Evolutionary methods:

Model Based Algorithms

- Algorithms for high dimensional data
- Subspace Clustering
- Projection Techniques
- Co-Clustering Techniques

Regression

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets)

may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

Types of Regression Methods

- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression.

1.4 DATA MINING APPLICATIONS

In Banking Industry data mining is used:

- Predicting Credit fraud
- Evaluation Risk
- Performing trend analysis
- Analyzing profitability
- Helping with direct marketing campaigns

In financial markets and neural networks data mining is used

- Forecasting stock prices
- Forecasting commodity-price prediction
- Forecasting financial disasters
- Forecasting offer prices
- Forecasting customer satisfaction

Data Mining for Telecommunications Industry used

- How does one retain customers and keep them loyal as competitors offer special offers and reduced rates?
- When is a high-risk investment, such as new fiber optic lines, acceptable?
- How does one predict whether customers will buy additional products like cellular services, call waiting, or basic services?
- What characteristics differentiate our products from those of our competitor?

1.5 ROLE OF DATAMINING IN MEDICAL

Breast Cancer (BC) is a major cause of concern worldwide. According to the latest statistics by GLOBOCAN, it was the second most frequently diagnosed cancer and the fifth cause of cancer mortality worldwide, responsible for 6.4% of all deaths. Among women, it is associated to the highest number of deaths due to cancer, with 521907 registered deaths in 2012. Though predominantly in women, BC can also occur in men. However, male BC is rare: it represents less than 1% of all cases .Further references to BC will pertain to female BC except where noted, since it is what this work will focus in.

Portugal follows these global trends, with BC being among the top three most frequently diagnosed cancers. Particularly for women, it was the cancer with highest rates of incidence and mortality. Solely in 2012, 6088 women were diagnosed with this disease, and 1570 died, which confirm the alarming scenario in Portugal. According to WHO (World Health Organization) projections, these numbers are expected to rise, with 1620 deaths by BC predicted for 2015. In diseases with high mortality rates, such as this one, survival prediction assumes an important role, since it aids clinicians to better define each patient's prognosis and the corresponding treatments to be attempted. In particular for BC, prognosis is

related to the patterns of recurrence. Cancer Recurrence (or Relapse) describes cancer that reappears after treatment, and in the specific case of BC, recurrence is very common, being experienced by about one third of patients after initial diagnosis. Therefore, establishing the patterns of recurrence is a crucial task to accurately predict the clinical behavior of this pathology. This enables a more personalized treatment for the patients, avoiding undesired overtreatment and adverse complications.

Despite the considerable advances in the study of BC in the last couple of decades, the underlying processes of recurrence have not yet been completely understood. Encompassed in this reality, this work conducts a data-driven research, attempting to construct a model of recurrence for patients with this condition. As detailed in the following section, our goal is to study the prognostic factors that define female BC recurrence, clarify the correlation between such factors and relapse patterns, and lastly, to provide a model to predict recurrence for a particular patient, based on her personal characteristics as well as her tumor expression.

1.6 OBJECTIVES

Our project aims to construct a model of metastatic BC. This is achieved by examining the behavior of BC relapses, in terms of the localization of the tumor and its other features. The primary goals of this work are the following:

1. Evaluate the pattern of metastatic dissemination in patients with BC

The first objective is to understand how the relapses are physically distributed, and their respective characteristics. BC prognosis is related to recurrence, but it even differs according to the site affected, namely bone only, visceral non hepatic, and visceral hepatic. Therefore, it is important to assess the behavior of BC recurrence metastases.

2. Establish the relation between the patterns of metastatic proliferation, patient's characteristics and BC subtypes

After analyzing the metastatic spread of BC, it will be measured the correlation between these data and the characteristics of the patients and their tumors. BC subtypes are defined via Immunohistochemistry (IHC) studies, used to determine the tumor features. The purpose of this goal is to determine how these characteristics aff ect the patterns of BC recurrence.

3. Build a model of BC recurrence

This work intends to define a recurrence pattern, based on the characteristics of both patients and tumors. To achieve this, we must construct a structure that generalizes the relations found in the previous goal. The fact that it is based on a real-world dataset means that this model may be able to support the decision-making process of clinicians, establishing more accurate predictions, following the paradigm of Personalized Medicine.

CHAPTER 2

LITERATURE SURVEY

[1]TITLE: Using three machine learning techniques for predicting breast cancer recurrence

AUTHOR: Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR

The above authors were discussed about using three machine learning techniques for predicting breast cancer recurrence. In this exploration were applied classification algorithms. They have explored risk factors for predicting breast cancer by using data mining techniques. Each method has its own limitations and strengths specific to the type of application. The number of recurrence and non-recurrence cases were been 117 and 430, respectively. In order to evaluate the validity of present results for making predictions regarding new data, 10-fold cross-validation was implemented in model building, evaluation, and comparison. It shows that SVM outperforms both Decision Tree and MLP in all the parameters of sensitivity, specificity and accuracy. SVM is the best predictor of breast cancer recurrence. Decision tree C4.5 outperformed decision tree C4.5 and ANNs. It shows the accuracy, sensitivity, and specificity comparison of decision tree C4.5, SVM and ANNs.

METHODOLOGY

1. C4.5 Decision tree
2. Artificial neural network and
3. Support vector machine

[2]TITLE: Predicting Breast Cancer Recurrence using effective Classification and Feature Selection techniques.

AUTHOR: Ahmed Iqbal Pritom ,Md. Ahadur Rahman Munshi ,Shahed Anzarus Sabab Shihabuzzaman Shihab

Research paper done by Ahmed Iqbal Pritom,shahed anzarus sababa, Md. Ahadur Rahman Munshi, Shihabuzzaman Shihab predicts whether the breast cancer is recurrent or not. They have used data sets from Wisconsin data sets of the UCI machine learning repository that have 35 attributes. After implementation of algorithms like C4.5 Decision Tree and Support Vector Machine (SVM) classification algorithm was implemented. Using proper attribute selection technique, any classification algorithm can be improved significantly. Attributes with less contribution in dataset often misguides the classification and results in poor prediction. In our work, we found Support Vector Machine giving much better output both before and after attribute selection. Area under ROC curve analysis showed results in our favor where Decision Tree showed much better improvement after feature selection method. We can consider the accuracy of J48 (AUC=0.745) 'Fair' and SMO (AUC = '0.529') should be considered as 'Poor'.

METHODLOGY:

1. **C4.5**
2. **Support Vector Machine**

[3]TITLE: Ensemble learning method for the prediction of breast cancer recurrence

AUTHOR: Daad Abdullah Almuhaidib, Hadil Ahmed Shaiba, Fatima MotebAlbusayyis, MashaelAbdulalimAlzaid, Najla Ghazi Alharbi, Sara Muhammad Alotaibi and Reem Mohammed Almadhi

The above authors were described whether the breast cancer is recurrent or not. They used the online WPBC dataset that is publicly available in the UC Irvine Machine Learning Repository. The dataset has been collected from the regular sample of the breast cancer tissue.The features in the dataset characterizes cell nucleus properties and were generated from image analysis of fine needle aspirates (FNA) of breast masses. The dataset has 33 features and consists of 198 instances with 148 non recurrence and 46 recurrence. It has shown that the best performing models are SVM with accuracy 0.6522, sensitivity 0.6250 and specificity 0.6593, then decision tree with accuracy 0.6261, sensitivity 0.63636 and specificity 0.62500, finally naïve Bayes with accuracy 0.5913, sensitivity 0.4889 and specificity 0.6571.The accuracy rate for SVM is 65%, for decision tree is 62% and for Naive Bayes is 59%.

METHODOLOGY

1. SVM
2. Naive Bayes
3. Decision tree

[4]TITLE: A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques

AUTHOR: Madhuri Gupta, Bharat Gupta.

The authors were discussed about A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning techniques. In this exploration were applied classification algorithms. In research work conducted, 10 fold cross validation is using to validate the classification model. The whole data is divided in 10 equal parts, where 9 parts are applied for training and 1 part for validation as testing the model. The process is repeated 10 times, with every time each of the 10 sub-samples uses at least once. The performance in terms of accuracy, MLP is better as compared to other techniques. MLP technique also performs better than other techniques when Cross Validation metrics is used in breast cancer prediction. In order to further improve accuracy.

METHODOLOGY

1. Multilayer perceptron (MLP),
2. Decision Tree (C4.5),
3. Support vector machine (SVM) and
4. K-nearest neighbor (KNN).

[5]TITLE:Prediction of Recurrence Cancer using J48 Algorithm.

AUTHOR: Dr Prof. NeerajBhargava, Sakshi Sharma, RenukaPurohit, Pramod Singh Rathore.

The authors were described about the Prediction of Recurrence Cancer using J48 Algorithm. They described the use of j48 data mining algorithm to predict the recurrence of cancer based on the dataset of breast cancer. If cancer comes back

after treatment, whether in the same place it first started, or in any other part of the body is called recurrent cancer. They use J48 data mining algorithm on the data set of breast cancer. J48 is a java implementation for C4.5 algorithm provided by Weka tool. Here, analyze the decision tree generated by the algorithm using 10-fold cross validation method to predict the recurrent events, based on the attributes such as node-caps, degree of malignancy, age, tumor-size, menopause, irradiate etc. Number of instances in the dataset are 286. They discussed about a new approach for the prediction of recurrence cancer using medical records of the patients, by applying J48 classification algorithm of data mining. J48 gives result in the form of decision tree.

METHODOLOGY

1. J48 Decision Tree

[6]TITLE: Clustering-based approach for detecting breast cancer recurrence

AUTHOR: Smaranda, Abdel Badeeh Salem, Florin Gorunescu,

Marina Gorunescu

The authors were discussed about Clustering-based approach for detecting breast cancer recurrence using data mining techniques. The exploration was applied by clustering algorithms. Aims to assess the effectiveness of three different clustering algorithms, used to detect breast cancer recurrent events. They applied the three clustering methods to the Wisconsin Breast Cancer database, consisting of 198 cases with two decision classes: non-recurrent-events 151 (76.26%) instances and recurrent-events 47 (23.73%) instances, with 34 input features (attributes). The 10-fold cross-validation has been used as a testing method. The classification performance training/testing - is computed 10 times, each time

leaving out one of the sub-samples and using that sub-sample as a test sample for cross-validation; therefore, each sub-sample is used 9 times as training sample and just once as testing sample. The performance of a classical k-means algorithm is compared with a much more sophisticated Self Organizing Map (SOM-Kohonen network) and a cluster network, closely related to both k-means and SOM. The best performance was obtained by SOM and k-means, their predicting accuracy ranging from 62% to 78%.

METHODOLOGY

1. K-means
2. Self Organizing Map (SOM-Kohonen network)
3. cluster network (Learned Vector Quantization (LVQ) method)

[7]TITLE: A Study on Prediction of Breast Cancer Recurrence using Data mining techniques.

AUTHOR: Uma Ojha, Dr. Savita Goel

The authors were discussed about the study on the prediction of breast cancer recurrence using data mining techniques. The exploration was applied by both clustering and classification algorithms. The result shows that on comparison, classification algorithms are better predictors than clustering algorithms. The classification algorithms were 0.7154 accurate as compared to the accuracy of 0.5257 of clustering algorithms. The performance of these algorithms is measured based on the accuracy, sensitivity and specificity. Probability of accuracy in results is measured in the range of 0 to 1 whereas 1 means 100% accuracy. The complete dataset of WPBC is divided into the ratio of 70:30 in classification algorithms. The 70% of data is used for training purposes and 30% of the dataset is used for testing

purposes. The two classification algorithms C5.0 and SVM achieve 81% accuracy, which is better than all algorithms.

METHODOLOGY

CLASSIFICATION

1. SVM
2. C5.0

CLUSTERING

1. FUZZ LOGIC
2. K-MEANS

[8]TITLE: Association Rule Mining Based Predicting Breast Cancer Recurrence on SEER Breast Cancer Data

AUTHOR: Umesh, Dr. B Ramachandra

The authors were discussed about association rule mining based predicting breast cancer recurrence on seer breast cancer data .The exploration was applied by association rule mining. This dataset contained population characteristics and included 17 input variables. They were gathered the data from the irregular sample of SEER breast cancer dataset. It can be seen from the confusion matrix, that 263 of 2143 records are characterized vaguely. 107 of the "RECURRENCE" cases have been classified as "NONRECURRENCE" (False Negatives). 156 of the "NONRECURRENCE" cases have been classified as "RECURRENCE" (False Positives). It demonstrates the examination of execution as for sensitivity, specificity and accuracy. 138 of the "RECURRENCE" cases have been classified

as “RECURRENCE” (True Positives).1742 of the “NONRECURRENCE” cases have been classified as “RECURRENCE” (True Negatives).

METHODOLOGY

1. **Association rule mining**

[9]TITLE: A Gene Signature for Breast Cancer Prognosis Using Support Vector Machine

AUTHOR: XiaoyiXu, Ya Zhang, Liang Zou, Minghui Wang, Ao Li

The authors were discussed about A Gene Signature for Breast Cancer Prognosis Using Support Vector Machine. A 70-gene signature had been discovered for breast cancer prognosis prediction and received a good performance.70-gene signature had been discovered for breast cancer prognosis prediction and received a good performance. Using the leave-one-out evaluation procedure on a gene expression dataset including 295 breast cancer patients, we discovered a 50-gene signature that by combing with SVM, achieved a superior prediction performance with 34%, 48% and 3% improvement in Accuracy, Sensitivity and Specificity, compared with the widely used 70-gene signature. The dataset of a series of 295 patients with primary breast cancer was adopted in this study. According to their clinical information, these patients were classified into two groups: metastatic disease group (good prognosis) with 208 patients and disease free group (poor prognosis) with 87 patients. When gene number decreased from 300 to 180, the accuracy increased from 73.32% to 92.88%.

METHODOLOGY

1. SVM

CHAPTER 3

SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

In Existing system, the authors used with feature selection and without feature selection. The first phase is without feature selection. In this phase, we constructed three prediction models using the training dataset for the naive Bayes, REP Tree, and K-nearest neighbor classifiers. Ten-fold cross-validation was used as the test option for the models. The models were constructed by using all the attributes contained in the WBC prognostic dataset without any process of feature selection. The results of the experiment show the accuracy, specificity, and sensitivity rate of the prediction models for the naive Bayes classifier, REP Tree classifier, and K-nearest neighbor classifier. The experiment results indicated that the accuracy percentage of the breast cancer recurrence prediction model that used REP Tree as the classifier was higher compared to the naive Bayes and IBK classifiers. REP Tree gave 76.3%, Naive Bayes gave 70% and IBK gave 66.3% accuracies. The dataset used in this experiment did not apply PSO feature selection. The next phase is with feature selection. In this phase, two stages were involved. The first stage was to perform feature selection based on the requirements. The second stage was to evaluate the performance of the models based on the selected features. The process of feature selection was executed by using the WEKA tool, which is the PSO Search that explores the attribute space using the particle swarm optimization (PSO) algorithm. The outcome of this phase was obtaining the best-fit features selected from the dataset. The results showed that out of 34 attributes, four attributes were found to be the best fit features - Cell Nucleus 10, Cell Nucleus 21, Cell Nucleus 31, and Time. Cell Nucleus was

computed based on ten real-valued features, namely, the radius (mean of distances from the center to points on the perimeter), texture (standard deviation of gray scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter∧2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension (‘‘coastline approximation’’ - 1).

The results show that when using PSO feature selection, all the classifiers out-performed their counterparts without feature selection in terms of the accuracy level. The best result of the accuracy level for breast cancer recurrence was exhibited by the naive Bayes classifier, which obtained 81.3% accuracy, followed by REP Tree, with 80% accuracy, and K-nearest neighbor (IBK) reaching the 75.0% accuracy level. However, in the experiment without PSO feature selection, REP Tree classifier provided the highest accuracy level of 76.3%, followed by naive Bayes at 70% and IBK at 66.3%

3.1.1 LIMITATIONS

- Based on this existing system, above algorithms are producing minimum accuracy rate.
- Low in Performance.
- Producing less specificity.
- Producing less sensitivity.

3.2 PROPOSED SYSTEM

Breast cancer starts to grow in the human body when cells in the breast are growing most in an unexpected manner. After these cells grow, it can be seen by x- ray. Basically, there are two types of breast cancer, cancer that spread into another area and cancer that can't spread into another area. Among the world women breast cancer is the first and the most leading of death of women and the accurate diagnosis have lots of advantage to prevent and detection of the disease. Data mining is a technique can support doctors in the decision making process. As breast cancer recurrence is high, good diagnosis is important. This research is going to be implemented by different data mining algorithms like logistic regression, MLP, Decision Tree, KNN, Support vector machine, and Random Forest So to get a more accurate value about the recurrence of breast cancer we are going to use data sets which were taken from the UCI machine learning repository and data was anaconda (jupyter note book).

Modules

- Data collection
- Data preprocessing
    1. Data cleaning
    2. Data splitting
    3. Feature selection

3.2.1 Advantages

- To improve the accuracy rate.
- We will try to evaluate some newer algorithms with better feature selection techniques.

CHAPTER 4

SYSTEM SPECIFICATION

4.1 SOFTWARE REQUIREMENTS

Operating System : Windows Pro

Front End : Anaconda Navigator

Language : Python

4.2 HARDWARE REQUIREMENTS

Processor : Intel Core i3

Hard Disk : 512 GB

RAM : 2 GB

4.3 ABOUT THE SOFTWARE

ANACONDA NAVIGATOR

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows you to launch applications and easily manage conda packages, environments and channels without using command-line commands. It is available for Windows, macOS and Linux, Anaconda is a package manager, an environment manager, a Python/R data science distribution, and a collection of over 1500+ open source packages. Anaconda is free and easy to install, and it offers free community support. Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window.

CONDA

Conda is an open source package management system and environment management system that runs on Windows, macOS and Linux. Conda quickly installs, runs and updates packages and their dependencies. Conda easily creates, saves, loads and switches between environments on your local computer. It was created for Python programs, but it can package and distribute software for any language.

Conda can be combined with continuous integration systems such as Travis CI and AppVeyor to provide frequent, automated testing of your code. The conda package and environment manager is included in all versions of Anaconda and Miniconda. Anaconda Repository. Conda is also included in Anaconda Enterprise , which provides on-site enterprise package and environment management for Python, R, Node.js, Java and other application stacks.

APPLICATIONS AVAILABLE IN NAVIGATOR

- JupyterLab
- Jupyter Notebook
- QTConsole
- Spyder
- VSCode
- GlueViz
- Orange 3 App
- Rodeo
- Rstudio

JUPYTER NOTEBOOK

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. The Notebook has support for over 40 programming languages, including Python, R, Julia and Scala. Jupyter Notebooks are a spin off project from IPython project, which used to have an IPython Notebook project itself.

PYTHON

Python is a widely used general-purpose, high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. Python is a programming language that lets you work quickly and integrate systems more efficiently. There are two major Python versions- **Python 2 and Python 3**. Both are quite different. Some of the advantages of using python:

- Emphasis on code readability, shorter codes, ease of writing
- Programmers can express logical concepts in fewer lines of code in comparison to languages such as C++ or Java.
- Python supports multipl**e** programming paradigms, like object-oriented, imperative and functional programming or procedural.
- There exist inbuilt functions for almost all of the frequently used concepts.

FEATURES

- Interpreted
- Platform Independent
- Free and open source
- Redistributable
- Embeddable
- Robust
- Rich Library support

DATA ANALYSIS IN PYTHON

Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of data-centric Python packages. Panda is one of those packages, and makes importing and analyzing data much easier.

CHAPTER 5

SYSTEM DESIGN

5.1 SYSTEM ARCHITECTURE

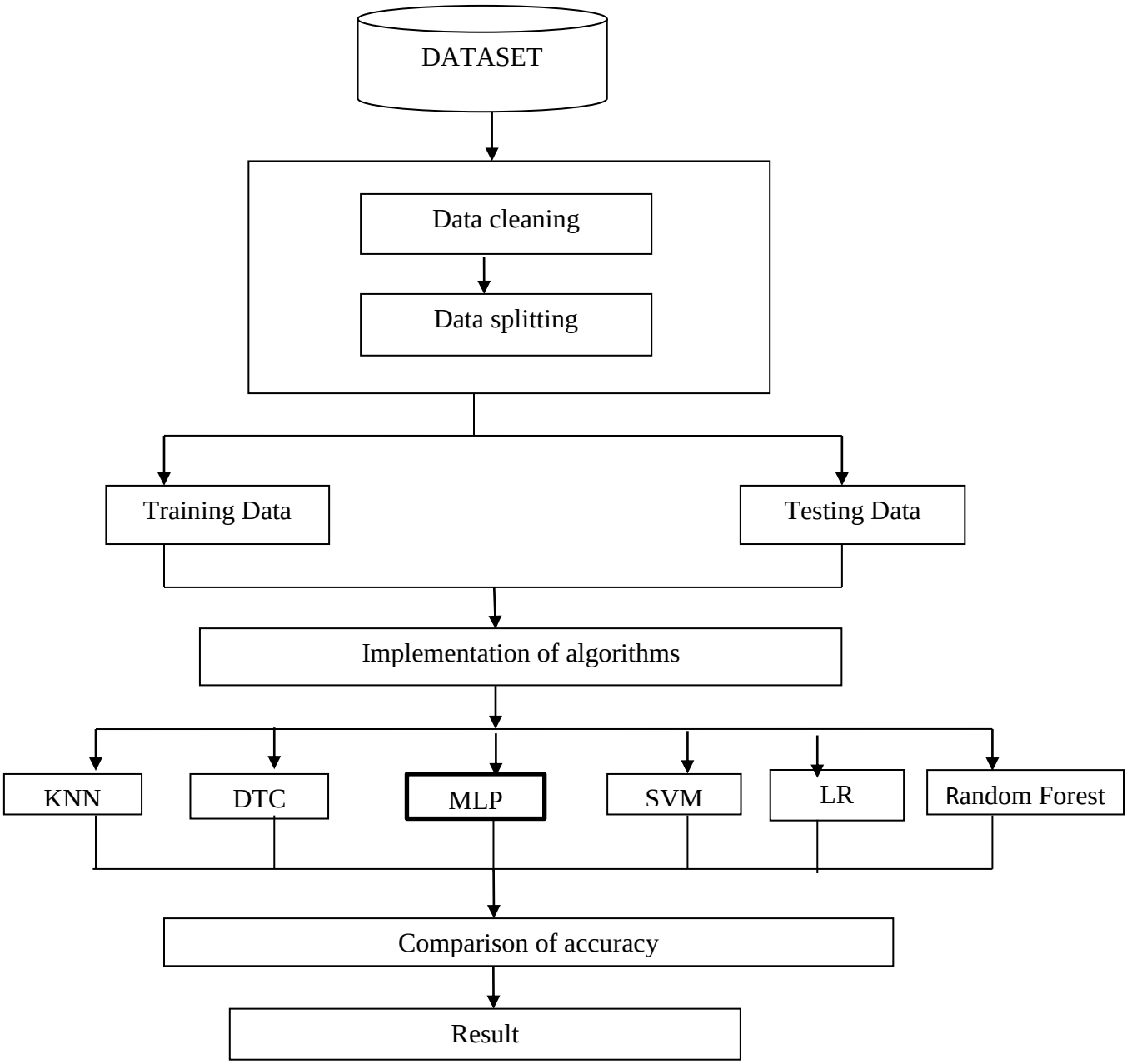

Figure 5.1 System Architecture

5.2 MODULAR DESCRIPTION

The proposed system consists of the following modules

- **Data collection**
- **Data preprocessing**
    4. Data cleaning
    5. Data splitting
    6. Feature selection

5.2.1 Data Collection

Data is a piece of information that should be collected carefully so that the collected information is useful. Data collection is an important step while doing experiments or researches. Data collection is the process of gathering or collecting information that is used for obtaining outcomes in experiments. We collected breast cancer data from the UCI public database. The Breast cancer dataset has 668 instances 31 attributes. The attributes names are Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concavity**,** Symmetry, Fractal dimension, etc.

Data: Data are any facts, numbers, or text that can be processed by a computer.

5.2.2 DATA PREPROCESSING

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent, to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing is used data based driven applications such as customer relationship management and ruled based application. Data preprocessing prepares raw data for further processing.

Data mapping: In this phase, the text data converted into numerical form. That means the data set is formatted into some regions of values. Here every parameter changed to some regions of numerical value.

Data cleaning: The integrated database went through the data cleaning process, in which we removed improper data entries, such as those that provided an irrelevant answer, in the database. To smooth noisy data, the tuples with improper data's were eliminated.

Data splitting: The structured dataset splitting into x and y sets, x contains factor attributes, y contains final choice. The final choice is based on satisfying the factors. Given data has been split into

- Training phase
- Testing phase

In Training phase has 80% of data. Testing phase has 20% of data

Feature selection: In this phase, the process of selecting a subset of relevant attributes. There are 35 factors presented in our dataset but 10 factors is enough to prove the patient affect the cancer or not.

The factors are:

- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Concavity
- Concave

- Symmetry
- Fractal dimension

From the above figure 4.1 illustrates with these classification algorithms for the prediction model, namely, MLP, SVM, Random Forest, Decision Tree, Logistic regression and K-nearest neighbor, were evaluated in the prediction of breast cancer recurrence by using the Wisconsin Prognostic Breast Cancer Dataset.

5.3 ALGORITHM DESCRIPTION

5.3.1K-NEAREST NEIGHBORS (IBK) ALGORITHM

To build a model able to predict target values of information instances in the testing set, for which only known the parameters. Text data are ideally suited for KNN classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories is trained to model aspect classification and this trained KNN is used for classification. The experimental results indicate that the proposed techniques have achieved about 85% accuracy.

5.3.2 SUPPORT VECTOR MACHINE

To build a model able to predict target values of information instances in the testing set, for which only known the parameters. Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories is trained to model aspect classification and this trained SVM is used for classification. The experimental results indicate that the proposed techniques have achieved about 62% accuracy.

5.3.3 RANDOM FORESTS

To build a model able to predict target values of information instances in the testing set, for which only known the parameters. Text data are ideally suited for random forest because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories is trained to model aspect classification and this trained random forest is used for polarity classification per aspect. The experimental results indicate that the proposed techniques have achieved about 95% accuracy. Web based data are applied to emotion cause extraction sub system and complementary feature selection method, based on the output of these features are merged.

5.3.4 MULTILAYER PERCEPTRON

To build a model able to predict target values of information instances in the testing set, for which only known the parameters. Text data are ideally suited for MLP because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories is trained to model aspect classification and this trained MLP is used for polarity classification per aspect. The experimental results indicate that the proposed techniques have achieved about 83% accuracy.

5.3.5 LOGISTIC REGRESSION

To build a model able to predict target values of information instances in the testing set, for which only known the parameters. Text data are ideally suited for logistic regression because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories is trained to model aspect classification and this

trained logistic regression is used for classification. The experimental results indicate that the proposed techniques have achieved about 83% accuracy

5.3.6 DECISION TREE CLASSIFIER

To build a model able to predict target values of information instances in the testing set, for which only known the parameters. Text data are ideally suited for decision tree because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories is trained to model aspect classification and this trained decision tree is used for polarity classification per aspect. The experimental results indicate that the proposed techniques have achieved about 87% accuracy.

5.4RECOMMEDATION SYSTEM

At last the final stage, breast cancer can be predicted based on the selecting factors. This research is achieved by MLP, SVM, Random Forest, Decision Tree, Logistic regression and K-nearest neighbor. Our goal of the research specifically noted on the accuracy level. Mainly focused on the factors are Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concavity**,** Symmetry and Fractal dimension. These 10 factors are the major factors for the breast cancer prediction. In this recommendation system, this part is an important phase for our outcome.

CHAPTER 6

CONCLUSION AND FUTUREWORK

6.1 CONCLUSION

A new approach for the prediction of recurrence cancer using medical records of the patients, by applying classification algorithm of data mining. Breast cancer dataset is used for the experimental purpose. Dataset has come up with the information of patients undergone the treatment of Breast Cancer. From the result of this experiment we can conclude that patient with specific range of attribute value have more chances of recurrence cancer. Main focus of this work conducted is to improve the prediction of breast cancer in order to increase the accuracy of diagnosis. Most of the studies are presented which have been proposed in several years and emphasis on the development of predictive models for breast cancer diagnosis/prognosis using machine learning methods and classification. The comparative analysis of four widely used machine learning techniques: Random Forest Tree, Support vector machine (SVM), K- Nearest Neighbor (KNN), MLP, Logistic Regression and Decision Tree Classifier are performed. The performance in terms of accuracy, Random Forest is better as compared to other techniques. In order to further improve accuracy.

6.2 FUTURE WORK

The proposed model can further used to increasing the prediction of accuracy level. Future works should consider additional parameters or addition of more symptoms based on cancer and other updates in dataset to give a better accuracy. They should also consider the issues of the classification and help to reach the better accuracy using other classification algorithm.

APPENDIX-I

SAMPLE SOURE CODE

Importing Data

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.metrics
import accuracy_score, classification_report, confusion_matrix
```

Reading CSV File

```
raw_data=pd.read_csv('breast-cancer-wisconsin-data.csv',delimiter=',')
raw_data.head(10)
raw_data.tail(10)
```

Visualizing The Data

```
print(cancer['feature_names'])
print(cancer['target_names'])
raw_data.shape
raw_data.info()
sns.pairplot(raw_data,hue='Diagnosis',vars=['Mean Radius','Mean Texture', 'Mean Perimeter','Mean Area','MeanSmoothness','Mean Compactness',
'Mean Concavity','Mean Concave points','Mean Symmetry','Mean Fractal dimension'])
```

```
sns.countplot(raw_data['Diagnosis'])
plt.figure(figsize=(20,9))
sns.heatmap(raw_data.corr(),annot=True)
```

Data Preprocessing

```
cancer_mapping={'Benign':0,'Malign':1}
raw_data.Diagnosis= raw_data..Diagnosis.map(cancer_mapping)
raw_data.duplicated()
```

Features Selection

```
y=raw_data['Diagnosis']
x=raw_data[['Mean
Radius','MeanTexture','MeanPerimeter','MeanArea','MeanSmoothness','MeanCom
pactness','MeanConcavity','Mean Concave points','MeanSymmetry','Mean Fractal
dimension']]
```

Data splitting

```
from sklearn.model_selection import train_test_split x_train, x_test, y_train,
y_test=train_test_split(x, y, test_size=0.2)
```

Implementing Algorithm

Support Vector Machine

```
from sklearn.datasets import load_breast_cancer
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
%matplotlib inline
cancer=load_breast_cancer()
svm=SVC()
svm.fit(x_train,y_train)
```

```
print('accuracy on the training subset:(:,3f)', format (svm.Score (x_train ,y_train)))
print('accuracy on the test subset: (:,3f)',format(svm.score(x_test, y_test)))
```

KNN

```
from sklearn.neighbors import KNeighborsClassifier
import mglearn
mglearn.plots.plot_knn_classification(n_neighbors=3)
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
cancer=load_breast_cancer()
knn=KNeighborsClassifier()
knn.fit(x_train,y_train)
print('accuracy of KNN n=5, on the training set:(:,3f)',format(knn.score (x_train,y_train)))
print('accuracy of KNN n=5, on the test set:(:,3f)',format(knn.score(x_test,y_test)))
training_accuracy=[]
test_accuracy=[]
neighbors_settings=range(1,11)
forn_neighbors in neighbors_settings:
clf=KNeighborsClassifier(n_neighbors=n_neighbors)
clf.fit(x_train,y_train)
training_accuracy.append(clf.score(x_train,y_train))
test_accuracy.append(clf.score(x_test,y_test))
plt.plot(neighbors_settings,training_accuracy, label='accuracy of the training set')
plt.plot(neighbors_settings,test_accuracy, label='accuracy of the test set')
plt.ylabel('accuracy')
plt.xlabel('Number of Neighbors')
plt.legend()
```

```
import matplotlib.pyplot as plt
%matplotlib inline
cancer=load_breast_cancer()
knn=KNeighborsClassifier()
knn.fit(x_train,y_train
```

Randomforest

```
from sklearn.datasets import load_breast_cancer
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
%matplotlib inline
cancer=load_breast_cancer()
x_train,x_test,y_train,y_test=train_test_split(cancer.data,cancer.target,  stratify=
cancer.target,random_state=42)
forest=RandomForestClassifier(n_estimators=100,random_state=0)
forest.fit(x_train,y_train)

print('accuracy on the training subset:(:,3f)',format(forest.score (x_train,y_train)))
print('accuracy on the test subset:(:,3f)',format(forest.score(x_test,y_test)))
```

Logistic Regression

```
from sklearn.datasets import load_breast_cancer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
 import matplotlib.pyplot as plt
%matplotlib inline
cancer=load_breast_cancer()
log_reg=LogisticRegression()
```

```
log_reg.fit(x_train,y_train)
print('accuracy on the training subset:(:,3f)',format
(log_reg.score(x_train,y_train)))
print('accuracy on the test subset:(:,3f)',format
(log_reg.score(x_test,y_test)))
importmglearn
mglearn.plots.plot_linear_regression_wave()
```

Decision Tree

```
from sklearn.datasets import load_breast_cancer
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
%matplotlib inline
cancer=load_breast_cancer()
tree=DecisionTreeClassifier(random_state=0)
tree.fit(x_train,y_train)
print('accuracy on the training subset:(:,3f)',format
(tree.score(x_train,y_train)))
print('accuracy on the test subset:(:,3f)',format(tree.score(x_test,y_test)))
```

MLP

```
from sklearn.datasets import load_breast_cancer
from sklearn.neural_network import MLPClassifier
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
%matplotlib inline
cancer=load_breast_cancer()
mlp=MLPClassifier(random_state=42)
```

```
mlp.fit(x_train,y_train)
print('accuracy on the training subset:(:,3f)',format(mlp.score(x_train,y_train)))
print('accuracy on the test subset:(:,3f)',format(mlp.score(x_test,y_test)))
```

APPENDIX-II

SAMPLE SCREENSHOTS

IMPORTING DATA

```
[47]: from sklearn.datasets import load_breast_cancer
      from sklearn.neighbors import KNeighborsClassifier
      from sklearn.linear_model import LogisticRegression
      from sklearn.svm import SVC
      from sklearn import tree
      from sklearn.tree import DecisionTreeClassifier
      from sklearn.neural_network import MLPClassifier
      from sklearn.ensemble import RandomForestClassifier
      from sklearn.model_selection import train_test_split

      import matplotlib.pyplot as plt
      import seaborn as sns

      %matplotlib inline
```

Figure 1

READING CSV FILE

```
In [4]: import pandas as pd
        raw_data=pd.read_csv('breast-cancer-wisconsin-data.csv',delimiter=',')
        raw_data.tail(10)
        raw_data.head(10)
```

Out[4]:

n s	Mean Concavity	Mean Concave points	Mean Symmetry	Mean Fractal dimension	...	Worst Texture	Worst Perimeter	Worst Area	Worst Smoothness	Worst Compactness	Worst Concavity	Worst Concave points	Worst Symmetry	Worst Fractal dimension	Diagnosis
0	0.30010	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0	0.1622	0.6656	0.7119	0.2654	0.4601	0.11890	Malign
4	0.08690	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902	Malign
0	0.19740	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758	Malign
0	0.24140	0.10520	0.2597	0.09744	...	26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300	Malign
0	0.19800	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678	Malign
0	0.15780	0.08089	0.2087	0.07613	...	23.75	103.40	741.6	0.1791	0.5249	0.5355	0.1741	0.3985	0.12440	Malign
0	0.11270	0.07400	0.1794	0.05742	...	27.66	153.20	1606.0	0.1442	0.2576	0.3784	0.1932	0.3063	0.08368	Malign
0	0.09366	0.05985	0.2196	0.07451	...	28.14	110.60	897.0	0.1654	0.3682	0.2678	0.1556	0.3196	0.11510	Malign
0	0.18590	0.09353	0.2350	0.07389	...	30.73	106.20	739.3	0.1703	0.5401	0.5390	0.2060	0.4378	0.10720	Malign
0	0.22730	0.08543	0.2030	0.08243	...	40.68	97.65	711.4	0.1853	1.0580	1.1050	0.2210	0.4366	0.20750	Malign

Figure 2

```
In [3]: cancer=load_breast_cancer()
        print(cancer['DESCR'])
```

```
Breast Cancer Wisconsin (Diagnostic) Database
=============================================

Notes
-----
Data Set Characteristics:
    :Number of Instances: 569

    :Number of Attributes: 30 numeric, predictive attributes and the class

    :Attribute Information:
        - radius (mean of distances from center to points on the perimeter)
        - texture (standard deviation of gray-scale values)
        - perimeter
        - area
        - smoothness (local variation in radius lengths)
        - compactness (perimeter^2 / area - 1.0)
        - concavity (severity of concave portions of the contour)
        - concave points (number of concave portions of the contour)
```

Figure 3

VISUALIZING THE DATA

```
In [62]: sns.pairplot(raw_data,hue='Diagnosis',vars=['Mean Radius','Mean Texture','Mean Perimeter','Mean Area','Mean Smoothne
```

```
C:\user\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for multidi
ensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted
as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

```
Out[62]: <seaborn.axisgrid.PairGrid at 0xc5b8aaeb38>
```

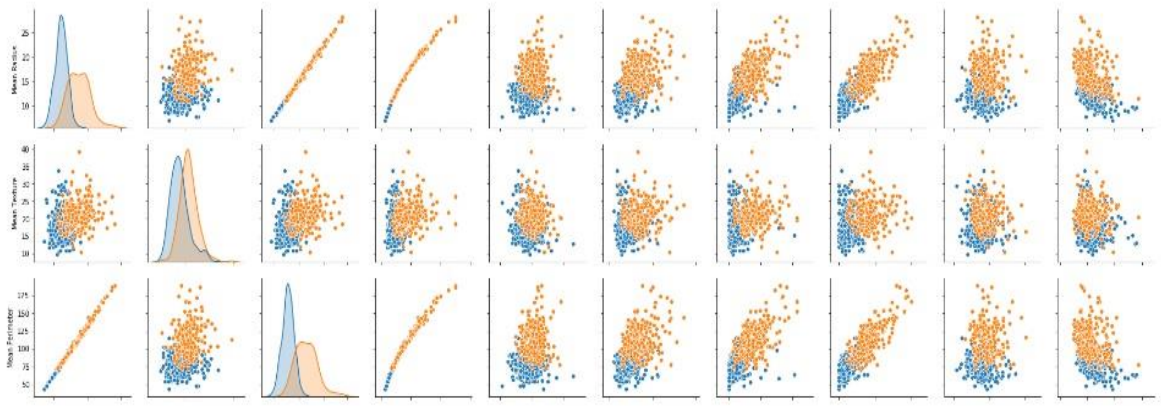

Figure 4

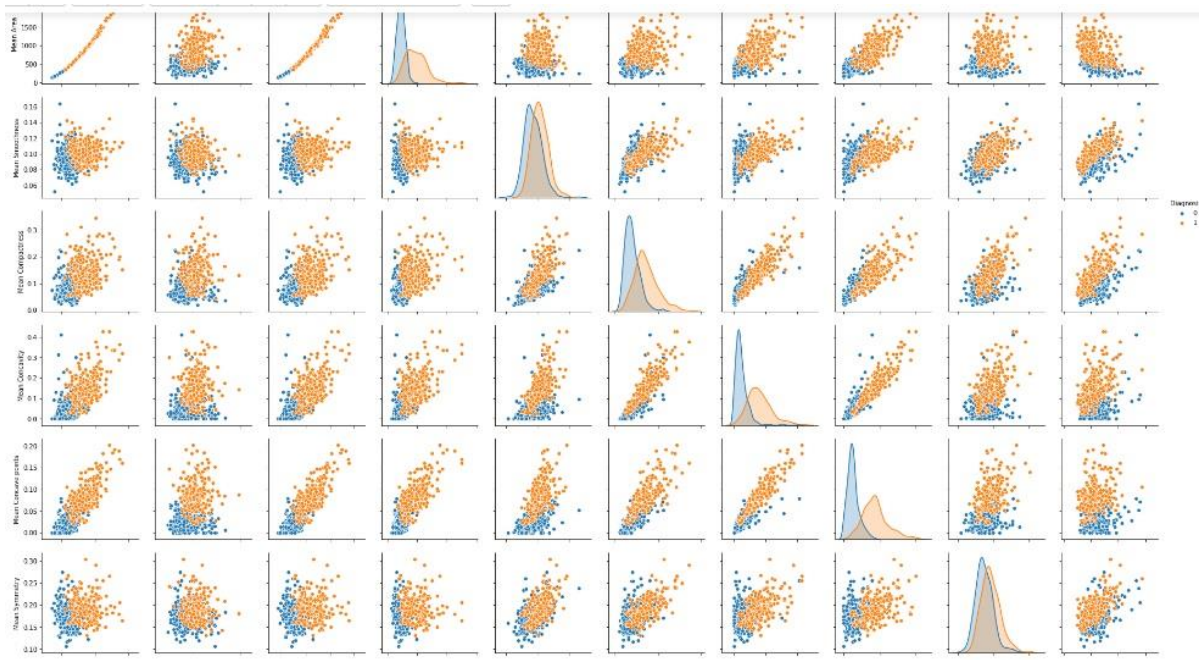

Figure 5

```
In [59]: sns.countplot(raw_data['Diagnosis'])

Out[59]: <matplotlib.axes._subplots.AxesSubplot at 0xc5bb2fa8d0>
```

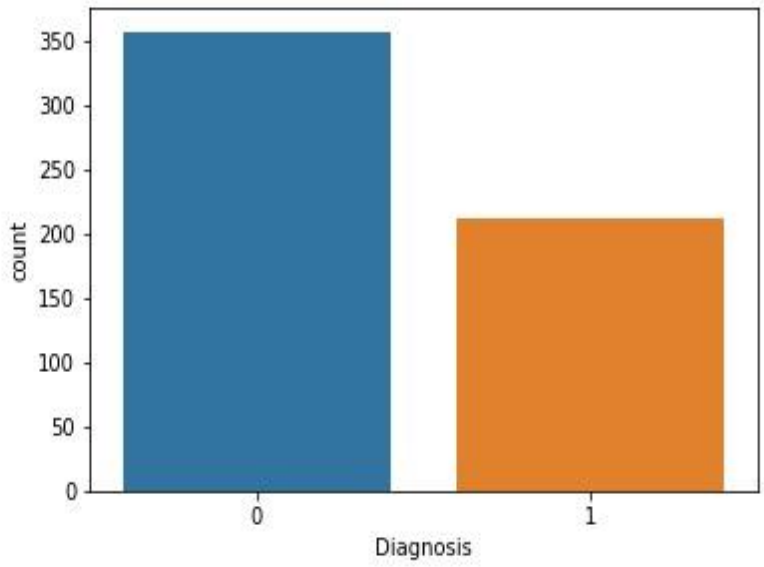

Figure 6

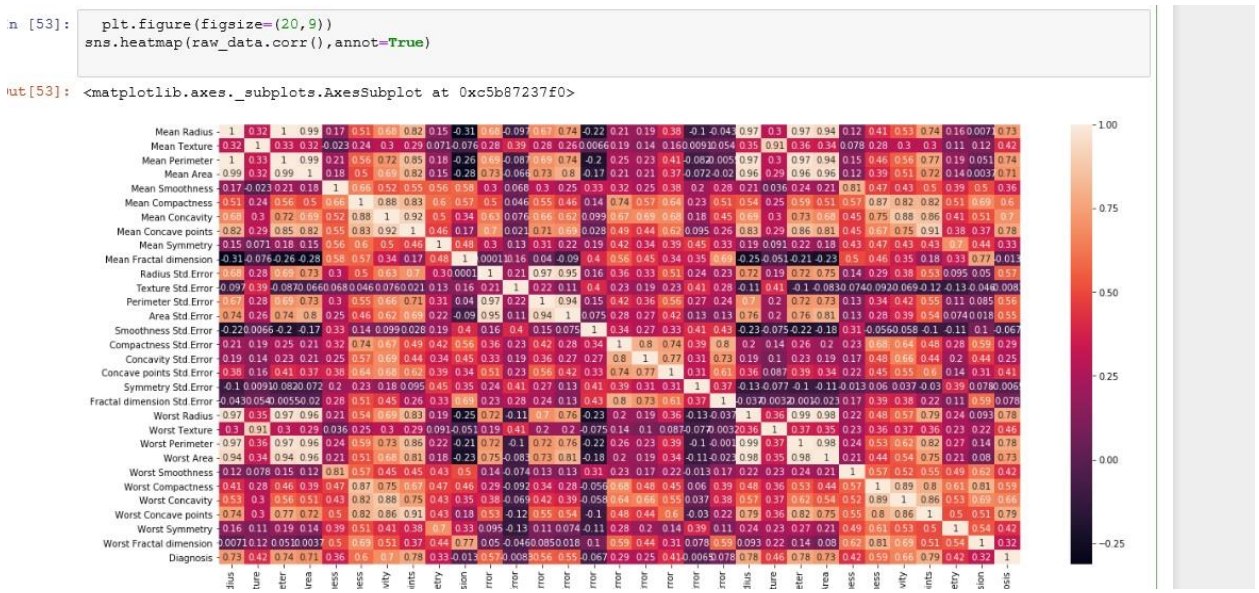

Figure 7

DATA PREPROCESSING

```
In [5]: print(cancer['feature_names'])

['mean radius' 'mean texture' 'mean perimeter' 'mean area'
 'mean smoothness' 'mean compactness' 'mean concavity'
 'mean concave points' 'mean symmetry' 'mean fractal dimension'
 'radius error' 'texture error' 'perimeter error' 'area error'
 'smoothness error' 'compactness error' 'concavity error'
 'concave points error' 'symmetry error' 'fractal dimension error'
 'worst radius' 'worst texture' 'worst perimeter' 'worst area'
 'worst smoothness' 'worst compactness' 'worst concavity'
 'worst concave points' 'worst symmetry' 'worst fractal dimension']

In [6]: print(cancer['target_names'])

['malignant' 'benign']

In [9]: cancer_mapping={'Benign':0,'Malign':1}
raw_data.Diagnosis=raw_data.Diagnosis.map(cancer_mapping)

In [8]: cancer.data.shape

Out[8]: (569, 30)
```

Figure 8

DATA SPLITTING

```
In [12]: y=raw_data['Diagnosis']
         x=raw_data[['Mean Radius','Mean Texture','Mean Perimeter','Mean Area','Mean Smoothness','Mean Compa
```

```
In [13]: x
```

Out[13]:

	Mean Radius	Mean Texture	Mean Perimeter	Mean Area	Mean Smoothness	Mean Compactness	Mean Concavity	Mean Concave points	M Symme
0	17.990	10.38	122.80	1001.0	0.11840	0.27760	0.300100	0.147100	0.2
1	20.570	17.77	132.90	1326.0	0.08474	0.07864	0.086900	0.070170	0.1
2	19.690	21.25	130.00	1203.0	0.10960	0.15990	0.197400	0.127900	0.2
3	11.420	20.38	77.58	386.1	0.14250	0.28390	0.241400	0.105200	0.2
4	20.290	14.34	135.10	1297.0	0.10030	0.13280	0.198000	0.104300	0.1
5	12.450	15.70	82.57	477.1	0.12780	0.17000	0.157800	0.080890	0.2
6	18.250	19.98	119.60	1040.0	0.09463	0.10900	0.112700	0.074000	0.1
7	13.710	20.83	90.20	577.9	0.11890	0.16450	0.093660	0.059850	0.2
8	13.000	21.82	87.50	519.8	0.12730	0.19320	0.185900	0.093530	0.2
9	12.460	24.04	83.97	475.9	0.11860	0.23960	0.227300	0.085430	0.2
10	16.020	23.24	102.70	797.8	0.08206	0.06669	0.032990	0.033230	0.1
11	15.780	17.89	103.60	781.0	0.09710	0.12920	0.099540	0.066060	0.1
12	19.170	24.80	132.40	1123.0	0.09740	0.24580	0.206500	0.111800	0.2
13	15.850	23.95	103.70	782.7	0.08401	0.10020	0.099380	0.053640	0.1

Figure 9

```
n [14]: y
```

```
ut[14]: 0      1
        1      1
        2      1
        3      1
        4      1
        5      1
        6      1
        7      1
        8      1
        9      1
        10     1
        11     1
        12     1
        13     1
        14     1
        15     1
        16     1
        17     1
        18     1
        19     0
        20     0
        21     0
        22     1
```

Figure 10

```
In [26]: from sklearn.model_selection import train_test_split
         x_train, x_test, y_train, y_test=train_test_split(x, y, test_size=0.2)

In [27]: x_train
```

Out[27]:

	Mean Radius	Mean Texture	Mean Perimeter	Mean Area	Mean Smoothness	Mean Compactness	Mean Concavity	Mean Concave points	Mean Symmetry	Mean Fractal dimension
63	9.173	13.86	59.20	260.9	0.07721	0.08751	0.059880	0.021800	0.2341	0.06963
159	10.900	12.96	68.69	366.8	0.07515	0.03718	0.003090	0.006588	0.1442	0.05743
215	13.860	16.93	90.96	578.9	0.10260	0.15170	0.099010	0.056020	0.2106	0.06916
477	13.900	16.62	88.97	599.4	0.06828	0.05319	0.022240	0.013390	0.1813	0.05536
345	10.260	14.71	66.20	321.6	0.09882	0.09159	0.035810	0.020370	0.1633	0.07005
395	14.060	17.18	89.75	609.1	0.08045	0.05361	0.026810	0.032510	0.1641	0.05764
439	14.020	15.66	89.59	606.5	0.07966	0.05581	0.020870	0.026520	0.1589	0.05586
236	23.210	26.97	153.50	1670.0	0.09509	0.16820	0.195000	0.123700	0.1909	0.06309
362	12.760	18.84	81.87	496.6	0.09676	0.07952	0.026880	0.017810	0.1759	0.06183
499	20.590	21.24	137.80	1320.0	0.10850	0.16440	0.218800	0.112100	0.1848	0.06222
300	19.530	18.90	129.50	1217.0	0.11500	0.16420	0.219700	0.106200	0.1792	0.06552
101	6.981	13.43	43.79	143.5	0.11700	0.07568	0.000000	0.000000	0.1930	0.07818
531	11.670	20.02	75.21	416.2	0.10160	0.09453	0.042000	0.021570	0.1859	0.06461
76	13.530	10.94	87.91	559.2	0.12910	0.10470	0.068770	0.065560	0.2403	0.06641
144	10.750	14.97	68.26	355.3	0.07793	0.05139	0.022510	0.007875	0.1399	0.05688

Figure 11

```
In [28]: x_test
```

Out[28]:

	Mean Radius	Mean Texture	Mean Perimeter	Mean Area	Mean Smoothness	Mean Compactness	Mean Concavity	Mean Concave points	Mean Symmetry	Mean Fractal dimension
399	11.800	17.26	75.26	431.9	0.09087	0.06232	0.028530	0.016380	0.1847	0.06019
429	12.720	17.67	80.98	501.3	0.07896	0.04522	0.014020	0.018350	0.1459	0.05544
492	18.010	20.56	118.40	1007.0	0.10010	0.12890	0.117000	0.077620	0.2116	0.06077
5	12.450	15.70	82.57	477.1	0.12780	0.17000	0.157800	0.080890	0.2087	0.07613
65	14.780	23.94	97.40	668.3	0.11720	0.14790	0.126700	0.090290	0.1953	0.06654
491	17.850	13.23	114.60	992.1	0.07838	0.06217	0.044450	0.041780	0.1220	0.05243
438	13.850	19.60	88.68	592.6	0.08684	0.06330	0.013420	0.022930	0.1555	0.05673
463	11.600	18.36	73.88	412.7	0.08508	0.05855	0.033670	0.017770	0.1516	0.05859
275	11.890	17.36	76.20	435.6	0.12250	0.07210	0.059290	0.074040	0.2015	0.05875
104	10.490	19.29	67.41	336.1	0.09989	0.08578	0.029950	0.012010	0.2217	0.06481
470	9.667	18.49	61.49	289.1	0.08946	0.06258	0.029480	0.015140	0.2238	0.06413
216	11.890	18.35	77.32	432.2	0.09363	0.11540	0.066360	0.031420	0.1967	0.06314
146	11.800	16.58	78.99	432.0	0.10910	0.17000	0.165900	0.074150	0.2678	0.07371
165	14.970	19.76	95.50	690.2	0.08421	0.05352	0.019470	0.019390	0.1515	0.05266
197	18.080	21.84	117.40	1024.0	0.07371	0.08642	0.110300	0.057780	0.1770	0.05340

Figure 12

```
: y_train

: 63      0
  159     0
  215     1
  477     0
  345     0
  395     0
  439     0
  236     1
  362     0
  499     1
  300     1
  101     0
  531     0
  76      0
  144     0
  367     0
  208     0
  182     1
  526     0
  535     1
```

Figure 13

KNN

```
In [20]: import mglearn
         mglearn.plots.plot_knn_classification(n_neighbors=3)

C:\user\Anaconda3\lib\site-packages\sklearn\utils\deprecation.py:77: DeprecationW
cated; Please import make_blobs directly from scikit-learn
  warnings.warn(msg, category=DeprecationWarning)
```

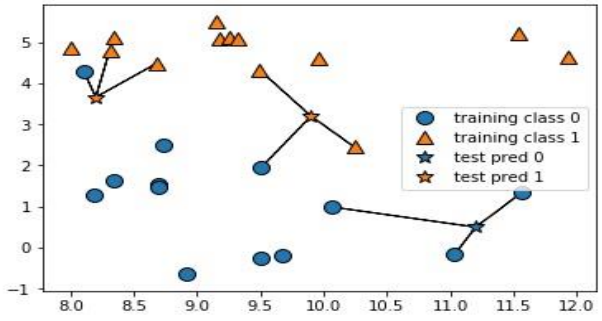

Figure 14

In [31]:
```
from sklearn.datasets import load_breast_cancer
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split

import matplotlib.pyplot as plt

%matplotlib inline

cancer=load_breast_cancer()
knn=KNeighborsClassifier()
knn.fit(x_train,y_train)
```

Out[31]:
```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=1, n_neighbors=5, p=2,
           weights='uniform')
```

In [32]:
```
print('accuracy of KNN n=5, on the training set:(:,3f)',format(knn.score(x_train,y_train)))
print('accuracy of KNN n=5, on the test set:(:,3f)',format(knn.score(x_test,y_test)))
```

```
accuracy of KNN n=5, on the training set:(:,3f) 0.9186813186813186
accuracy of KNN n=5, on the test set:(:,3f) 0.8596491228070176
```

Figure 15

In [36]:
```
training_accuracy=[]
test_accuracy=[]
neighbors_settings=range(1,11)
for n_neighbors in neighbors_settings:
    clf=KNeighborsClassifier(n_neighbors=n_neighbors)
    clf.fit(x_train,y_train)
    training_accuracy.append(clf.score(x_train,y_train))
    test_accuracy.append(clf.score(x_test,y_test))
plt.plot(neighbors_settings,training_accuracy, label='accuracy of the training set')
plt.plot(neighbors_settings,test_accuracy, label='accuracy of the test set')
plt.ylabel('accuracy')
plt.xlabel('Number of Neighbors')
plt.legend()
```

Out[36]:
```
<matplotlib.legend.Legend at 0xc5b75e8ac8>
```

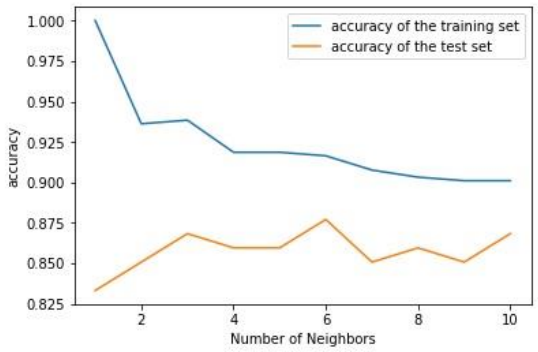

Figure 16

LOGISTIC REGRESSION

```
]: from sklearn.datasets import load_breast_cancer
   from sklearn.linear_model import LogisticRegression
   from sklearn.model_selection import train_test_split

   import matplotlib.pyplot as plt
   %matplotlib inline

   cancer=load_breast_cancer()

   log_reg=LogisticRegression()
   log_reg.fit(x_train,y_train)
```

```
]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

```
]: print('accuracy on the training subset:(:,3f)',format(log_reg.score(x_train,y_train)))
   print('accuracy on the test subset:(:,3f)',format(log_reg.score(x_test,y_test)))
```

```
accuracy on the training subset:(:,3f) 0.9208791208791208
accuracy on the test subset:(:,3f) 0.8859649122807017
```

Figure 17

```
: import mglearn
  mglearn.plots.plot_linear_regression_wave()
```

```
w[0]: 0.393906  b: -0.031804
```

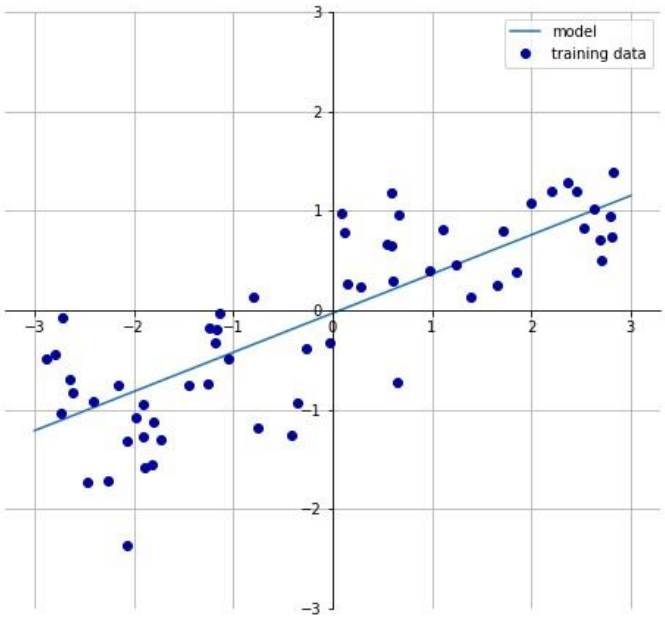

Figure 18

DECISION TREE

In [41]:
```
from sklearn.datasets import load_breast_cancer
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split

import matplotlib.pyplot as plt
%matplotlib inline

cancer=load_breast_cancer()

tree=DecisionTreeClassifier(random_state=0)
tree.fit(x_train,y_train)

print('accuracy on the training subset:(:,3f)',format(tree.score(x_train,y_train)))
print('accuracy on the test subset:(:,3f)',format(tree.score(x_test,y_test)))
```

```
accuracy on the training subset:(:,3f) 1.0
accuracy on the test subset:(:,3f) 0.8771929824561403
```

Figure 19

RANDOMFOREST

```
from sklearn.datasets import load_breast_cancer
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split

import matplotlib.pyplot as plt
%matplotlib inline

cancer=load_breast_cancer()
x_train,x_test,y_train,y_test=train_test_split(cancer.data,cancer.target,stratify=cancer.target,random_state=42)

forest=RandomForestClassifier(n_estimators=100,random_state=0)
forest.fit(x_train,y_train)

print('accuracy on the training subset:(:,3f)',format(forest.score(x_train,y_train)))
print('accuracy on the test subset:(:,3f)',format(forest.score(x_test,y_test)))
```

```
C:\user\Anaconda3\lib\site-packages\sklearn\ensemble\weight_boosting.py:29: DeprecationWarning: numpy.core.umath_test
s is an internal NumPy module and should not be imported. It will be removed in a future NumPy release.
  from numpy.core.umath_tests import inner1d

accuracy on the training subset:(:,3f) 1.0
accuracy on the test subset:(:,3f) 0.958041958041958
```

Figure 20

MLP

```
In [43]: from sklearn.datasets import load_breast_cancer
         from sklearn.neural_network import MLPClassifier
         from sklearn.model_selection import train_test_split

         import matplotlib.pyplot as plt
         %matplotlib inline

         cancer=load_breast_cancer()

         mlp=MLPClassifier(random_state=42)
         mlp.fit(x_train,y_train)

         print('accuracy on the training subset:(:,3f)',format(mlp.score(x_train,y_train)))
         print('accuracy on the test subset:(:,3f)',format(mlp.score(x_test,y_test)))

         accuracy on the training subset:(:,3f) 0.8849765258215962
         accuracy on the test subset:(:,3f) 0.8321678321678322
```

Figure 21

SVM

```
         accuracy on the test subset:(:,3f) 0.8321678321678322

In [45]: from sklearn.datasets import load_breast_cancer
         from sklearn.svm import SVC
         from sklearn.model_selection import train_test_split

         import matplotlib.pyplot as plt
         %matplotlib inline

         cancer=load_breast_cancer()

         svm=SVC()
         svm.fit(x_train,y_train)

         print('accuracy on the training subset:(:,3f)',format(svm.score(x_train,y_train)))
         print('accuracy on the test subset:(:,3f)',format(svm.score(x_test,y_test)))

         accuracy on the training subset:(:,3f) 1.0
         accuracy on the test subset:(:,3f) 0.6293706293706294
```

Figure 22

REFERENCES

[1]Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and RazaviAR"Usingthree machine learning techniques for predicting breast cancer recurrence"

[2] Ahmed Iqbal Pritom ,Md. Ahadur Rahman Munshi ,Shahed Anzarus Sabab Shihabuzzaman Shihab" Predicting Breast Cancer Recurrence using effective Classification and Feature Selection techniques"

[3] Daad Abdullah Almuhaidib ,Hadil Ahmed Shaiba , Fatima MotebAlbusayyis ,MashaelAbdulalimAlzaid, Najla Ghazi Alharbi, Sara Muhammad Alotaibi and Reem Mohammed Almadhi "Ensemble learning method for the prediction of breast cancer recurrence"

[4] Madhuri Gupta,Bharat Gupta"A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques"

[5]Dr Prof. Neeraj Bhargava ,Sakshi Sharma ,RenukaPurohit , Pramod Singh Rathore" Prediction of Recurrence Cancer using J48 Algorithm."

[6] Smaranda ,Abdel-Badeeh Salem ,Florin Gorunescu ,Marina Gorunescu"Clustering-based approach for detecting breast cancer recurrence "

[7] Uma Ojha, Dr. Savita Goel "A Study onPrediction of Breast Cancer Recurrence using Data mining techniques".

[8] Umesh , Dr. B Ramachandra "Association Rule Mining Based Predicting Breast Cancer Recurrence on SEER Breast Cancer Data".

[9] XiaoyiXu, Ya Zhang, Liang Zou, Minghui Wang, Ao Li "A Gene Signature for Breast Cancer Prognosis Using Support Vector Machine"