# ABSTRACT

Cancer is an incurable and high-risk disease, formed by the growth of abundance of cells in a very unrestrained and uncharacteristic behavior due to the collection of cells exiting from DNA harm. Cancer structure is dissimilar for all kinds of cancer. The process of diagnosis remains usually almost the same for all. Data Mining Algorithms provide us rapid and accurate grouping and is a promising tool for grouping of the tumors.

Data mining method defines in this paper shows affords innovation and method to alternate voluminous records into recommended records for achieving a decision. By the usage of statistics mining constructions it wishes a lot much less funding for the forecast of the sickness with more accuracy and precision. Information mining mechanisms can reply exchange addresses that expectedly being used a lot time abolishing to choose. This evaluates pretty a variety classifiers and algorithms are used for the expectation of Cancer.

The algorithm discussed in this paper is Random Forest with the Performance Analysis of Different Prediction Algorithms for Breast Cancer Dataset. Random Forest is the most used algorithm in the data mining strategy and one of the oldest algorithms to be used for this purpose. With time the algorithm is changed as per the needs of the researchers to obtain the results, Implementation and diagram of scientific desire help to assist clinical practitioner and experts in prognosis and prediction.

Approach of data mining improves the medical decisions by providing information which is buried in the data. Hidden patterns can be examined in complex and voluminous collections of data and these patterns escape traditional statistical techniques for evaluation. In this context, data is classified and modeled in unsupervised and supervised learning methods by approach of data mining.

Early detection of breast most cancers is some distance less difficult to cure. Tree mainly primarily based statistics mining technique for early detection of breast cancer. Breast most cancers prognosis differentiates benign (lacks plausible to invade neighboring tissue) from malignant (ability to invade neighboring tissue) breast tumors. This paper additionally discusses a range of archives mining methods that have been utilized for breast most cancers diagnosis, and moreover summarizes breast most cancers in generic (types, danger factors, signs and symptoms and signs and symptoms and treatment).

Clinicians and computer scientists both collaborate with each other for looking at the data mining applications. This research work shows the utilization of data mining techniques and pattern into framework of risk prediction can be further extended for other medical domains. Various aspects related to data mining are cleaning, filtering used together with pattern recognition techniques and elimination of redundant features without losing the important characteristics of the data domain. The whole implementation is carried out with using three different data mining algorithms and UCI repository data sets with different parameters to check the compatibility of algorithm. There for the conclusion obtained with this approach is that the more accurate and optimized results are shown by Random Forest.

# CHAPTER-1

**INTRODUCTION**

This chapter gives the brief introduction about the past production and implementation of the data mining and its sub algorithms. This chapter is consisting of history and introduction of various parts of the data mining strategy.

## INTRODUCTION OF DATA MINING:

In the 1990s, the time length “Data Mining” used to be once introduced, on the other hand facts mining is the evolution of a quarter with an large history. Explicit hands-on data investigationhas progressively been extended ,computerized records processing,and  exceptional laptop science discoveries such as Neural networks, Clustering, Genetic algorithms (1950s), Decision bushes (1960s) and Supporting vector machines (1990s). Data mining origins are traced once more to three families’ line: Classical statistics, Artificial Genius and Machine learning. Data mining is an interdisciplinary subfield of laptop science and statistics with an common cause to from the extract documents (with smart methods) from a dataset and critically trade the information into a comprehensible form for in a similar fashion use. Data mining is the contrast step of the “Knowledge discovery in databases” manner or KDD.

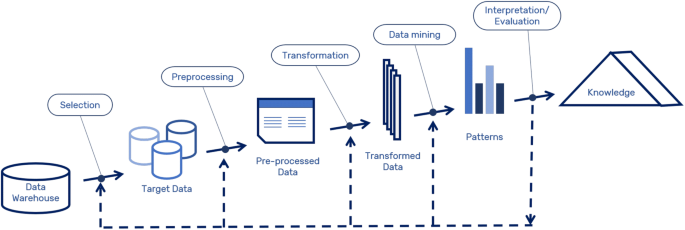


Figure 1.Data mining in KDD process

## FUNCTIONALITIES IN DATA MINING:

In this process, therefore it has multiple functionalities into it which have their own working methods, data mining process is the collection of various machine learning algorithms and techniques and methods to treat the raw data and to find out the suitable information out if it. These functionalities are developed according to the dataset got and the quality of information has to be extracted. No two functionalities are same but can be merged to obtain the better results and their comes the role of machine learning which is used to build the models to carry out the operations on the data set.

Figure 2.Functionalities in the Data mining process

Data Mining Functionalities

Regression and Prediction

Classification Clustering

Mining frequent patterns, Association and Correlation

Outlier analysis

Characterization & Discrimination

Figure 3. Data Mining and its algorithms



Data mining



Regression

Classification

Supervised learning



**Clustering**

Unsupervised learning



Gradient Boosting

AdaBoost

Random forest

Linear regression



Association

Top-k Association

FPGrowth

Targeted Association

* **K-means**
* K-mediods
* K-modes
* PAM
* CLARAN

S

* CLARA
* FCM
* FCMdC
* Fanny



Grid-based method

* STING
* Wave Cluster
* CLIQUE
* Optigrid

Model-based method

* EM
* COBWEB
* CLASSIT
* SOMs

Density-based method

* DBSCAN
* OPTICS
* DBCLASD
* DENCLUE
* Spectral
* Subtractive
* Mean shift



Naïve Bayes

CNN

Logistic Regression

ANN

SVM

Rules

Baysian Classifie

Decision tree

KNN



Hierarchical method

* BIRCH
* CURE
* ROCK
* Chameleon
* Echidna
* Diana
* Agnes

**Partitioning method**

Multilayer perceptron

## SUPERVISED LEARNING

In this Learning, a calculation incorporates the ideal arrangements known as names. Essentially, in this classification, we educate or train the framework utilizing information that is involved marks. From that point onward, the framework is furnished with another arrangement of information to such an extent that an Algorithm examinations the preparation information and gives a right outcome from the information with very much characterized marks.

### Classification

Classification is the process of finding the class of the data elements. It basically classifies the data elements as per their matches and put them under the same class. This division is known as the discrete values. This method deals with the prediction of discrete values. Classification came under the Predictive Task.

### Regression

Regression is the information mining capacity which is utilized to anticipate a number. Benefit, deals, contract rates, house estimations, area and so forth, could all be anticipated utilizing relapse strategy however the necessity to complete this is the final product must be given to whom it must be looked at. Relapse went under the Predictive Task.

## UNSUPERVISED LEARNING

This method is applied on those data sets which don’t have any output values or the output is unknown.

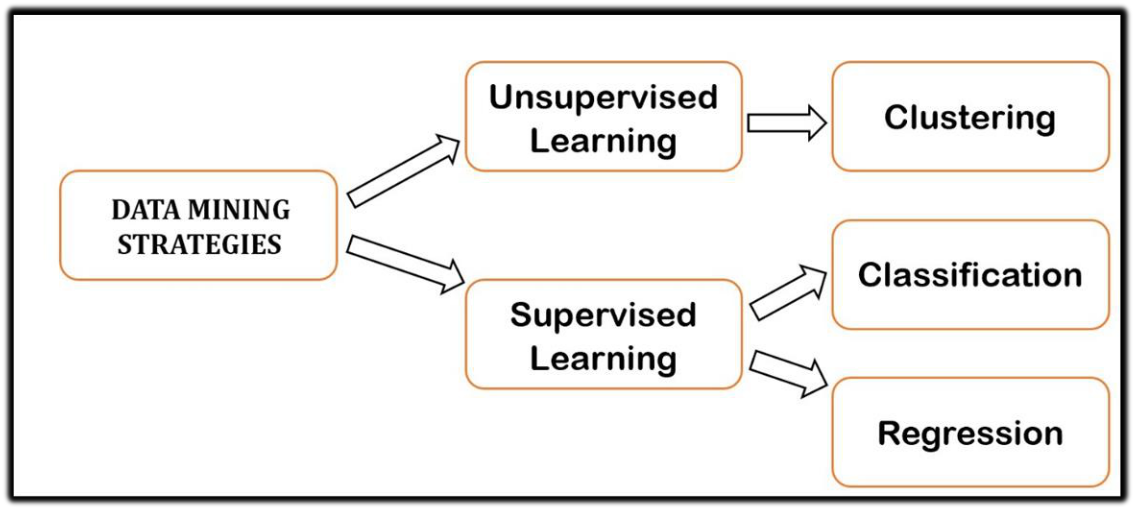
### Association

Association rule is such an acknowledging where there is no name on the data. Alliance is done to find the association between the limits. Association rule is illuminating not the insightful strategy, generally used to discover the relationship concealed in the immense datasets. The associations are by and large tended to in kind of rules or progressive thing sets. Association went under the Predictive Task.

### Clustering

Classification, Regression, Association all are part to deal with the data, like classification where the data is classified into different different classes, clustering on the other hand makes the clusters of data elements with the similarity measures obtained by calculating the distance between them. Clustering is the main focus of this paper there for it is discussed briefly. Clustering came under the Descriptive Task.

Fig 2 shows that data mining uses two methods: regulated and solo learning. In coordinated learning, a planning set is used to learn model limits while in performance learning no readiness set is used (e.g., k-infers gathering is independent). Each data mining technique fills an other need dependent upon the showing objective [1]. The two most essential showing objections are organization and assumption. Request models expect unmitigated names (discrete, unordered) while assumption models anticipate steady regarded limits.

 Figure 4. Strategies of Data Mining

**1.1.1 Data Mining and Knowledge Discovery**

Data mining is looking for covered, authentic, and perhaps accommodating models in huge enlightening assortments. Data Mining is connected to finding unsuspected/already dark associations among the data. Data Mining is focus piece of Knowledge Discovery Database (KDD). Various people treat Data Mining as an identical word for KDD since it's an essential piece of KDD measure. The uniting objective of the KDD collaboration is to eliminate data from data concerning tremendous informational collections. Data mining is the case of sorting out colossal dataset to recognize plan and develop relationship to handle issue through data assessment.

The terms Knowledge Discovery in Databases (KDD) and Data Mining are consistently used correspondingly. KDD is the route toward changing the low-level data into critical level data. Thus, KDD suggests the nontrivial extraction of inferred, effectively dark and possibly accommodating information from data in informational collections. While data mining and KDD are habitually treated as indistinguishable words yet in authentic data mining is a critical development in the KDD cycle. In the KDD cycle, the data burrowing methods are for removing plans from data. The models that can be found depend on the data mining tasks applied. Generally, there are two kinds of data mining endeavors: unquestionable data mining tasks that portray the general properties of the current data, and perceptive data mining tasks that attempt to do figures reliant on open data.

Data mining ought to be conceivable on data which are in quantitative, scholarly, or media structures. Data mining applications can use assorted kind of limits to assess the data. They join alliance (plans where one event is related with another event), course of action or way assessment (plans where one event prompts another event), request (ID of new models with predefined targets) and bundling (get-together of undefined or tantamount articles)

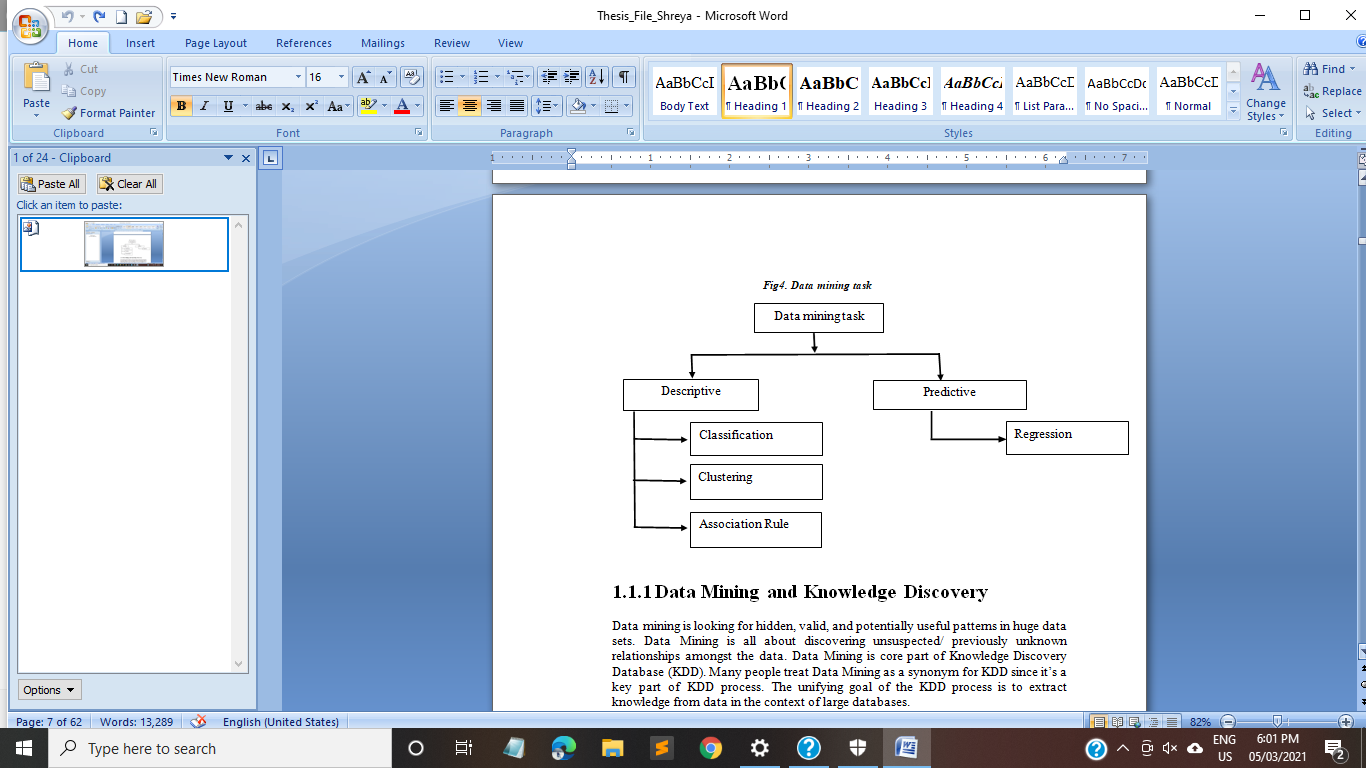


Figure 5.Data Mining Task

Data mining involves some of the following key steps [6]-

**(1) Problem definition**: The first step is to identify goals. Based on the defined goal, the correct series of tools can be applied to the data to build the corresponding behavioral model.

**(2) Data exploration**: If the quality of data is not suitable for an accurate model then recommendations on future data collection and storage strategies can be made at this. For analysis, all data needs to be consolidated so that it can be treated consistently.

**(3)Data preparation**: The purpose of this step is to clean and transform the data so that missing and invalid values are treated and all known valid values are made consistent for more robust analysis.

**(4) Modeling**: Based on the data and the desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighborhoods and clustering but also next generation techniques such as decision trees, networks and rule based algorithms. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.

**(5) Evaluation and Deployment**: Based on the results of the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration

Fig 1.1 shows Knowledge discovery as a process and consists of an iterative sequence of the following steps:

**1. Data Cleaning -** To eliminate commotion or unimportant information**.**

**2. Data Integration -** Where different information sources might be joined.

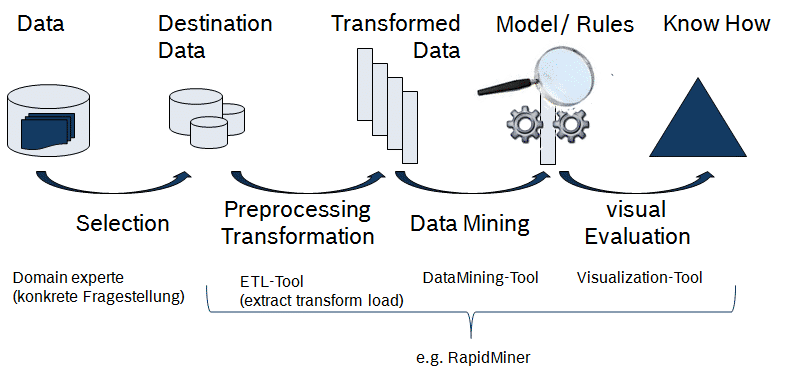
**3. Data Selection -** Where information pertinent to the examination task are recovered from the data set.

**4. Data Transformation -** Where information are changed or solidified into structures proper for mining by performing outline or collection tasks.

**5. Data Mining -** A fundamental interaction where keen techniques are applied to separate information designs.

**6. Pattern Evaluation** - To distinguish the really fascinating examples addressing information dependent on some intriguing quality measures.

**7. Knowledge Presentation -** information portrayal procedures are utilized to introduce the mined information to the client.



Although data mining has been around for more than two decades, its potential is only being realized now. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The prediction has attracted substantial attention given the potential implications of successful projecting in a business context.

There are two major types of predictions: one can either try to predict some unavailable data values or pending trends or predict a class label for some data. The latter is considered as classification[1]. Once a classification model is built based on a training set, the class the label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is, nonetheless, more often referred to as the forecast of missing numerical values, or increase/ decrease trends in time-related data. It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction[2]. Prediction can also be used for identification of distribution trends based on available data.

Data mining is a statistical process whereby data is taken from a data warehouse and compiled, organized and interpreted[3]. The purpose of data mining is to search in large amounts of data for valuable information and to understand the results. Data mining is the process of using data from a wide variety of data warehouses to uncover previously unexplored data and to make decisions. Researchers who wish to conduct data mining studies can choose from a number of computer programs, some of which require payment before use, while others are free[4].

## PREDICTION ALGORITMS:

The prediction as its name proposed, is one of the data mining procedures that discover the associations between self-sufficient components and associations between subordinate variables[6]. Which implies and huge computations to learn. Judicious Analytics is a piece of bleeding edge data examination that incorporates the use of various techniques, for instance, AI, real estimations, and other data mining strategies to measure future events reliant on genuine data[2]. Blameless Bayes is a direct yet amazingly inconceivable figuring for farsighted showing. The model is included two sorts of probabilities that can be resolved direct from your readiness data. The probability of each class and the unexpected probability for each class given each x worth.

**DECISION TREE**

A decision tree is an analytical ML representation that decides the purpose assessment of the latest example based on dissimilar factor and element principles acknowledged data. The inner module of a decision tree defines the several elements; the branches of all other elements. The elements, which help in analyzing the assessment of the reliant variable, are known as the self-sufficient variables in the particular.

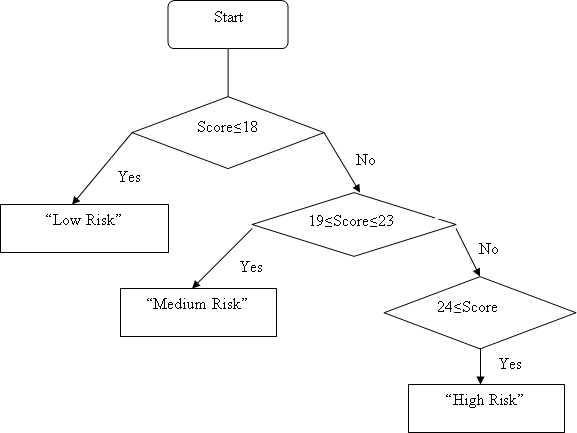


Figure 5. Decision Tree for Breast Cancer by risk factor score

**Decision Tree Classification Algorithm**

A Supervised learning performance used for mutual classification and Regression troubles, generally the favored intended to solve organization troubles. The classifier, anywhere inner nodes characterize all components, undergrowth symbolizes the decision convention and all leaf nodes characterize to the conclusion.

Here Decision tree, nearby two nodes, Decision Node & Leaf Node. Decision nodes use to accumulate any assessment of various branches, the production individual’s decisions as well as not to restrain any auxiliary nodes

**Root Node**: The root node is the decision tree that gets initiates and defines the intact dataset, which auxiliary split many more reliable sets.

**Leaf Node**: An ultimate productivity node cannot be differentiated auxiliary.

**Splitting**: The method of separating &distribute the decision node in the sub-nodes according to the specified circumstances.

**Sub Tree**: A tree that is formed through distributing the hierarchy.

**Pruning**: The method of removing the redundant undergrowth from the hierarchy.

**Parent/Child node**: The root node of the tree be identified as the parent node and residual nodes are identified as the child nodes.



### Figure 6. Illustrated example of binary decision tree

**Decision Tree Algorithm:**

**Step-1:** Create hierarchy with the root node, says S, which contains the entire dataset.

**Step-2**: Locate the finest feature from the datautilizings of element assortment compute.

**Step-3**: Separate the S into subsets that having probable principles for the best elements.

**Step-4**: Produce the decision tree node, which contains the best elements.

**Step-5**: Recursively, build the latest decision trees employing the subgroup of the dataset produced in step -3 maintain this method pending level reach then we cannot auxiliary sort the nodes and call the finishing node as a leaf node.

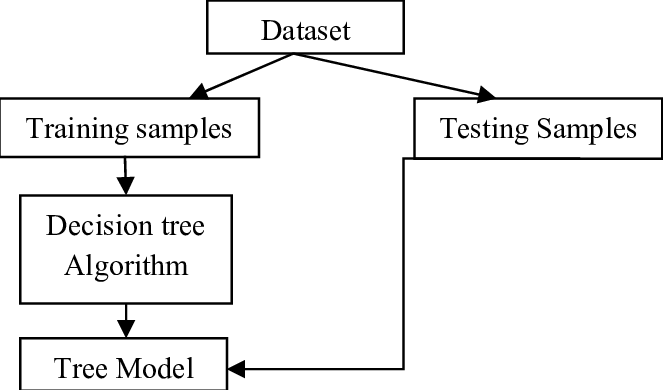
****

Figure.7 Workflow of decision tree

**Random Forest**

A supervised learning algorithm, random forest algorithm form decision trees take place statistics records and modules subsequently, achieve prediction from all of them and then go for the optimum outcome utilizing the range. It is an all-together process, that is improved than a single decision tree, which also diminishes the over-fitting by usual of the result. Random forests decision forests are a group learning technique for game plan, backslide and various tasks that work by building countless decision trees at getting ready time and yielding the class that is the strategy for the classes or mean/ordinary assumption for the individual trees.

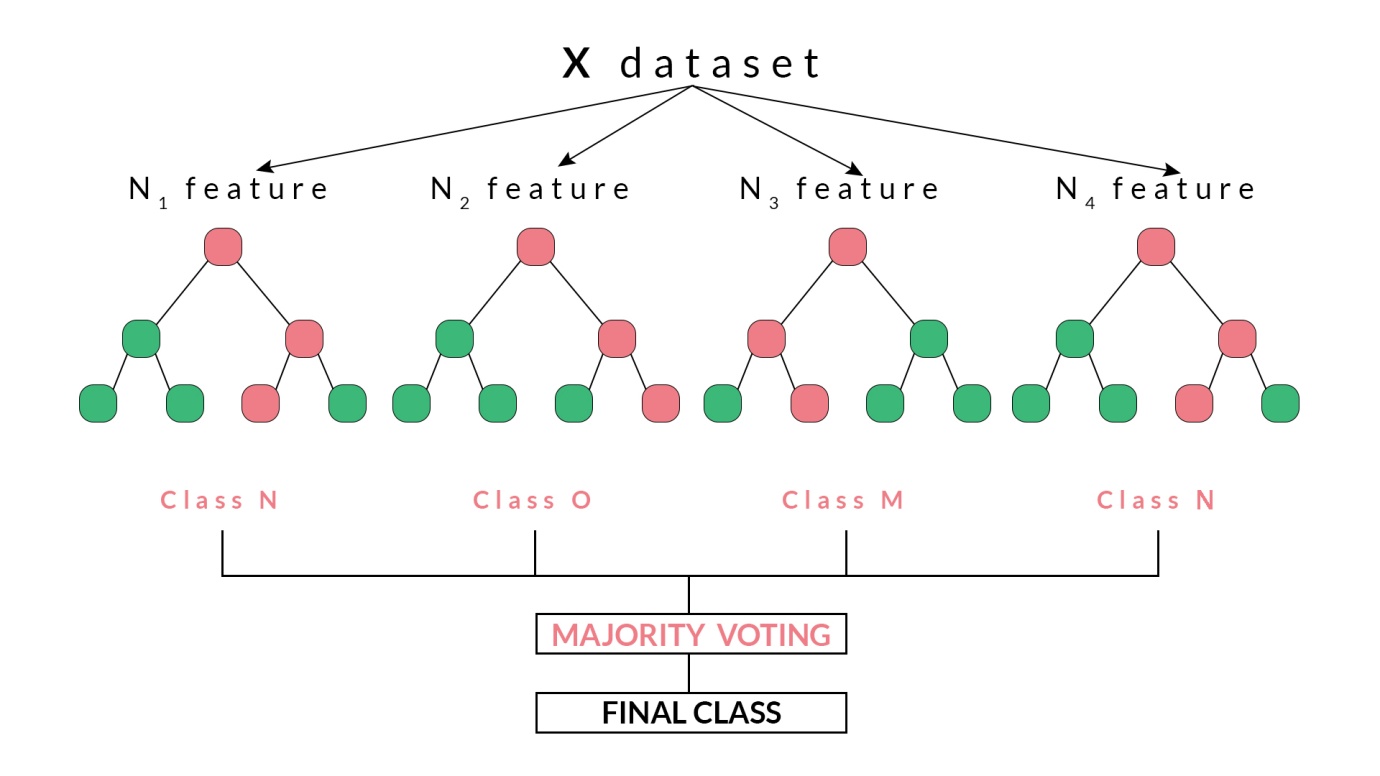


Figure 8. Random Forest Seletion Process Workflow

**Random Forest Algorithm:**

**Step 1**: Initially, create through the collection of accidental modules to a specified

dataset.

**Step 2**: Subsequently, the algorithm will assemble a decision tree on behalf of all modules.

**Step 3**: Third step, selection resolve be performed for each predicted conclusion.

**Step 4**: Eventually, choose the mainly selected prediction outcome as the absolute prediction consequence.

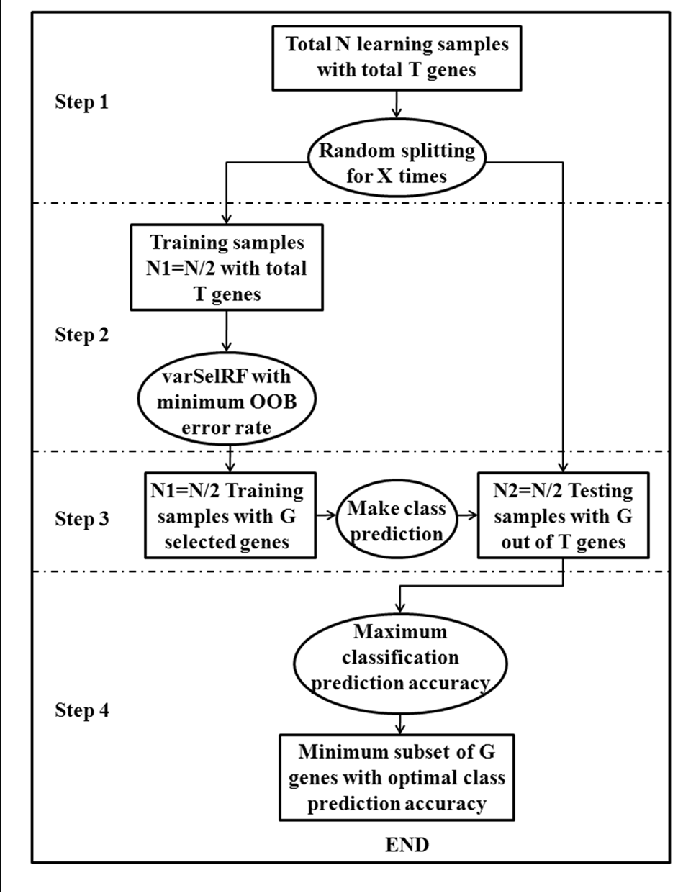
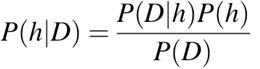


Figure 9. Workflow of Random Forest

**Naive Bayes**

Naive Bayes Bayes is an authentic request strategy reliant on Bayes Theorem. It is one of the least demanding oversaw learning counts. Blameless Bayes classifier is the fast, accurate and reliable count. Blameless Bayes classifiers have high precision and speed on immense datasets.

Naïve Bayes classifier expects that the effect of a particular component in a class is self-governing of various features. For example, a development up-and-comer is charming or not depending upon his/her compensation, past advance and trade history, age, and territory. Whether or not these features are connected, these features are at this point considered independently [7]. This assumption enhances estimation, and that is the explanation it is considered as naïve. This assumption that is called class unforeseen opportunity.



**•P(h):** the probability of hypothesis h being substantial (paying little regard to the data). This is known as the previous probability of h.

**•P(D):** the probability of the data (paying little notice to the hypothesis). This is known as the prior probability.

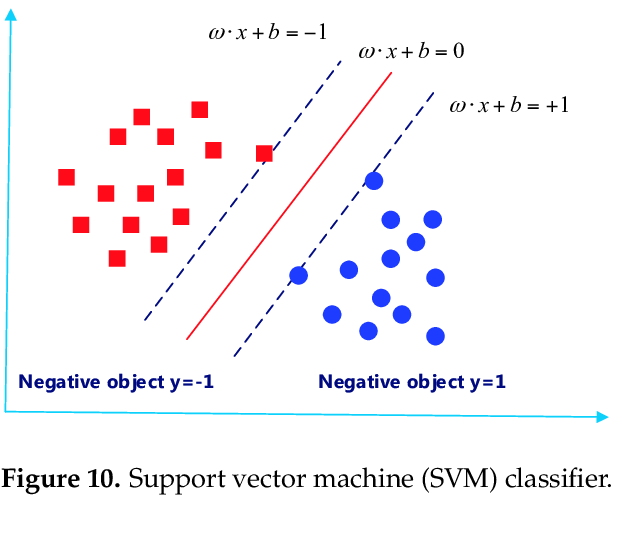
**•P(h|D**): the probability of theory h given the data D. This is known as back probability.

**•P(D|h):** the probability of data d given that the hypothesis h was substantial. This is known as back probability.

**Support Vector Machine**

SVM iis iused ifor iaccessing iinformation iand ipatternifor iregression ianalysis iand iclassification. iThe iobjective iis ito ifind imost iaccurate iclassification ifunction ito irecognize iindividuals ifrom ithe itwo iclasses iin ithe itraining idataset. iIt idepends ion imathematical ifunctions iand iutilized iin imodel icomplex, iand ireal iworld iproblems. iIt iworks iwell ion idata isets ithat ihave imany iattributes iand iensures ithat ithe ibest isuch ifunction iis ifound iby imaximizing ithe imargin ibetween ithe itwo iclasses. iIt idifferentiates idata iby isearching ithe ibest ihyper iplane ithat idivides iall idata ipoints iof ione iclass ifrom ithose iof ithe iother iclass.

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyper-plane.



**SVM Algorithm**

Import the dataset

• Explore the data to Fig out what they look like

• Pre-measure the data

• Split the data into characteristics and imprints

• Divide the data into getting ready and testing sets

• Train the SVM computation

• Make a couple of assumptions

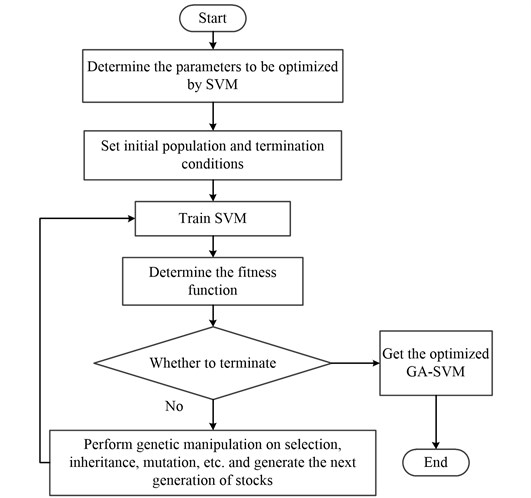
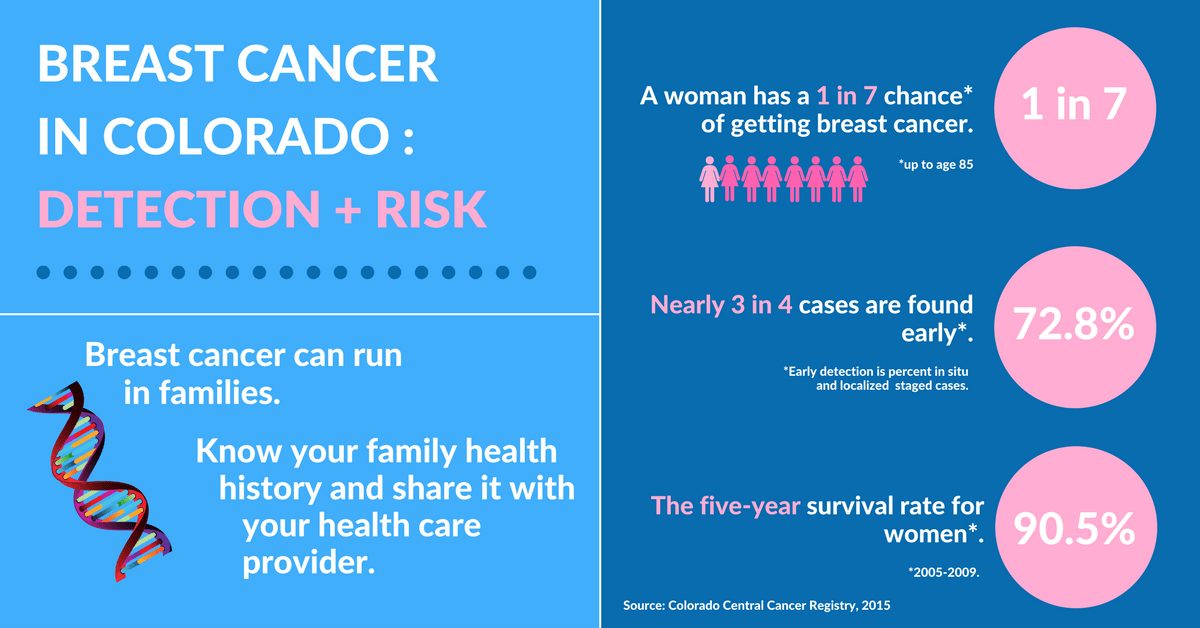
• Evaluate the delayed consequences of the count

Figure 11. Workflow of SVM

**1.3 Overview on breast cancer**

****

### Cancer is maybe the most broadly perceived disorders on earth that results in a lot of death[3]. Illness is achieved by uncontrolled advancement of cells in any of the tissues or parts of the body. Harmful development may occur in any piece of the body and may spread to a couple of various parts. Simply early area of sickness at the kind stage and contravention from spreading to various parts in compromising stage could save a person's life[8]. There are a couple of variables that could impact a person's tendency for infection. Tutoring is a huge marker of monetary status through its relationship with occupation and lifestyle factors[9]. Different assessments in made countries have shown that infection recurrence vacillates between people with different levels of guidance. A high pace of breast illness has been found among those with critical levels of guidance however an opposite alliance has been found for the recurrence of threatening developments of the stomach, lung and uterine cervix[3].

### Such differentiations in dangerous development chances related with tutoring in like manner reflect in the qualifications in lifestyle components and receptiveness to both characteristic and business related carcinogens[10]. This assessment portrays the connection between threat event model and risk levels of various components by creating a peril estimate structure for different kinds of sickness which helps in prognosis[6]. A colossal number of grandmothers, mothers, young ladies capitulate to breast threatening development reliably. The human body includes a considerable number of cells each with its own outstanding limit. When there is unregulated improvement of any of these cells it is named as dangerous development. In this, telephones disengage and grow uncontrollably, outlining a bizarre mass of tissue called as tumor[8]. Tumor cells create and assault stomach related, uncertain and circulatory structures upsetting the bodies' regular working. Regardless of the way that every single tumor isn't cancerous[8].

### Threat is portrayed by such a cell that is affected and more than 200 sorts of infections are known. This paper is based on Breast illness. Breast illness is the most broadly perceived sort of harmful development among females across the world [2]. Continuous years have seen a genuine improvement in perseverance rates for women with breast harm, which can be mainly credited to an expansive screening and redesigned treatment[10]. The new advances in data grouping and limit techniques have made it serviceable for various clinical associations and crisis facilities to keep huge proportions of data relating to their medicalrecords identifying with medication and signs of a disease[2]. Authoritatively, data mining is the route toward running fantastic counts on data to eliminate important information. The uses and potential outcomes of these methodology have found its expansion in clinical data.

### Expecting aftereffect of an affliction is a troublesome task. Data mining strategies will overall develop the figure segment. Robotized gadgets have made it possible to assemble immense volumes of clinical data, which are made available to the clinical assessment groups[2]. The results being a growing conspicuousness of data mining techniques to perceive models and relationship among enormous number of elements, which make it possible to expect the aftereffect of the disease using pre-existential datasets[10]. This paper presents the feasible helpful energies between data mining procedures and breast threat examination. Data mining system incorporates the usage of complex data assessment devices to discover effectively dark, genuine models and associations in gigantic instructive record. These devices can consolidate authentic models, mathematical count and AI methods in early distinguishing proof of cancer[3]. In portrayal learning, the learning plan is given a lot of gathered models from which it is needed to acquire capability with a technique for masterminding disguised models. In alliance learning, any relationship among features is searched for, not just ones that predict a particular class regard. In grouping, social affairs of models that have a spot together are sought[7].

### 

### Figure 12. Breast Cross territory (Bottom) Enlarged Breast (Top)

### In numeric assumption, the outcome to be expected is genuinely not a discrete class yet a numeric sum. In this examination, to describe the data and to mine progressive models in instructive assortment Decision Tree computation is used[6]. A decision tree is a stream chart like tree structure, where every inside center point means a test on an attribute, each branch tends to a consequence of the test and each leaf center holds a class label[2]. The top most centers are the root center. The trademark assessment of the data is attempted against a decision tree. A way is followed from root to leaf center, which holds the class estimate for that data. Decision trees can be conveniently changed over into gathering rules[2].

### This decision tree is used to make standard models in the dataset. The data and thing sets that happen routinely in the data base are known as progressive patterns[6]. The progressive models that is most basically related to express dangerous development types and are valuable in expecting the infection and its sort is known as Significant constant model. Using this colossal models made by decision tree the instructive assortment is packed in like way and peril scores are given. Breast infection has become the fundamental clarification of death in women in made countries. Breast sickness is the second most customary purpose behind dangerous development passing in women worldwide[9]. The high pace of breast harm in women has extended basically during the a few numerous years. In this paper we have discussed distinctive data mining approaches that have been utilized for early acknowledgment of breast cancer[5]. Breast Cancer Diagnosis is perceiving of great from unsafe breast hitches. We have pushed toward the finding of this disorder by using Data mining strategy.

### Data mining is a crucial development during the time spent data disclosure in informational indexes in which shrewd strategies are applied to remove patterns[6]. The best technique to diminish breast illness passings is to remember it earlier. This paper discusses the early acknowledgment of breast threat in three critical steps of choosing the breast illness. They join (I) combination of instructive assortment, (ii) preprocess of the enlightening list and (iii) course of action. Data mining and AI depend upon request which is the most key and critical task[4]. Various preliminaries are performed on clinical datasets using different classifiers and feature assurance strategies.

### 

### Figure 13. Types of Breast Cancer

Out of the two types of breast sickness, for instance perilous and thoughtful, the undermining tumor makes when cells in the breast tissue segregate and create without the customary controls on cell downfall and cell division. Industrialized nations like the United States, Australia, and countries in Western Europe saw the most important rate speeds of breast cancer[2]. In spite of the way that breast harmful development is the ensuing driving purpose behind sickness passing in women, still the perseverance rate is high at whatever point it is perceived early[10]. With early assurance, 97% of women make due for quite a while or more In the clinical consideration industry, it is basic to appreciate the consistent enhancements of such tumor[1]. There should be the availability of definite and exact data, so a model with exact model urges the experts to expect and examine the infection whether it is liberal or hurtful toward the starting stage. This will genuinely save time for the specialists and improve their efficiency[3]. This paper mainly inspects the probability to recognize the breast sickness condition whether it is altruistic or compromising even at early stage[5]. The figure condition relies upon the credits related to the breast illness.

**Table. 1 Different types of Cancer with %**

|  |  |
| --- | --- |
| **Different types of cancer in women** | **Percentages (%)** |
| Lungs and Bronchus | 26% |
| Breast | 14% |
| Colon & Rectum | 9% |
| Pancreases | 7% |
| Ovary | 5% |
| Non-Hodgkin Lymphoma | 3% |
| Leukemia | 4% |
| Uterine Corpus | 3% |
| Liver | 2% |
| Brain | 2% |
| Other nervous system | 2% |

**1.3.1 Breast Cancer Risk Factors**

Studies have shown that your risk for breast cancer is due to a combination of factors. The main factors that influence your risk include being a woman and getting[2]. Most breast cancers are found in women who are 50 years old or older. Some women will get breast cancer even without any other risk factors that they know of. Having a risk factor does not mean you will get the disease, and not all risk factors have the same effect. Most women have some risk factors, but most women do not get breast cancer. If you have breast cancer risk factors, talk with your doctor about ways you can reduce the chances and about screening for breast cancer[5]. Symptoms of breast cancer include a lump in the breast, bloody discharge from the nipple and changes in the shape or texture of the nipple or breast. Its treatment depends on the stage of cancer. It may consist of chemotherapy, radiation, hormone therapy and surgery.

**1.3.1.1 Uncontrollable Risk Factors**

1. **Age:** Age is the same for so many diseases; the risk factor for breast cancer goes higher as ever-increasing age. About seven out of ten persistent breasts.
2. **Family History**: Women who have diagnosed with breast cancer having a superior threat of emergent the syndrome.
3. **Genetics**: Concerning almost 10% of disease is consideration genetic, give rise by atypical genes conceded from parent to child.
4. **Race**: Ashen women be additional predictable to have cancer rather than other person, Hispanic, and Asian women.
5. **Being Overweight**: Overweight and heavy women comprise advanced threat of being examined by breast cancer compared to women who sustain a well mass, heaviness, especially later than menopause.
6. **Getting older:** The risk increases with age mainly breast cancers are diagnosed after age 55 years.
7. **Having dense breasts:** Intense & weighty breasts have more & more connective tissue than cause’s fatty tissue.

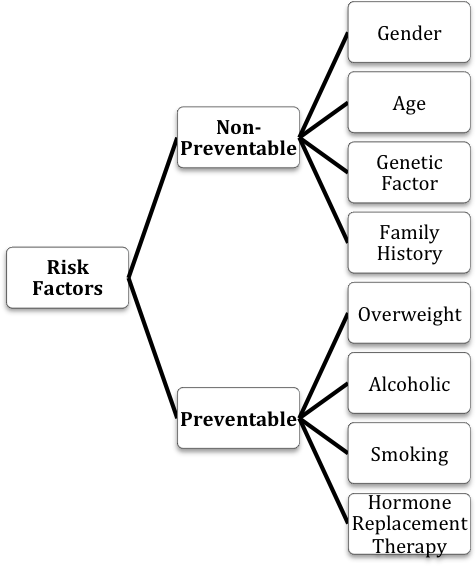


Figure 14. Risk Factors of Breast Cancer

**1.3.1.2 Controllable Risk Factors**

1. **Gender.** Being a woman is the principle threat factor for making breast illness. Regardless of the way that men can get breast harmful development, too, women's breast cells are ceaselessly changing and growing, for the most part as a result of the development of the female synthetic substances estrogen and progesterone. This activity puts them at significantly more genuine peril for breast dangerous development.
2. **Age.** Just becoming more established is the second greatest danger factor for breast disease. From age 30 to 39, the danger is 1 out of 228, or .44%. That leaps to 1 of every 29, or just shy of 3.5%, when you are in your 60s.
3. **Family history of breast cancer.** When a first-degree relative (mother, daughter, sister) who has had breast cancer, or you have multiple relatives affected by breast or ovarian cancer (especially before they turned age 50), you could be at higher risk of getting breast cancer.
4. **Race.** White women possibly will undoubtedly make breast harm than are Black women. Asian, Hispanic, and Native American women have a lower threat of making and passing on from breast threatening development.
5. **Radiation therapy to the breast.** Having radiation therapy to the breast area as a child or young adult as treatment for another cancer significantly increases breast cancer risk. The increase in risk seems to be highest if the radiation was given while the breasts were still developing (during the teen years).
6. **Breast cellular changes.** Unusual changes in breast cells found during a breast biopsy (removal of suspicious tissue for examination under a microscope) can be a risk factor for developing breast cancer. These changes include overgrowth of cells (called hyperplasia) or abnormal (atypical) appearance.
7. **Exposure to estrogen.** Because the female hormone estrogen stimulates breast cell growth, exposure to estrogen over long periods of time, without any breaks, can increase the risk of breast cancer[6]. Some of these risk factors are not under your control, such as: starting menstruation (monthly periods) at a young age (before age 12) going through menopause (end of monthly cycles) at a late age (after 55) exposure to estrogens in the environment (such as hormones in meat or pesticides such as DDT, which produce estrogen-like substances when broken down by the body)
8. **Pregnancy and breastfeeding.** Pregnancy and breastfeeding reduce the overall number of menstrual cycles in a woman’s lifetime, and this appears to reduce future breast cancer risk. Women who have never had a full-term pregnancy, or had their first full-term pregnancy after age 30, have an increased risk of breast cancer. For women who do have children, breastfeeding may slightly lower their breast cancer risk, especially if they continue breastfeeding for 1 1/2 to 2 years. For many women, however, breastfeeding for this long is neither possible nor practical.
9. **DES exposure.** Women who took a medication called diethylstilbestrol (DES), used to prevent miscarriage from the 1940s through the 1960s, have a slightly increased risk of breast cancer. Women whose mothers took DES during pregnancy may have a higher risk of breast cancer as well.

**Table.2 Risk scores for the attributes that represent the significant patterns.**

|  |  |  |
| --- | --- | --- |
| **Attributes** | **Values** | **Risk score** |
| Age | x<30  30 < x < 40  40< x < 60 | 3  4  5 |
| Living Area | Urban  Rural | 5  3 |
| Habits | Smoking Alcohol Chewing Hot beverage | 3  5  3  2 |
| Occupational Hazards | Radiation Exposure Chemical Exposure Sunlight Exposure  Thermal Exposure | 3  3  2  2 |
| Anemia | Yes No | 3  1 |
| Weight Loss | Yes No | 2  1 |
| Family History of Cancer | Yes No | 5  1 |

### 

**1.3.2 Symptoms of Breast Cancer**

People who have breast threatening development will from the start warning only a few of results and signs. Presence of these signs and results don't actually infer that the individual has breast threat. The huge signs of breast harm are as exhibited in Fig.5 and few are depicted in this sub portion.

1. **Breast Lump:** It is the most broadly perceived symptom of breast danger. Simple, hard mass with inconsistent edges fringes to be hazardous anyway breast threat can moreover be sensitive, fragile or changed.
2. **Change in Nipple:** Nipple withdrawal (turning inward) or areola discharge (other than breast milk) moreover can be a huge sign of breast danger.
3. **Skin Dimpling:** Puckering of skin on the breast is similarly seen as one of the results of breast danger.

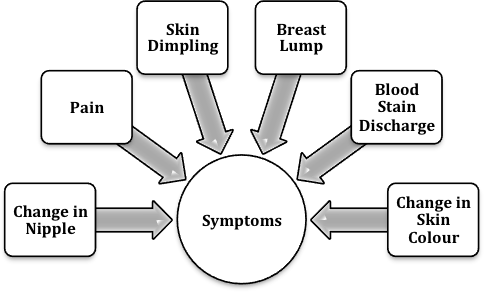


Figure 15. Symptoms of Breast Cancer

# CHAPTER-2 LITERATURE REVIEW

**2.1 Research Work Done By Various Experts on Breast Cancer Prediction System**

Data mining is so versatile in behavior that the real world problems can be solved using the data mining process, with just a collection of data and with a proper implementation model, one can find the solution to the problem and can innovate the existing process too. Not just the innovation the existing process of finding the solution could also be enhanced by making changes in the algorithms and models of implementation. Here some of the work done on the various fields for tackling the real world problem by using data mining process and some enhancement in the existing models and algorithms of data mining, specifically the clustering part is reviewed. The research works done by scholars are discussed here. A Literature survey was done for the project. Several research papers were studied and some concepts for each paper were declared and incorporated in this research paper.

Mücahid Mustafa Saritas et al[1] demonstrate comparing the precision of the division, defined a principally distinct from correctness of 86.95 with ANN and 83.54 with Naïve Bayes using continuously and manages the arguments beginning the patients, therefore significant with the purpose of the algorithms may well be worn for premature breast cancer discovery by by means of ANN and Naïve Bayes base on biosignatures. Glucose, Resistin, Age, BMI, HOMA, Leptin, Insulin, Adiponectin, MCP-1.

David Goerlitz et al [5]demonstrated to be linked through the pathogenesisof breast cancer and submit as biosignatures for detection and prognosis. Relative genomic hybridization reads have been accounted for IBC rather than non-IBC breast tumors.

Charles Edeki et al [7] defines with the aim of any of the data mining and statistical learning algorithms in the breast cancer datasets out performed on supplementary approach in such a manner that acknowledged the finest algorithm, none of the algorithms accomplish defectively as to be impassive by the potential calculation representation in the breast cancer.

Layla Abdel-Ilah et al [8] develops that the decision tree algorithms are well-built candidates for breast cancer detection. Random Forest proved to be the most appropriate decision tree algorithm for breast cancer categorization because it achieved the highest accuracy in both distributions. Time-series analysis method which belongs to the data mining. all the attributes are individually considered for the analysis process and to predict the upcoming values and stand of those parameters compared to the past result, the predicted result shows the increase in PM10 is 45.90% for upcoming years, Ozone is increasing 88.3% from 2011-2018, as well as the CO will increase 68.83% in forthcoming year.

Amorim and Mirkin et al [9] develops that there is an extensively seen batching gadget that is used for cutting edge mechanics, ailments and electronic thinking application purposes. Here k is a positive entire number tending to the amount of packs. The pre-arranged count with the assessment of k identical to 2. This locations there is two bundles where one gathering contains relevant data to Cancer and another contains remaining data that suggests non critical data. This is the most basic and principal subjects of data mining. It is considered as the standard data mining issue that hopes to find the normal things or models from the data conveyance focus. There are different sorts of counts, used to mine intriguing persistent models from data bases like association rules, bundles, portrayals and associations.

Majoor et al [6] taken apart the age peril factor in every country. The breast dangerous development is clearly relating to the age peril factor. The shade of the women moreover checks out causing breast harm in women. The effect of Family history, and communicated that the women has high risk probability of getting breast threatening development when anyone of her close by blood association having this affliction.

Kamińska et al [3] proposed distinctive threat factors which causes breast harm. This investigation tracked down that the women have high risk of breast harmful development when they are with united breast tissue. In case there is unexpected change in the size of breast tissue, it may provoke breast harmful development. This work moreover uncovered that the ones who defer the essential youth still to age 30 has high peril. This investigation also obliged the pregnancy women not to consume the diethylstilbestrol medicine because of risk of breast harmful development.

Jayalakshmi and Santhakumaran et al [8] anticipated that a generally perceived proper meaning of information mining can be characterized as "Information mining is the non-insignificant extraction of verifiable beforehand obscure and possibly valuable data about information". Information mining Has a few fields to examination of information like arrangement, grouping, connections, affiliation rule and so forth and has been utilized seriously and widely by numerous associations. Also, In-medical care, information mining is getting progressively well known. Information mining gives the approach and innovation to examination the valuable data of information for dynamic. Information pre-preparing is an imperative undertaking of information mining. It fundamentally utilized for making examination suitable and furthermore making information proper for bunching by dodging copy records and adding missing information as per past recorded information. The principle advantages of information pre-handling decreases memory.

V.Krishnaiah et al [6] explain that a sample of lung cancer disease calculation organizations by means of data mining classification methods. The indispensable illustration all the direction of analyzing the long-suffering, Lung cancer convoy by the IF-THEN law, The Decision Trees & Neural Network. On behalf of the identification of Lung Cancer Disease Naïve Bayes find the optimum fallout and ensue best Decision Trees.

Jyotismita Talukdar et al [4] predicted the be revealed of breast cancer most competently, arranged the main elements of Breast Cancer for fast, efficient, robust and correct analysis so that it can be correctly medicated point in time. This paper fundamentally talks about the likelihood to recognize the breast disease condition whether it is kind or threatening even at beginning phase. The expectation condition depends on the ascribes identified with the breast malignancy. There are 10 ascribes in the informational collection utilized in this paper. There are three significant advances that have been utilized in this paper for example assortment of datasets, information preprocessing and arrangement. This paper clarifies the different periods of information mining that is performed on the dataset. We have utilized WEKA as an information mining device.

# Table.3 Comprehensive Review on Breast Cancer Prediction Exhausting Data Mining Techniques

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Technique** | **Author & Ref**  **No.** | **Ye ar** | **Description** | **Advantages** | **Disadvantages** |
| Performanc e comparison of Naïve  bayes and J48  decision trees process. | William s | 2015 | J48 choice bushes exhibited a better accuracy with lesser mistakes expenses related to that of the naïve bayes’ method through the usage of the usage of  17 attributes with LASUTH breast most cancers fact[2]s. | 1. Great outcomes in assessments. 2. J48 selects most discriminatory features. 3. High tolerant of noisy data. | 1. Computational cost of J48 is high. 2.J48 reduces unambiguousness to the related model. |
| Mining the microarray structure of genes related  information | Jauhari and Rizvi | 2014 | SVM classification based on Gene importance and enrichment ranking techniques[5]. | 1. High accuracy because of mining gene expression data. 2. Diagnosis of complex proteins in gene data. | 1. It is difficult to collect gene expression data. 2. 458 genes are identified yet which   cause breast cancer. |
| Classificati on based on Taxonomy of patients history (Taxofinder  ) | Kang and Haghigh | 2016 | Taxonomy learning approach collects domain related concepts and represent it via graphs. Then analytic algorithm learn the taxonomy related classification[3]. | 1.This method has no certainty predefined lexico syntactic designs that dependably require additional examination | 1. The dependency of events and   attributed are uncategorized.   1. Very complicated. |
| probability of breast cancer Recurrence used for classificatio n  algorithms. | Pritom and Sabab | 2016 | Ranker algorithm is used to foretell the recurrence probability of breast cancer. The best feature selection approach is used to pinpoint the recurrence possibility[7]. | 1. Extracts meaningful technique for efficient classification | 1. This is not adoptable for real time database implementation. |
| DNN and multi criteria decision making combined with Analytical  Hierarchy Process | Soliman | 2016 | The output of DNN Multilayer Perceptron layer is integrated with AHP which works based on multiple criteria decision[4]. | 1. Accuracy, efficiency, less time for classification tasks. 2. Strong against noisy data. 3. useful for large and small datasets | 1. DNN over fit with attributes and mislead and can cause health complication. 2. need of huge memory.   . |
| K-means algorithm is integrated with PSO | Salma and Doreswa my | 2016 | Select top N strong features from high dimensional breast cancer dataset for better classification of breast cancer disease[10]. | 1. Very helpful approach for attribute selection. 2. Robust and balanced than conventional techniques. | 1. Not applicable for small data. |
| NN used for classificatio n using genome variability. | Boutorh and Guessou m | 2015 | The dimensionality feature reduction of SNP was achieved through hybrid mechanism of Association Rule Mining (ARM) and Neural  Networks[6]. | 1. Association rule method is suitable for extracting hidden information from dataset. 2. This hybrid approach is useful to discover   relationship patterns in complex SNP data. | 1. Very easy to understand. But it is time consuming. 2. It is not suitable for large dataset. |
| Biclusterin g and BP neural network algorithm | Chen | 2016 | Ultrasound imaging features are biclustered and BP NN algorithm was applied to classify the breast tumors[4]. | 1. Effective Treatment can be achieved. 2. A robust approximator function classifies effective prediction. 3. adaptable to new attributes of dataset. | 1. Time complexity of algorithm is more. |
| multivariat e statistical approach combined with AI | Jhajhari a | 2016 | The features are extracted using PCA algorithm and the resustant features and relevant data is trained using ANN approach[3]. | 1. PCA involves in dimensionality reduction when there is need. 2. When PCA works along with ANN, the   resultant results will be effective. |  |
| Random tree classifier and ANN | Kaushik and Kaur | 2016 | Features extracted from the mammogram dataset are used to predict the biopsy results by using random forest classifier and NN[9]. | 1. Better performance and reduced dimension. 2. Highly accurate classification. | 1. It’s more complex. 2. It’s more difficult to implement.  3. High computation cost |
| Gene expression data are  used for classificatio n using Elitism PSO | Nagpa and Shrivas | 2015 | EPSO is feature selection based classifier based on lazy IBK[10]. | 1. Exact selection of gene makes easy prediction of breast cancer. 2. High dimensional data can be mined effectively. | 1.Not applicable for small data |
| Sequential rule and pattern matching | Cheng | 2017 | Wealthy set of sequential guidelines are hired for the type set of regulations[2]. | 1. Minimal help recollect is used to healthy the sequential patterns. | 1. Inconsistent and Incomplete records 2. bizarre visits of patients influences sequential approach 3.Imbalance issue   4.Post-analysis issue |
| Ensemble approach for feature selection  mechanism | Dhakate | 2015 | The ensemble approach is used for feature selection and the accuracy of results are compared without using the technique with  same dataset[6]. | 1.It provides better accuracy when compare with other techniques | 1.Relatively higher computation cost |
| Compariso n of J48, NB, SMO,  IBK with three dataset. | Salama | 2012 | The combined classification of MLP and J48 along with PCA is greater than other  classifiers employing WBC data set[10]. | 1. This work examined various combination of classifiers. | 1. Some  amalgamation of classifier over fit with data set and some will be under fit with data set. |
| En based clustering and fuzzy logic implementa tion  combined with PCA | Nilashi | 2017 | Fuzzy rules are applied with the EM based clustering Mammographic mass datasets. The PCA supports to avoid multi-collinearity in dataset[3]. | 1. Prediction accuracy was high. | 1. Practical issues at the time of implementation. |
| fuzzy- rough nearest neighbor method. | Onan | 2015 | Fuzzy-tough instance choice take away the noise of records. Then consistency-based totally totally sincerely feature choice is carried out. Fuzzy-hard nearest neighbor set of guidelines is used for the elegance[7]. | 1.A combined approach 2.Dealt the uncertainty data very well   1. Assisted with noisy data. 2. works fast for small training sets | 1. It suffers with large training data. |
| GA is used for FS and multiple classificatio n approach is employed. | Alicovic | 2017 | Genetic algorithms are used for putting off of useful and amazing abilities. Discrete person and multiple classifier coordination have been used to make specific technique for breast most cancers cataloguing[8]. | 1. Individual accuracy and diversity. 2. Very sensitive and high accuracy. 3. Subset relationship is maintained. | 1. More time complexity. 2. individual observation   might be a member of no  classes or too many classes |
| Foraging lung cancer using Ant colony optimizatio  n | Christop her | 2015 | The various attributes are taken to compute the multi objective function and determines the increased risk of disease[3]. | 1. Changes of attributes can be considered. 2. Multi objective function involved in prediction | 1. Attributes cannot be increased much. 2.Computational complexity. |
| prognostica ting breast cancer using random forest  classifier | Nguyen | 2013 | Hybrid technique of random forest classifier and function desire method classifies the breast most cancers records[3]. | 1. Perform good with large dataset. 2. Good attribute selection approach. 3. It measure importance of each feature. | 1. smaller groups are Favored over larger groups. |
| FS  approach based on  Filter and Wrapper | Tomar | 2015 | The filter approach filters the features based on the ranking and wrapper method validate it with SVM and cross validation approach[4]. | 1. Non-regular data can be handled easily. | 1. High algorithmic complexity. 2. More memory requirement. |
| K-means clustering combined with SVM | Zheng | 2014 | The similarities of breast tumors are calculated using K-SVM approach. K-means detects the hidden information of tumors and SVM classifies the new data  with the existing pattern[2]. | 1. This method applied pattern matching system using SVM. 2. The required attributes only can be selected from   large attribute set. | 1. Accuracy is less. |

**Table.4 Comparative Detailed Study based on the previous research papers**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Publication** | **Method** | **Cancer type** | **No of patients** | **Type of data** | **Accuracy** | **Validation method** | **Important features** |
| Ayer T | ANN | Breast cancer | 62,219 | Mammographic, Demographic | AUC = 0.965 | 10-fold cross validation | Age, mammography findings |
| Waddell | SVM | Multiple myeloma | 80 | SNPs | 71% | Leave-one-out cross validation | snp739514, snp521522, snp994532 |
| Listgarten J | SVM | Breast cancer | 174 | SNPs | 69% | 20-fold cross validation | snpCY11B2 (+) 4536 T/C snpCYP1B1 |
| Stajadinovic | BN | Colon carcinomatosis | 53 | Clinical, pathologic | AUC = 0.71 | Cross-validation | Primary tumor histology, nodal staging, |
| Exarchos K | BN | Oral cancer | 86 | Clinical, imaging tissue genomic, blood genomic | 100% | 10-fold cross validation | Smoker, p53 stain, extra-tumor spreading, TCAM, SOD2 |
| Kim W | SVM | Breast cancer | 679 | Clinical, pathologic, epidemiologic | 89% | Hold-out | Local invasion of tumor |
| Park C | Graph-based SSL | Colon cancer, | 437 | Gene expression, PPIs | 76.7% | 10-fold cross validation | BRCA1, CCND1, STAT1, CCNB1 |
| Park C | SSL Algo | Breast cancer | 374 | Clinical, population | 80.7% | 10-fold cross validation | BRCA1, CCND1, STAT1, CCNB1 |
| Tseng C-J | SVM | Cervical cancer | 168 | Clinical, population | 68% | Hold-out | pathologic\_S, pathologic\_T, cell |
| Eshlaghy A | SVM | Breast cancer | 547 | Clinical, pathologic | 95% | 10-fold cross validation | type RT target summary  Age at diagnosis, age at menarche |

# CHAPTER-3

**PROBLEM IDENTIFICATION AND PROPOSED WORK**

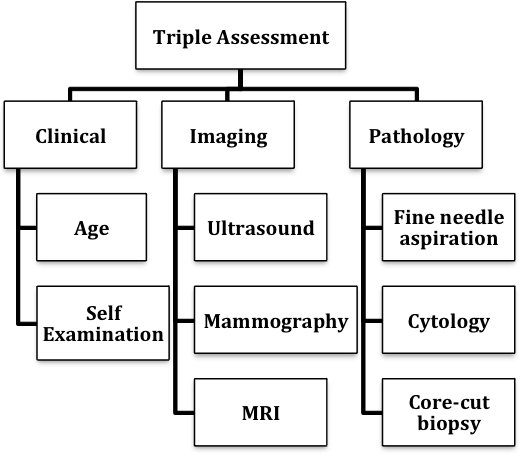
## 3.1 INTRODUCTION:

In this chapter the overview of problem identification and proposed work is been given. With the experiments, literature survey, study of research paper and implementation of various algorithms under the prediction methods[2]. Very few systems use the available clinical data for prediction purposes and even if they do, they are restricted by the large number of association rules that apply[9]. Diagnosis of the condition solely depends upon the Doctors' intuition and patient's records[7]. There are many ways that a medical misdiagnosis can present itself[11]. Whether a doctor is at fault, or hospital staff, a misdiagnosis of a serious illness can have very extreme and harmful effects.[2]

## PROBLEM STATEMENT:

**3.2.1 Existing System**

Malignant growth conclusion situation in medical clinics is manual. For instance when a patient is enlisted he/she needs to go through radiology test measure for example X-beams, CT or MRI. Radiologist gives his comments on the test report[7]. After this interaction a specialist audits the X-beams/CT/MRI and gives his remarks[3]. In certain kinds of malignant growth the analysis depends on an official conclusion by the specialists for example breast and lung cancer[5], however in different kinds of disease like carcinoma some different tests are additionally required like biopsy. In a manual framework the radiologist and the specialist analyze cancer[1]. This interaction is delayed as after the radiologist's survey the specialist needs to audit likewise and give his/her comments lastly tell if the malignant growth is available or not[6]. The need is to mechanize this interaction to make the malignant growth finding effective and quick with the utilization of cutting edge technology[6]. Test information under investigation is quality articulation information of malignant growth type leukemia.

Figure 16. Diagnostic Tests for Breast Cancer 

**3.2.2 Drawback of Existing System**

**Late detection**

Detection is not possible at earlier stage which affects the quality of service provided to patients.

**Accuracy Issues**

An electronic framework alone doesn't guarantee precision, and the distribution center information is just pretty much as great as the information section that made it.

**Automation**

The framework isn't completely computerized; it needs specialists for full analysis for genuine patients.

**Time Consuming**

In the current framework, pragmatic utilization of different gathered information is tedious.

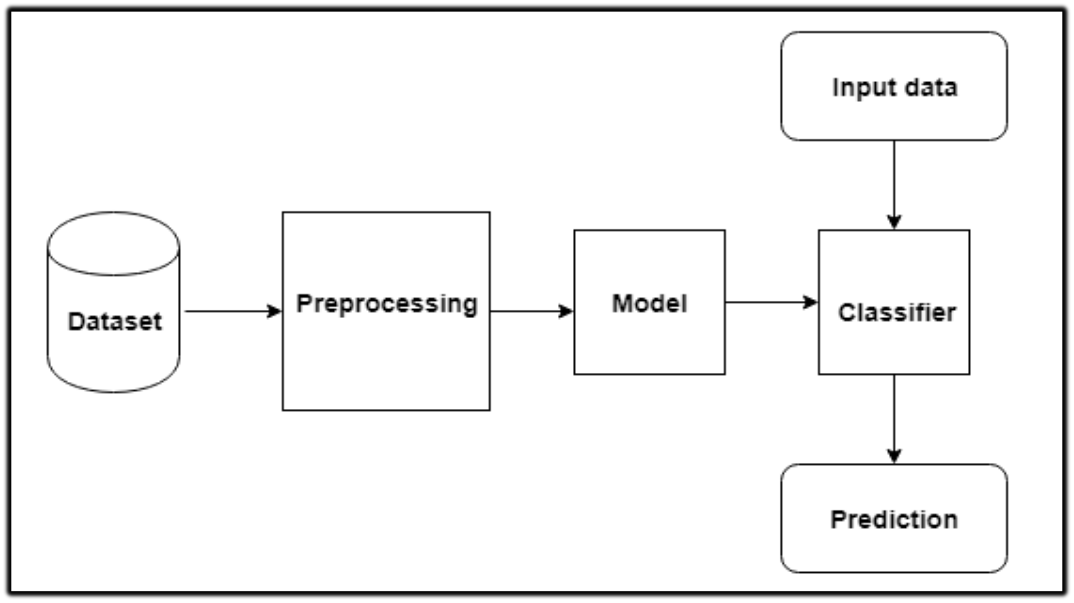


Figure 17. Simplified Representation of Proposed System

## 3.3 EXISTING METHODOLOGY:

In the wake of being resolved to have breast threat, a patient will have diverse treatment decisions depending upon the period of breast sickness and the expert treating. Different factors are turned out extraordinary all around treatment, including the sort, patients' age and general prosperity. Fig 16 shows the stream layout tending two or three arrangements of choices available for a patient in the wake of being resolved to have breast infection. Breast sickness can be found after signs appear, anyway for certain women early breast threat have no signs. Thus performing screening test before any results make is amazingly principal. Early recognizable proof of breast harmful development (before presence of signs), in the limited stage, grows the 5-year perseverance rate to 98 %.The tests that can be performed are gathered into a triple evaluation standard as shown in Fig 18. Breast threatening developments will overall be greater and bound to be spread past the breast if appearances are otherworldly. Strangely, breast threat that is found during screening tests is more disposed to be more humble and bound to the breast.

**Training Data (missing values, null values, irrelevant features**)

**Train an Algorithm**

**Classification & Prediction Algorithms (Random Forest, Logistic Regression, SVM)**

**Model is produced**

**Inaccurate Model/Inefficient Model**

**Inaccurate Model**

**Wrong Diagnosis / Prediction of Cancer**

Figure 18. Problem Statement

## PROPOSED WORK:

The Breast Cancer Prediction System can be performed by following the procedure which is represented in Fig.1 which indicates the research methodology for deploying a classification model required for the prediction of the breast cancer in patients. The model structures a central method for doing the breast cancer related disease prediction by utilizing Naive Bayes classifier, Random Forest and Support Vector Machine algorithms for prediction. Firstly, data is preprocessed by removing inconsistent, redundant data then a proposed model should be trained with the clinical datasets and the prediction is made. In this phase, the training stage incorporates preparing the three classification algorithms namely Naïve Bayes, Random Forest and Support Vector Machine utilizing the coronary illness dataset and a prediction model is built.

The overall process proposed prediction system is based on the following three steps:-

* **Data collection**
* **Data pre-processing and**
* **The classification and prediction of data.**

**3.5 PROPOSED ALGORITHM**

* + 1. **Installation**: Install Scientific Python Development Environment, Jupyter Notebook on the system and then create a workspace directory for a machine learning code and datasets.
    2. **Import the packages**: Import all the necessary packages of Python such as Pandas, NumPy, Matplotlib and Scikit-Learn according to the requirements.
    3. **Loading of dataset**: Load the Indian Liver Patients dataset from UCI repository in CSV (comma separated value) format using Pandas in Python workspace, Jupyter Notebook.
    4. **Data Analysis**: Apply info ( ), describe ( ), value\_counts ( ), head ( ) and tail ( ) methods in order to analyze the data and to get a quick description of the data.
    5. **Data Preprocessing**: The stored data contain some missing values so it is necessary to remove it by using dropna ( ), drop ( ) or isnull.sum( ) method. Also, replace ( ) method can be utilized to replace missing values with the new values.
    6. **Data Formatting**: Apply data info ( ) method for the detail of categorical data to quantitative data.
    7. **Data Visualization**: Apply sns.swarmplot ( ), grid( ), and heatmap ( ) methods to show the relationship among several variables and for the proper visualization of data.
    8. **Splitting of Training and Test data**: Apply train\_test\_split ( ) method in order to split the training and test data from the whole dataset.
    9. **Training the Model**: Apply fit ( ) method to train the model using training data.
    10. **Apply Classification Algorithm**: Random Forest algorithm is applied on the dataset which is collected from UCI repository.

.

* + 1. **Feature Importance**: To access a feature importance score for each feature, the feature\_importances\_variable is used.
    2. **Evaluation of Accuracy**: Apply accuracy\_score ( ) method to evaluate an accuracy of a Random Forest classifier after the data preprocessing is done.
    3. **Model Generation**: Finally, an accurate model is generated that can be utilized to perform several tasks.

**Cancer Dataset (csv file**)

**Load Data in Jupyter Notebook**

**Data Analysis**

**Data Preprocessing**

**Data Visualization**

**Splitting of Training & Testing Data**

**Training the Model**

**Apply Prediction Algorithm**

**Evaluation of accuracy, training & Testing Time**

**Cross Validation (k-folds)**

**Final Model Generation**

Figure 19. Proposed Methodology

## 3.6 DATA SET DESCRIPTION & DETAILS:

The database for this research work is taken from UCI (University of California) repository. The database records were divided into seven rows: mean radius, mean perimeter, mean area etc. Cleveland Clinic Foundation dataset of UCI repository is used as training dataset and testing dataset. This database contains attributes including ID number, Diagnosis (M = malignant, B = benign). Ten real-valued features are computed for each cell nucleus 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), symmetry, fractal dimension (coastline approximation), one attribute as output and the remaining are input attributes. Missing, redundant, inconsistent data have all been resolved by assuming the issues. The target class of the dataset denotes the presence of heart disease with a yes or no in Boolean values Data sets are the essential part to carry out any machine learning module implementation to obtain the results and to check whether the machine learning method is working properly or not. The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository [10] is used to differentiate benign (non- cancerous) from malignant (cancerous) samples. Table 2 shows a brief description of the dataset that is being considered.

### Table.5 Description of Breast Cancer Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Columns | Rows | Type |
| Wisconsin Breast Cancer  (Original) | 33 | 569 | 2 |

## 3.6.1 BREAST CANCER DATA:

In this dataset it contains following attributes which helps in obtaining highest performance and accuracy along with lesser training time and testing time. ID number,  
Diagnosis(M = malignant, B = benign), Ten real-valued features are computed for each cell nucleus 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), symmetry, fractal dimension.



Figure 20. Breast Cancer Dataset

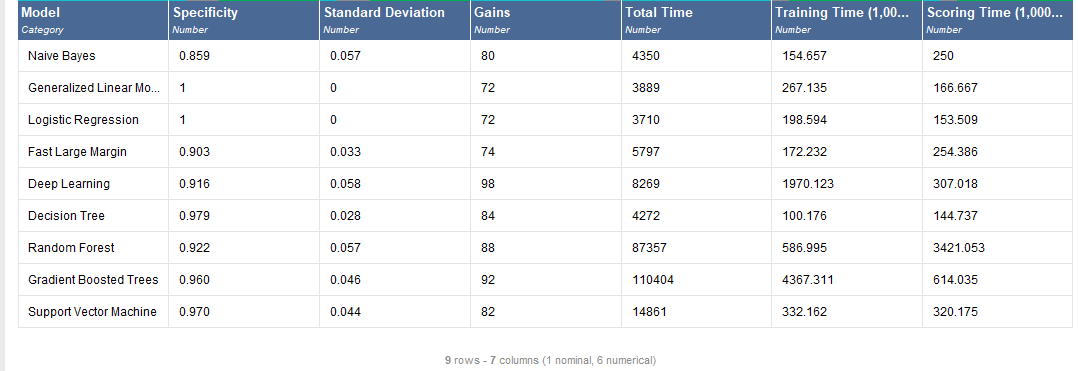


Figure 21. Dataset outcomes after applying RapidMiner

# CHAPTER-4 IMPLEMENTATION ENVIRONMENT

In this chapter the implementation environment setup done to carry out the processing on data set is discussed and with that the work done for the survey work and its setup is also discussed.

## 4.1 ENVIRONMENT SETUP FOR PROCESSING:

* + - Windows 10 Home (64 bit)
    - 4 GB RAM
    - Jupyter notebook (IDE)
    - Anaconda Navigator Distribution (for environment setup of Jupyter notebook)
    - RapidMiner Tool

## 4.2 LANGUAGE/TOOL USED:

* + - Python 3.7
    - RapidMiner Tool

## 4.2.1 RapidMiner Tool:

Rapid Miner is a data science software information that gives an acclimatize climate to information examination, information readiness, AI, profound learning, text mining, and prescient investigation. It is utilized for business and business applications and for research, schooling, preparing, fast prototyping, and application advancement and supports all means of the AI cycle including information planning, results representation, model approval and streamlining.Rapid Miner in this topic analyzes medical datasets and calculated the interpretation of data mining techniques applied to these datasets. The certain data mining methods are presented with their detailed description of parameters which used for experiment. Basically, the visualization of each algorithm’s experiment shown for medical datasets.



Figure 22. The RapidMiner Tool Overview

There are various advantages of using Weka tool such as:

• Portability, since it is completely carried out in the Java programming language and subsequently runs on practically any cutting edge processing stage.

• A far reaching assortment of information pre-handling and demonstrating procedures.

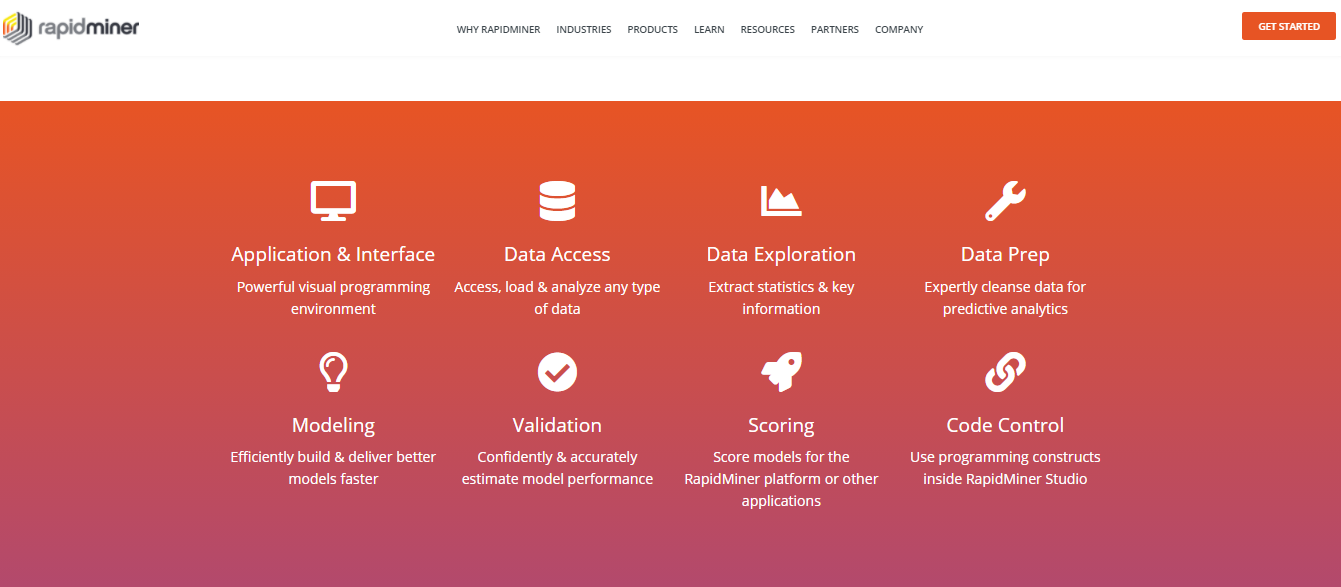
• Ease of utilization because of its graphical UI.

Figure 23. Workbench of Weka tool

## Jupyter Notebook:

Jupyter notebook is the IDE used for the python programming language. Jupyter notebook is very popular when comes to the python programming due to its easy to access GUI and the clarity of codes written in it.

Jupyter Notebook is not like any other IDE, there are individual cells for writing the code and it also has many functionality of commenting, making headings and sub-headings etc., Jupyter notebook is generally used for the machine learning code writing and for the scientific development of code due to rich in libraries provided by the python programming language.

Jupyter Notebook is the open source IDE and the libraries it support has a vast collection from normal level to very high end scientific libraries. A snapshot of the jupyter notebook is given below with some lines of code for better overview of it.

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. Project Jupyter is a non-profit, open-source project, born out of the Python project  in 2014 as it evolved to support interactive data science and scientific computing across all programming languages. Jupyter will always be 100% open-source software, free for all to use and released under the liberal terms of the Modified BSD license. Jupyter is developed in the open on GitHub, through the consensus of the Jupyter community. For more information on our governance approach.

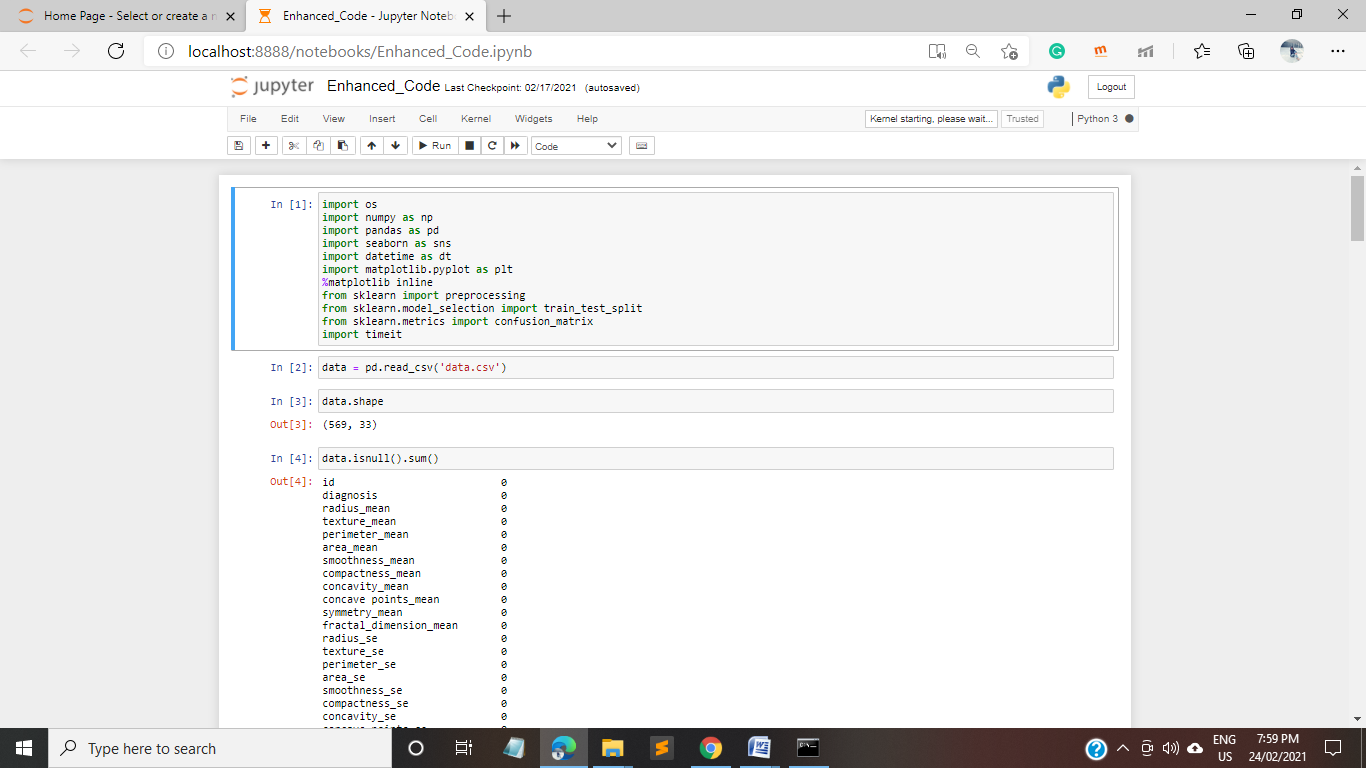


Figure 24. Homescreen of Jupyter Notebook

## 4.3PARAMETERS OF ANALYSIS AND COMPARISON:

* + - Accuracy
    - Number of Iterations
    - Time taken
    - Training Time taken
    - Heatmap

**4.3.1 Performance Parameters**

The metrics referenced underneath are utilized by scientists to assess forecast models and show their exhibition results. Following are the definite boundaries:-

**True Positive Rate**

TP = number of examples predicted positive that are actually positive.

**False Positive Rate**

FP = false positive number of examples predicted positive that are actually negative.

**True Negative Rate**

TN = true negatives: number of examples predicted negative that are actually negative.

**False Negative Rate**

FN = false negatives: number of examples predicted negative that are actually positive.

**Correctly classified instance**

Correctly classified instances means the sum of TP and TN.

**Incorrectly classified instance**

Incorrectly classified instances means the sum of FP and FN

**Accuracy**

The exactness of a classifier on a given test set is the level of test set tuples that are accurately arranged by the classifier

Accuracy = (True Positive + True Negative)/ (Positive + Negative)

**Precision**

Precision is the proportion of the examples which truly have class x divided by total classified as class x.

Precision = True Positive /(True Positive + False Positive)

**Recall**

Recall is the ratio of relevant instances found in the search result to the total of all relevant instances.

Recall = True positive / (True positive + False negative)

**F-measure**

A measure that combines precision and recall is the harmonic mean of precision and recall balanced.

F-measure = 2\*(Precision\*Recall)/ (Precision + Recall)

**Confusion Matrix**

A confusion shows the quantity of right and mistaken expectations made by the model contrasted and the real orders in the test information. The lattice is n-by-n, where n is the quantity of classes as demonstrated in Table.6. From that we determined the exactness of every arrangement calculations. The formula for calculating the Accuracy,

Accuracy = Number of correctly classified samples / Total number of samples.

**Table.6 Confusion Matrix**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | **Actual Class** | | |
|  | **MATRIX** |  |  |  |  |
|  |  |  |  |  |  |
|  |  | **C1** |  | **C2** | |
|  |  |  |  |  |  |
|  | **C1** | True Positive |  | False Positive | |
|  |  |  |  |  |  |
|  | **C2** | False Positive |  | True Negative | |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

A **confusion matrix** is acquired to figure the exactness of order. A disarray lattice shows the number of occasions have been doled out to each class. To foresee the likelihood of patients having breast malignancy, a disarray network made and appeared in *Table.6*, where A denotes patients with breast cancer, and B denotes patients with no breast cancer.

**Table.7 Confusion matrix using Random Forest Classifier**

|  |  |  |
| --- | --- | --- |
| **N=171** | **A (Predicted Yes)** | **B (Predicted No)** |
|  |  |  |
| **A (Actual Yes)** | **101** | **7** |
|  |  |  |
| **B (Actual No)** | **5** | **58** |
|  |  |  |

# CHAPTER-5

**RESULT ANALYSIS AND OUTCOMES**

**5.1 DATA MINING ALGORITHMS**

Various algorithms formulate due to different research works on data mining. These techniques are straightforwardly utilized for developing frameworks or to find crucial inferences and conclusions from the resulted dataset. Various well-known techniques are Support vector machine; K-means, Naïve Bayes, Artificial neural network etc are discussed.

**5.1.1 Naïve Bayes Algorithm**

A least difficult procedure for building classifiers and a probabilistic classifier relies upon

Bayes' hypothesis. Bayes hypothesis is determined as follows:

P (C|X) = P (X|C) \* P(C)/P(X)

Where C is the class with the end goal that P(X) is steady for all classes and X is information tuple. All classifiers guess that the estimation of a specific element is autonomous of the estimation of some other component, given the class variable. Notwithstanding the way that it guess an unreasonable condition that characteristic qualities are restrictively autonomous, it performs shockingly well on voluminous datasets where this condition is expected and holds.

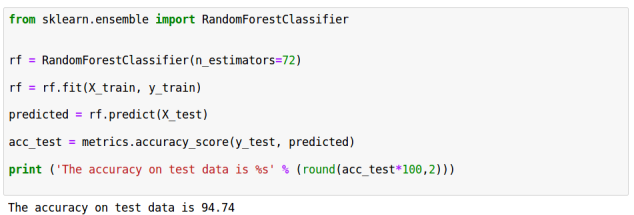


Figure 25. shows Accuracy of iNaïve iBayes

**Advantages:**

1. Easy to carry out.

2. Requires a limited quantity of preparing information to assess the boundaries.

3. Good outcomes got in the majority of the cases.

**5.1.2 Support Vector Machine**

SVM is used for accessing information and pattern for regression analysis and classification. The objective is to find most accurate classification function to recognize individuals from the two classes in the training dataset. It depends on mathematical functions and utilized in model complex, and real world problems [10]. It works well on data sets that have many attributes and ensures that the best such function is found by maximizing the margin between the two classes. It differentiates data by searching the best hyper plane that divides all data points of one class from those of the other class.

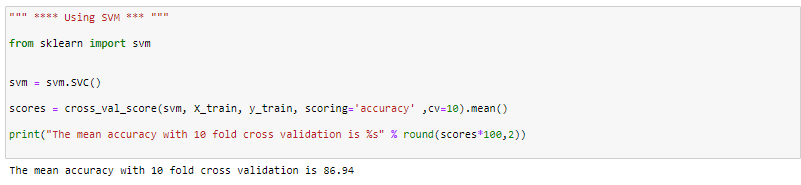


Figure 26. shows Accuracy ofiSupport iVector iMachine i(SVM)

**Advantages:**

1. Memory efficient because it is a subset of training points.
2. Provides high dimensional input space.
3. Simple to manage decision rule complexity.

**5.1.3 Random Forest**

Random forest is a combination of decision trees that can be modeled for prediction and behavior analysis. The choice tree in a backwoods can't be pruned for testing and thusly, supposition choice. The inconsistent woods strategy can oversee gigantic illuminating records because of its capacity to work with different parts hustling to thousands.

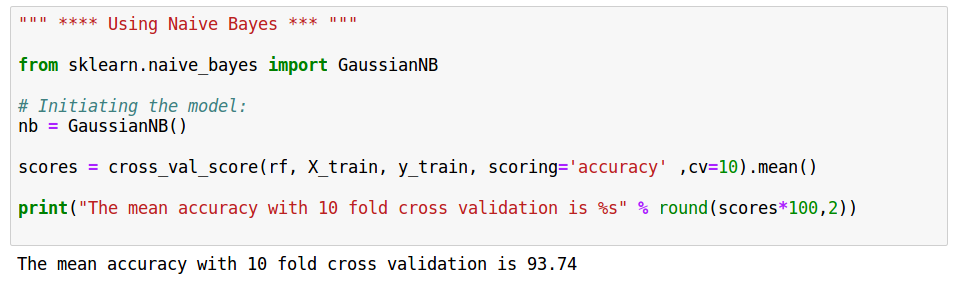
****

Figure 27. shows Accuracy of iRandomiForest

**Advantages:**

1. It is maybe the most exact learning computations available. For a few, instructive records, it conveys a significantly exact classifier.

2. It runs adequately on immense informational indexes.

3. It can manage a large number of data factors without variable retraction.

4. It gives examinations of what variables are huge in the course of action.

Graph.1 displays the graphical representation of results using the Naïve Bayes and SVM classifier technique and Random Forest. The graph clearly represents Random Forest approach is the best with high precision.

**Graph.1iComparingiAlgorithmsiOniTheiBasisiOfiDetailediAccuracy**

**Table.8 Detailed accuracy of Random Forest Classifier**

|  |  |
| --- | --- |
| **Parameters** | **Values** |
|  |  |
| **Correctly Classified Instance** | **159** |
|  |  |
| **Incorrectly Classified Instance** | **12** |
|  |  |
| **Accuracy** | **92.98%** |
|  |  |
| **TP Rate** | **92.9%** |
|  |  |
| **FP Rate** | **7.02%** |
|  |  |
| **Precision** | **95.28%** |
|  |  |
| **Recall** | **93.51%** |
|  |  |
| **F-measure** | **94.34%** |
|  |  |

**Table.9 Performance Comparisons of Random Forest, SVM, Naïve Bayes Algorithms**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm classification** | **Correctly Classified Instance** | **Incorrectly Classified Instance** | **Accuracy** |
| **Naïve Bayes classifier** | **93.74%** | **6.26%** | **93.74%** |
| **Support Vector Machine** | **86.94%** | **13.06%** | **86.94%** |
| **Random Forest**  **Classifier** | **94.15%** | **5.85%** | **94.15%** |

**Table. 10 Comparison Table of Random Forest, SVM, Naïve Bayes Algorithms**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameters** | **Random Forest** | | **Support Vector Machine** | | | **Naïve Bayes** | |
| **Dataset** | **Small Dataset**  **(300 rows)** | **Large Dataset**  **(1140 rows)** | **Small Dataset**  **(300 rows)** | **Large Dataset (1140 rows)** | | **Small Dataset (300 rows)** | **Large Dataset (1140 rows)** |
| **Accuracy(in %)** | **90.42%** | **98.5%** | **84.17%** | | **87.58%** | **92.81%** | **97.99%** |
| **Training Time(in ms)** | **0.81** | **0.82** | **0.81** | | **0.83** | **0.81** | **0.81** |
| **Testing Time(in ms)** | **0.30** | **0.30** | **0.32** | | **0.35** | **0.44** | **0.32** |

**Table.11 Comparison Table of Random Forest with basic parameters and cross validation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameters** | **Random Forest** | | **Random Forest(After CV,10 folds)** | |
| **Dataset** | **Small Dataset (300 rows)** | **Large Dataset (1140 rows)** | **Small Dataset (300 rows)** | **Large Dataset (1140 rows)** |
| **Accuracy(in %)** | **90.42%** | **98.5%** | **91.11%** | **98.83%** |
| **Training Time(in ms)** | **0.82** | **0.80** | **0.81** | **0.81** |
| **Testing Time(in ms)** | **0.30** | **0.30** | **0.30** | **0.31** |

**Table.12 Comparison Table of parameters used in different Classification Algorithm for Breast Cancer Prediction**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S. N.** | **Algorithm** | **Correctly Classified Instances** | **Incorrectly Classified Instances** | **Kappa Statistic** | **Mean Absolute Error** | **Root Mean Squared Error** | **Relative Absolute Error** | **Root Relative Squared Error** |
| **1** | **J48** | **75.8333 %** | **24.1667 %** | **0.6289** | **0.2155** | **0.3632** | **49.7768 %** | **78.0737 %** |
| **2** | **Naïve Bayes** | **67.7083 %** | **32.2917 %** | **0.513** | **0.2251** | **0.397** | **52.0046 %** | **85.3312 %** |
| **3** | **Decision Table** | **71.0417 %** | **28.9583 %** | **0.5585** | **0.2901** | **0.367** | **67.0138 %** | **78.8864 %** |
| **4** | **Random Forest** | **77.0833 %** | **22.9167 %** | **0.6459** | **0.2422** | **0.3324** | **55.9451 %** | **71.4604 %** |
| **5** | **Random Tree** | **66.875 %** | **33.125 %** | **0.4886** | **0.2178** | **0.4565** | **50.3176 %** | **98.1204 %** |
| **6** | **REP Tree** | **64.5833 %** | **35.4167 %** | **0.4477** | **0.2727** | **0.4117** | **63.0007 %** | **88.5017 %** |
| **7** | **LMT** | **68.6916 %** | **31.3084 %** | **0.5702** | **0.099** | **0.2737** | **46.7534 %** | **84.3316 %** |
| **8** | **Decision Stump** | **52.5000 %** | **47.5 %** | **0.2581** | **0.3461** | **0.4164** | **79.9546 %** | **89.4966 %** |

**Table.13 Comparison Table of parameters on the basis of datasets**

**For small dataset (rows-300, columns-32)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameters** | **Support Vector Machine** | **Naïve Bayes** | **Random Forest** |
| **Accuracy** | 84.17% | 92.81% | 92.82 |
| **Training Time** | 0.80sec | 0.81sec | 0.80sec |
| **Testing Time** | 0.32sec | 0.44sec | 0.31sec |

**For large dataset (rows-1140,columns-32)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameters** | **Support Vector Machine** | **Naïve Bayes** | **Random Forest** |
| **Accuracy** | 87.58% | 97.99% | 98.5% |
| **Training Time** | 0.85sec | 0.86sec | 0.86sec |
| **Testing Time** | 0.35sec | 0.32sec | 0.30sec |

**5.2 Detailed results from RapidMiner Tool**

**5.2.1 Naïve Bayes:**

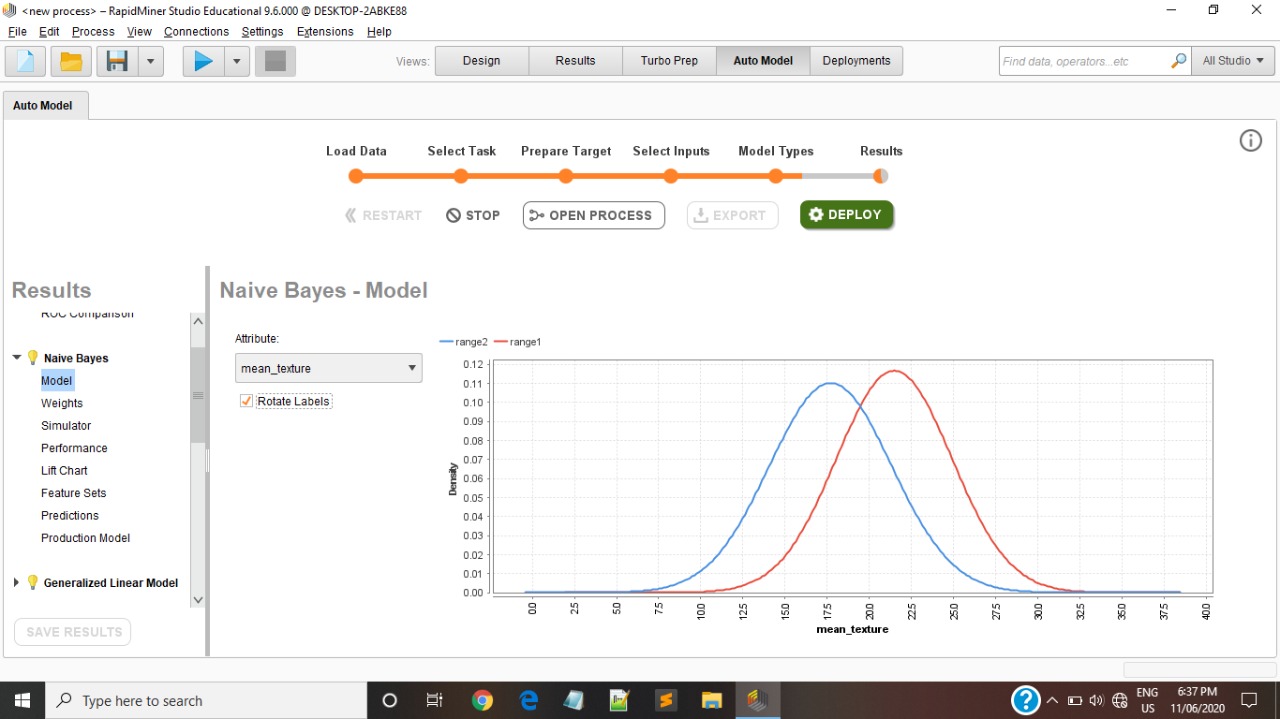
****

Figure 28. Naïve Bayes Model

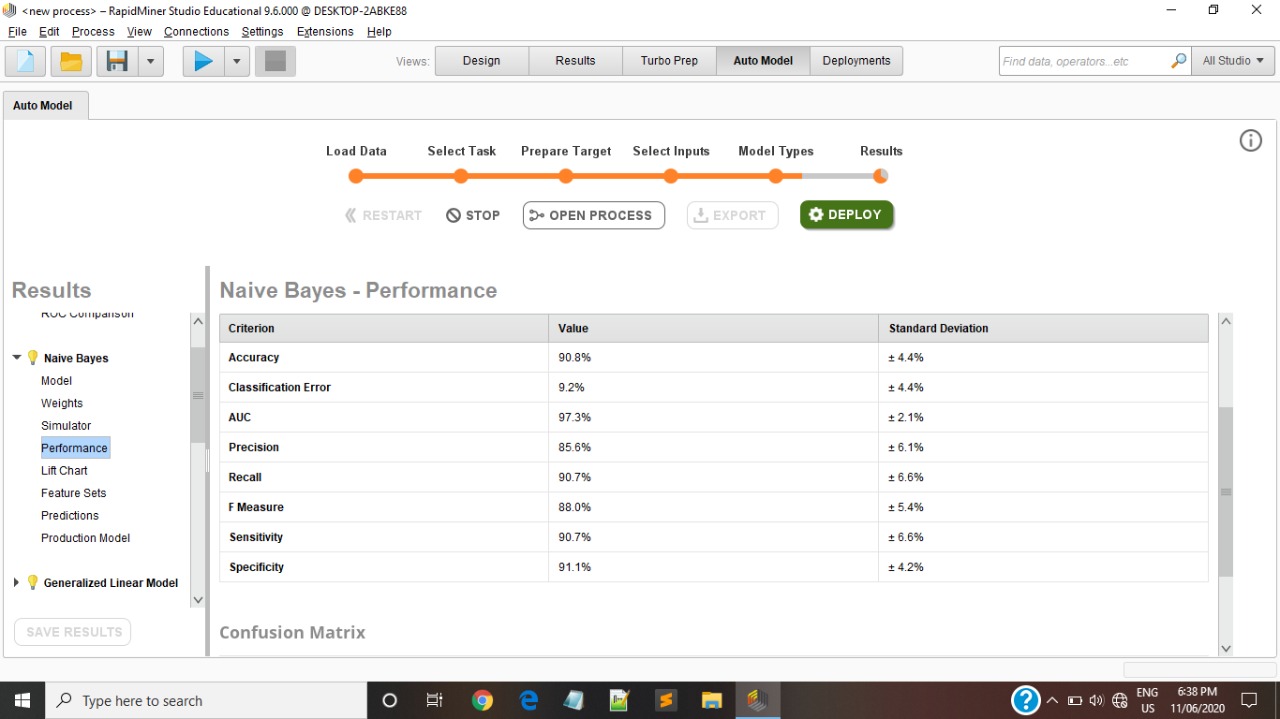
****

Figure 29. Naïve Bayes Performance

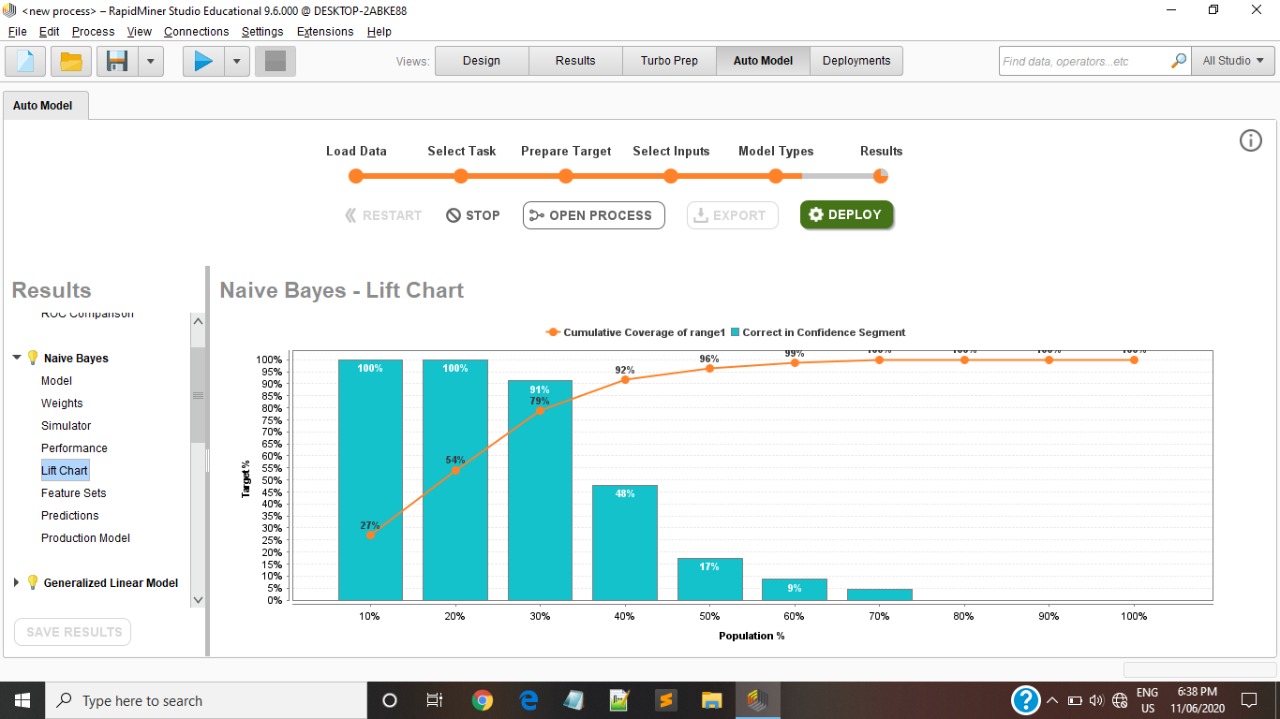
****

Figure 30. Naïve Bayes Lift Chart

**5.2.2 Support Vector Machine:**

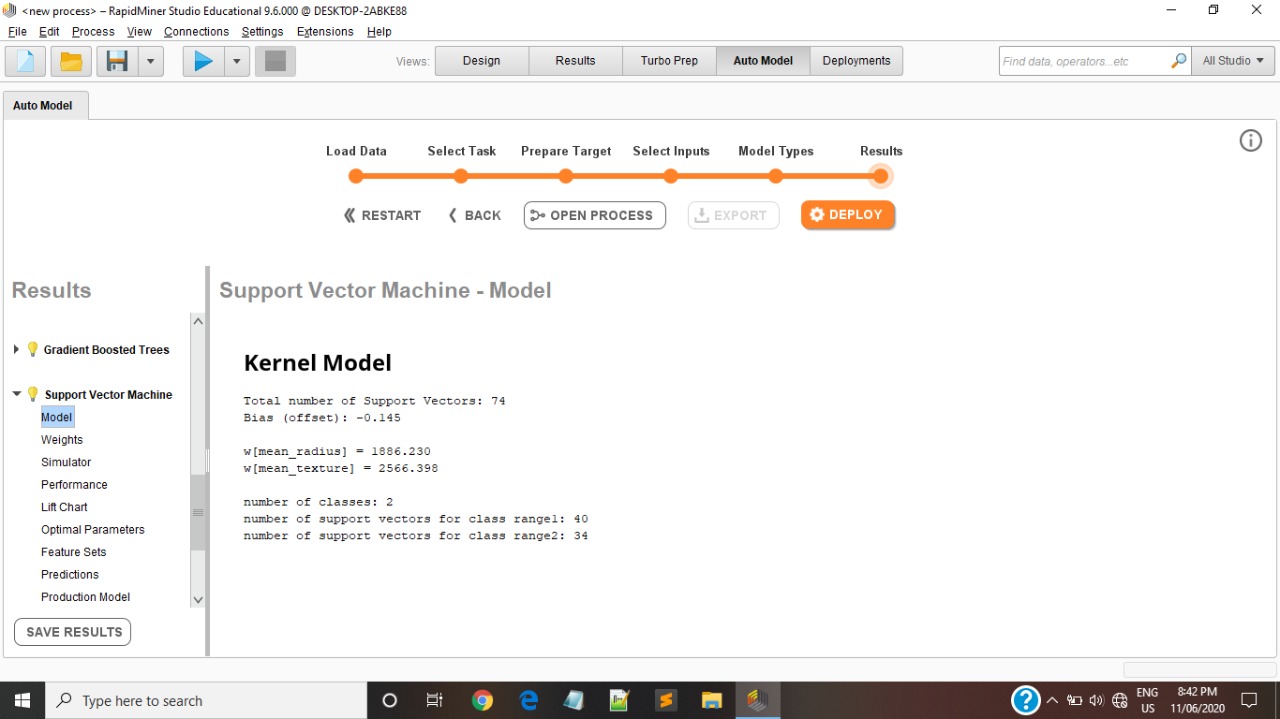
****

Figure 31. SVM Model

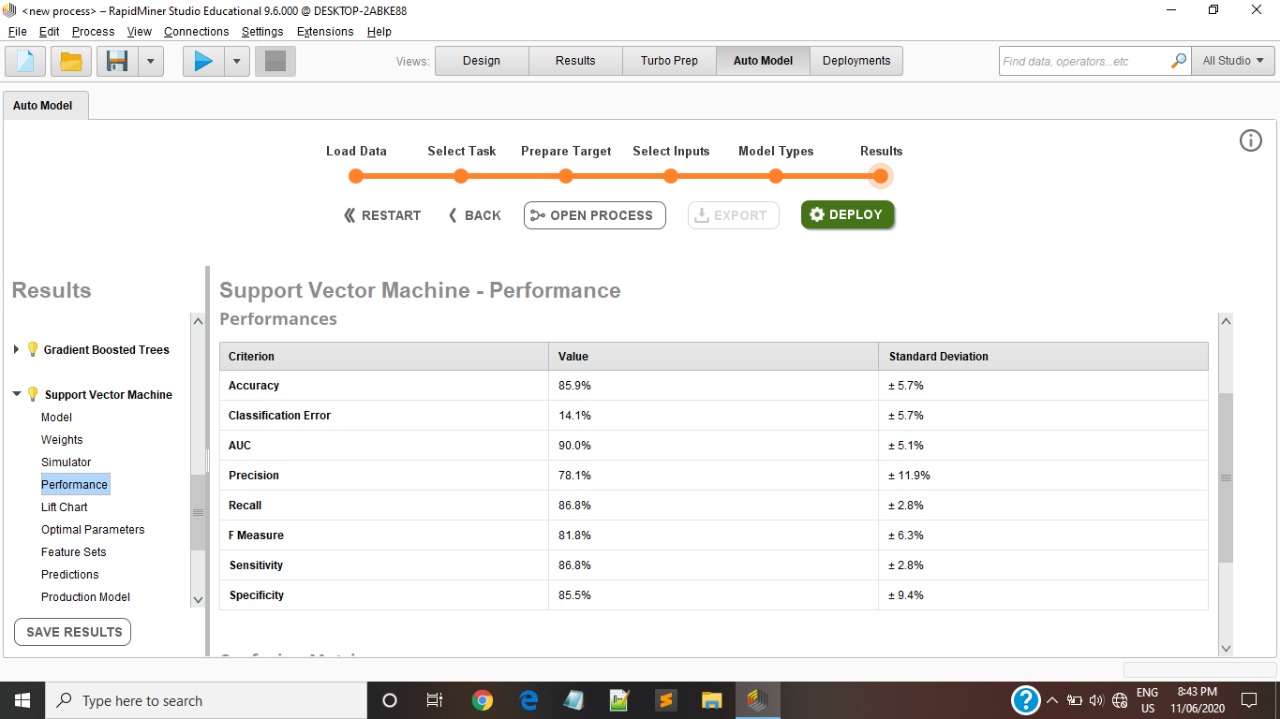
****

Figure 32. SVM Performance

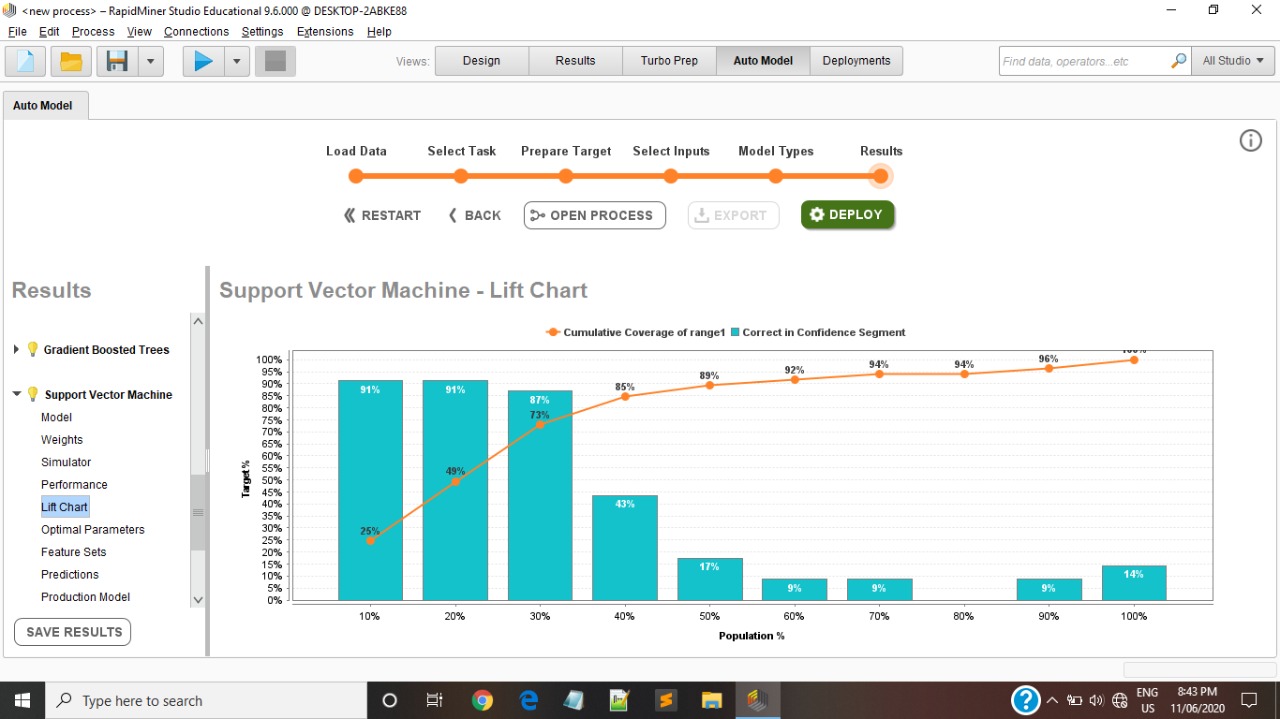
****

Figure 33. SVM Lift Chart

**5.2.3 Random Forest:**

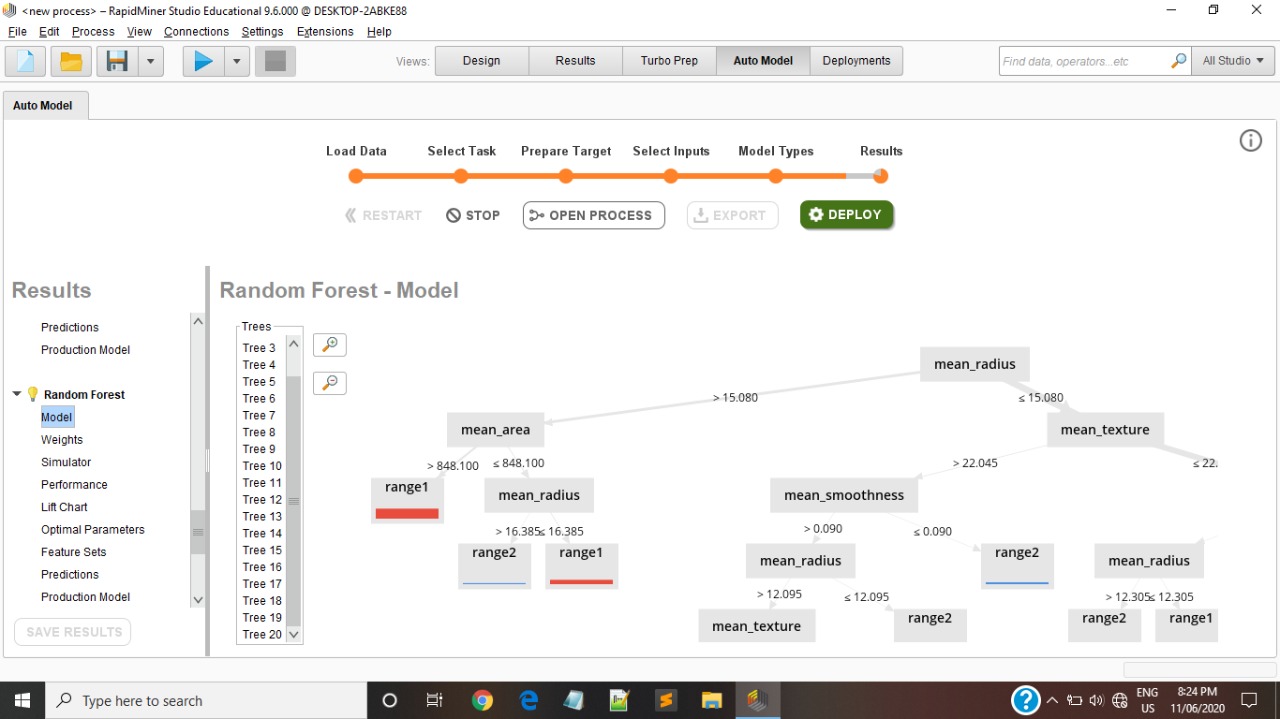
****

Figure 34. Random Forest Model

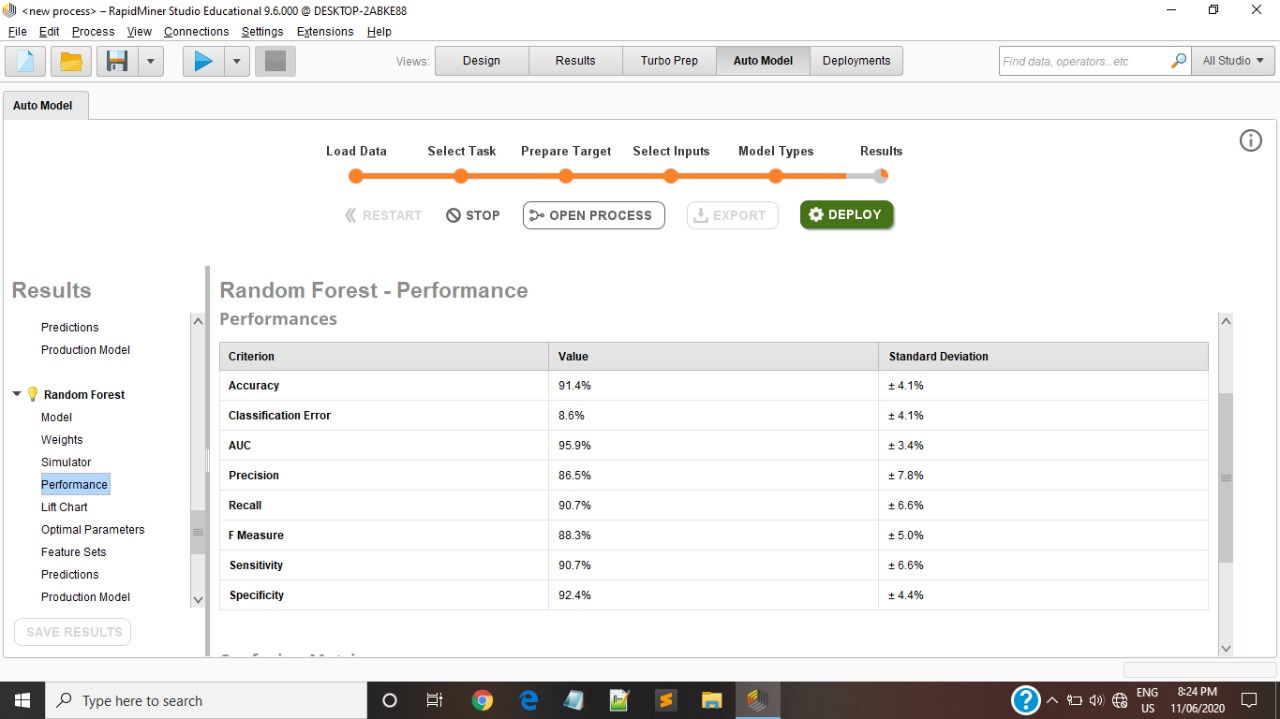
****

Figure 35. Random Forest Performance

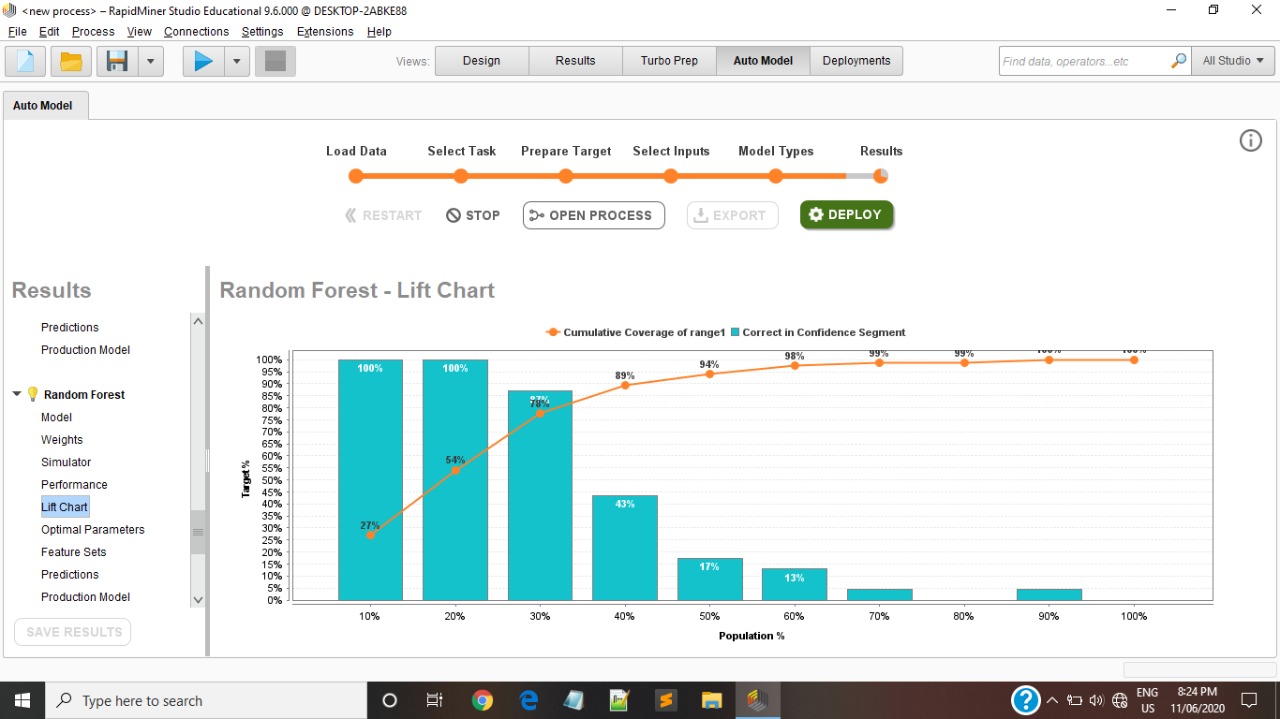
****

Figure 36. Random Forest Lift Chart

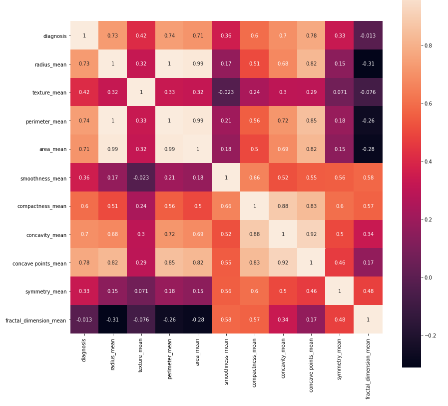
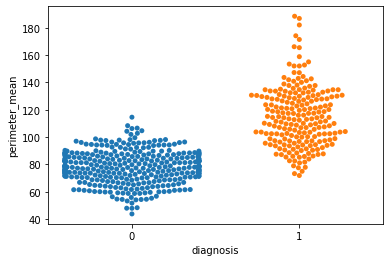
****

Figure 37.A iheat imap shows i(or iheatmap) iidata ivisualization itechnique ithat idefines imagnitude iof ia iphenomenon ias icolor iin itwo idimensions. iThe ivariation iin icolor imay ibe iby ihue ior iintensity, igiving iobvious ivisual icues ito ithe ireader iabout ihow ithe iphenomenon iis iclustered ior ivaries iover ispace.

**Graph.2 iPerimeter\_mean iv/s iDiagnosis**



# CHAPTER-6 CONCLUSION

**6.1 Conclusion**

Cancer iis ipotentially ifatal idisease. iDetecting icancer iis istill ichallenging ifor ithe idoctors iin ithe ifield iof imedicine. iEven inow ithe iactual ireason iand icomplete icure iof icancer iis inot iinvented. iDetection iof icancer iin iearlier istage iis icurable. iIn ithis iwork iwe ihave ideveloped ia isystem icalled idata imining ibased icancer iprediction isystem. iThe imain iaim iof ithis imodel iis ito iprovide ithe iearlier iwarning ito ithe iusers iand iit iis ialso icost iand itime isaving ibenefit ito ithe iuser. iIt ipredicts ithree ispecific icancer irisks.. iSpecifically, iCancer iprediction isystem iestimates ithe irisk iof ithe ibreast, iskin, iand ilung icancers iby iexamining ia inumber iof iuser iprovided igenetic iand inon-genetic ifactors. iThis isystem iis ivalidated iby icomparing iits ipredicted iresults iwith ithe ipatient’s iprior imedical irecord iand ialso ithis iis ianalyzed iusing iRapidMiner itool. iThis iprediction isystem iis iavailable iin ionline, ipeople ican ieasily icheck itheir irisk iand itake iappropriate iaction ibased ion itheir irisk istatus. iThe iperformance iof ithe isystem iis ibetter ithan ithe iexisting isystem. i i

In ithis ipaper iwe idiscussed iabout ithree ialgorithms iSVM, iRandom iForest iand Naïve Bayes. iWe iimplemented ithem ion idifferent icancer idatasets ito icompare itheir accuracies iand ito ifind iout iwhich ione iproposes ithe ibest isolution. iAbove iin ithe given itable ithe iaccuracies iof ithe idata iset icorresponding ito ithe ialgorithm iare given.. Hence, ithis ishows ithat iRandom iForest ialgorithm igives ius ithe imore accurate results ithan iSVM iand iNaïve iBayes ialgorithm iin iboth iPython iand iR language.. We assume that the automated perfect estimate of breast danger ensures preventive measures for the contamination. We need beneficial applications for area and neutralization of breast danger and risk.

**6.2 Future work**

Since majority of previous studies used artificial neural networks (ANNs) for breast cancer detection, our aim was to do a classification of breast cancer cases on malignant or benign by decision tree classifiers implemented in Weka software on of proper treatment methods for a patient diagnosed with heart disease. Results obtained from this study decision tree algorithms are strong candidates for breast cancer detection. Random Forest proved to be the most suitable decision tree algorithm for breast cancer classification because it achieved the highest accuracy in both distributions.

In future it is needed to accumulate the data from various regions across the world and make a more exact and general farsighted model for breast threat end. Future examination will similarly zero in on get-together data from a later time span and find new anticipated that prognostic elements should be associated with a decision tree. The work can be expanded and overhauled for the computerization of Breast threat investigation.

# References

[1] K. C. Gast *et al.*, “Accuracy of self-reported cancer treatment data in young breast cancer survivors,” *J. Patient-Reported Outcomes*, vol. 3, no. 1, 2019, doi: 10.1186/s41687-019-0114-5.

[2] F. Paquin, J. Rivnay, A. Salleo, N. Stingelin, and C. Silva, “Multi-phase semicrystalline microstructures drive exciton dissociation in neat plastic semiconductors,” *J. Mater. Chem. C*, vol. 3, pp. 10715–10722, 2015, doi: 10.1039/b000000x.

[3] C. Shah and A. G. Jivani, “Comparison of data mining classification algorithms for breast cancer prediction,” *2013 4th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2013*, no. July, 2013, doi: 10.1109/ICCCNT.2013.6726477.

[4] J. Talukdar and S. K. Kalita, “Detection of Breast Cancer using Data Mining Tool (WEKA),” *Int. J. Sci. Eng. Res.*, vol. 6, no. 11, pp. 1124–1128, 2015.

[5] K. Hamdi *et al.*, “miRNAs in Sera of Tunisian patients discriminate between inflammatory breast cancer and non-inflammatory breast cancer,” *Springerplus*, vol. 3, no. 1, pp. 1–9, 2014, doi: 10.1186/2193-1801-3-636.

[6] D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: A comparison of three data mining methods,” *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, 2005, doi: 10.1016/j.artmed.2004.07.002.

[7] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[8] P. Ramachandran, N. Girija, and T. Bhuvaneswari, “Early Detection and Prevention of Cancer using Data Mining Techniques,” *Int. J. Comput. Appl.*, vol. 97, no. 13, pp. 48–53, 2014, doi: 10.5120/17069-7492.

[9] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, “Diagnosis of Breast Cancer using Decision Tree Data Mining Technique,” *Int. J. Comput. Appl.*, vol. 98, no. 10, pp. 16–24, 2014, doi: 10.5120/17219-7456.

[10] M. Kaya Keleş, “Breast cancer prediction and detection using data mining classification algorithms: A comparative study,” *Teh. Vjesn.*, vol. 26, no. 1, pp. 149–155, 2019, doi: 10.17559/TV-20180417102943.

[11] B. Zheng, S. W. Yoon, and S. S. Lam, “Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms,” *Expert Syst. Appl.*, vol. 41, no. 4 PART 1, pp. 1476–1482, 2014, doi: 10.1016/j.eswa.2013.08.044.

[12] R.Preetha, S. Vinila Jinny,” A Research on Breast Cancer Prediction using Data Mining Techniques” ISSN 0973-0176 Volume 9, Number 1 (2016).

[13] A. E. M. Reed, J. R. Kutasovic, S. R. Lakhani, and P. T. Simpson, "Invasive lobular carcinoma of the breast: morphology, biomarkers and’omics," Breast cancer research, vol. 17, p. 12, 2015.

[14] K.Arutchelvan, Dr.R.Periyasamy, “CANCER PREDICTION SYSTEM USING DATAMINING TECHNIQUES’, Tamilnadu, India, Nov-2015

[15] Thomas, Tanishk, Nitesh Pradhan, and Vijaypal Singh Dhaka. "Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey." *2020 International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2020.

[16] Neelam Singh Santosh Kumar Singh Bhadauria ,” Early Detection of Cancer Using Data Mining “,(2016), pp. 47-52.

[17] Nyante, S.J., et al., The association between mammographic calcifications and breast cancer prognostic factors in a population‐based registry cohort. Cancer, 2017. 123(2): p. 219-227.

[18] Shu, X., et al., Associations of obesity and circulating insulin and glucose with breast cancer risk: a Mendelian randomization analysis. International journal of epidemiology, 2018.

[19] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011 Mar-Apr;61(2):69-90. doi: 10.3322/caac.20107.

[20] Vertriest C, Berardi G, Tomassini F, Vanden Broucke R, Depypere H, Cocquyt V, et al. Resection of single metachronous liver metastases from breast cancer stage I–II yield excellent overall and disease-free survival. Single center experience and review of the literature. Dig Surg. 2015;32:52–9.

[21] Laronga, C., Chagpar, A. B., & Vora, S. R. (2016). Patient education: breast cancer guide to diagnosis and treatment (beyond the basics). UpToDate. Waltham. https://www.uptodate.com/contents/breast-cancer-guide-todiagnosis-and-treatment-beyond-the-basics. Accessed, 7-12017.

[22] Devi, R.D.H.; Devi, M.I. Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer. Int. J. Adv. Eng. Technol. 2016, 12, 93–98.

[23] Bazazeh D.; Shubair R. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 2016, pp. 1-4

**APPENDIX A**

**PUBLISHED SURVEY PAPER**

**APPENDIX B**

**PRINT OUT OF PLAGIARISM REPORT**