Software Defect Prediction Using Machine Learning

P.L.S.Tejaswini, K.Sree Varsha, P.Yasaswini, Sangeetha Yalamanchili

Abstract— Software defect prediction analysis is an important problem in the software engineering community. Software defect prediction can directly affect the quality and has achieved significant popularity in the last few years. This software prediction analysis helps in delivering the best development and makes the maintenance of software more reliable. This is because predicting the software faults in the earlier phase improves the software quality,efficiency, reliability and the overall cost in SDLC. Developing and improving the software defect prediction model is a challenging task and many techniques are introducing for better performance. Supervised ML algorithms have been used to predict future software faults based on historical data[1]. These classifiers are Naïve Bayes (NB), Support Vector Machine(SVM) and Artificial neural network(ANN). The evaluation process showed that ML algorithms can be used effectively with a high accuracy rate. The comparison is made with other machine learning algorithms to finds the algorithms which gives more accuracy. And the results show that machine learning algorithms gives the best performance. The existence of software defects affects dramatically on software reliability, quality, and maintenance cost. Achieving reliable software also is hard work, even the software applied carefully because most time there is hidden errors. In addition, developing a software defect prediction model which could predict the faulty modules in the early phase is a real challenge in software engineering. Software defect prediction analysis is an essential activity in software development. This is because predicting the bugs prior to software deployment achieves user satisfaction, and helps in increasing the overall performance of the software. Moreover, predicting software defects early improves software adaptation to different environments and increases resource utilization.

Keywords: Naïve Bayes, Support Vector Machine(SVM), Artificial Network Network(ANN), Software Defect Analysis, Reliability, Anaconda Navigator.

I. INTRODUCTION

Software Defect Prediction is an important issue in software development and maintenance processes, which concerns with the overall of software success. Predicting and finding the bugs in the earlier phase in SDLC makes the software more reliable, efficient and better quality when compared with finding bugs in the later stages. However, developing a software defect prediction model is not an easy task and many new tools and methods are introducing in the machine learning for better performance. These classifiers are Naïve Bayes(NB) and Support Vector Machine(SVM) and Artificial Neural Networks(ANN). The development procedure demonstrated that ML calculations can be utilized adequately with a high precision rate.A programming deformity is a blunder, bug, imperfection, issue, breakdown or errors in programming that makes it make a mistaken or unpredicted result. Issues are basic properties of a framework. They show up from structure or assembling, or outside condition. Programming blemishes are customizing mistakes which cause distinctive execution contrasted and expectation. The dominant parts of the flaws are from source code or plan, some of them are from the mistaken code producing from compilers. For programming designers and customers, programming deficiencies are a perilous issue. Programming abandons not only diminish programming quality, increment cost yet in addition postpone the advancement plan. Programming deficiency anticipating is proposed to settle this kind of trouble[4].

Revised Version Manuscript Received on 16 September, 2019.

P.L.S.Tejaswini, IV/IV B.Tech student, IT Department, .V.R Siddhartha Engineering College, Vijayawada. Andhra Pradesh, India
(Email: i.tejaswini.paladugu@gmail.com)
K.Sree Varsha, IV/IV B.Tech student, IT Department, V.R Siddhartha Engineering College, Vijayawada. Andhra Pradesh, India
(Email: Sreevarshakathi126@gmail.com)
P.Yasaswini, IV/IV B.Tech student, IT Department, .V.R Siddhartha Engineering College, Vijayawada. Andhra Pradesh, India
(Email: yasaswini2831@gmail.com)
Dr.Sangeetha Yalamanchili, Associate Professor, IT Department, .V.R Siddhartha Engineering College, Vijayawada. Andhra Pradesh, India
(Email: sangeetha18.yalamanchili@gmail.com)

II. APPLICATIONS

This Software defect prediction helps in deploying the products which are defect free and satisfy the customers.

It helps in reducing the cost of testing the software. It helps to identify the errors in the early stages.

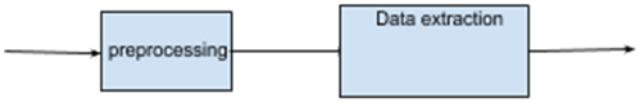
In less time we can find the number of bugs.

Helps to deploy the products earlier than expected time.

III. METHODOLOGY

Pre Processing:

We need to do the preprocessing of the data so that the features of the data can be extracted. After preprocessing load the data into the next step



Machine Learning:

Machine learning is one of the applications in Artificial Intelligence. It helps the system to learn from experience and through programming, and it also helps in improving their ability to learn. The main aim of machine learning is to develop programs for the systems so that they learn by accessing the given data.

Supervised Machine Learning:

Managed learning, with regards to computerized reasoning (AI) and AI, is a kind of framework where both info and wanted yield information are provided[3]. Information and yield information are named for arrangement.

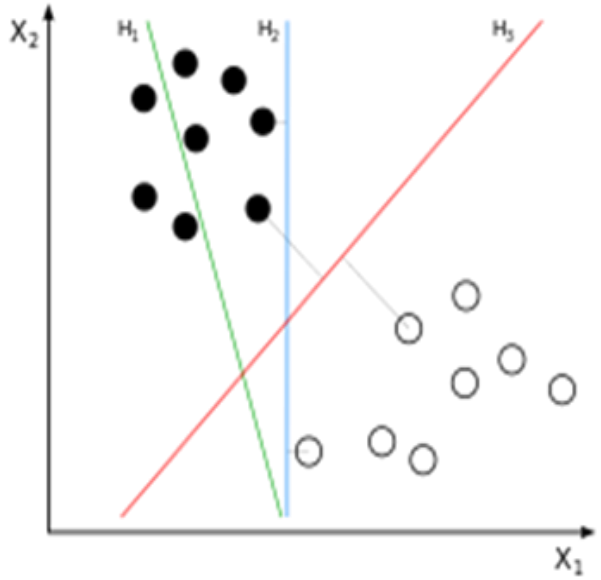
IV. CLASSIFICATION

Classification is nothing but identifying the categories to which the given data set belongs to by training the given data with instances having different categories.

Naïve bayes:

Naïve Bayes is one of the order methods which depends on Bayes Theorem. Gullible Bayes classifier accept that the nearness of any component in a class isn't identified with the nearness of some other element.

space and mapping them to find the class and they are separated by a line having the gap.



Artificial Neural Network:

$$P\left(\frac{h}{d}\right) = \frac{p(d|h)p(h)}{p(d)}$$

Above,

Artificial Neural Networks (ANN) helps in designing the system to simulate the way, how the human brain analyzes and processes the given information. This technology helps in solving the problems that are impossible to solve by human brains and that beyond the imagination.

P(h|d) is the posterior probability of class (c, target) given predictor (x,attributes).

P(h) is the prior probability of class.

P(d|h) is the likelihood which is the probability of predictor given class. P(d) is the prior probability of predictor.

Gaussian Naïve Bayes :

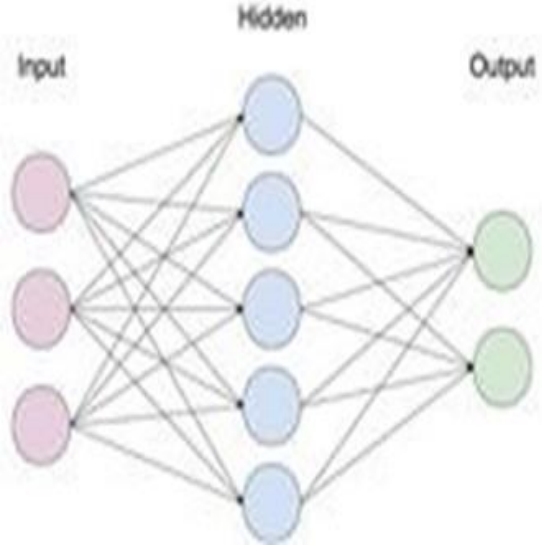
A Gaussian Naive Bayes algorithm is a special type of Naive Bayes algorithm. It is used when the features have continuous values. It's also assumed that all the features are following a Gaussian distribution i.e, normal distribution.

Support Vector Machine:

support vector machines (SVMs) is a directed learning strategy. SVM gains from preparing the dataset and it is utilized for grouping. On the off chance that we think about a lot of preparing models, each case having a place with one of two classes, an SVM calculation assembles a model that aides in anticipating whether the model falls into one class or the other one. Basically, we can think about an SVM model by speaking to the model in

ANN have self-learning capabilities that help them to analyze a large amount of data with best results.

Artificial Neural Networks (ANN) are paving the way for many applications to develop to use in all sectors of the economy. Artificial Intelligence (AI) platforms that are built on ANN are disrupting the traditional way of doing things. From translating web pages into other languages, groceries online to conversing with chatbots to solve problems, online bank transactions, online shopping, etc. AI platforms simplify these transactions and make services available for all efficiently at very low cost.



Datasets:

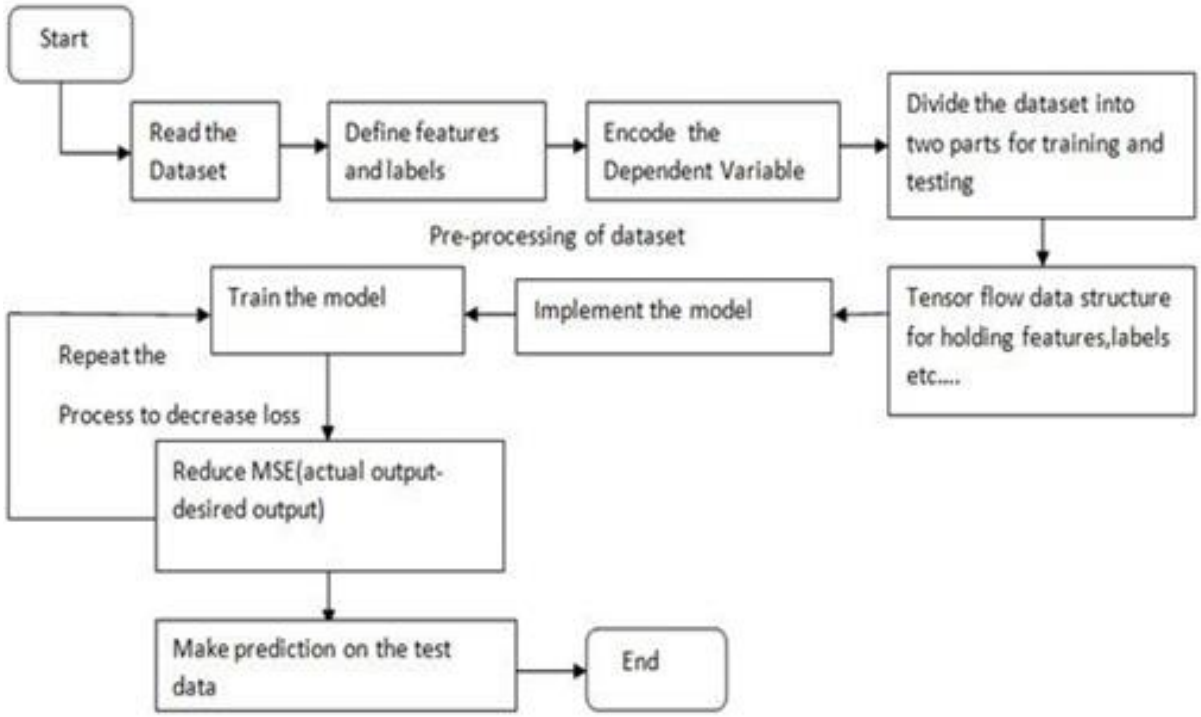
The dataset contains the historic data of the software application contains defects. It has attributes that have an effect on the application. The features in the datasets were defined in the 70s in an attempt to objectively characterize code features that are associated with software quality.

Time Estimator	Line Count	Count of Lines of Con	Count of blank lines	LOCode & Comment	Unique operator	Unique Operands	Total Operators	Total Operands	Branch Count	Problems
1.3	2	2	2	2	1.2	1.2	1.2	1.2	1.4	FALSE
1	1	1	1	1	1	1	1	1	1	TRUE
1187.7	65	10	6	0	18	25	107	64	21	TRUE
635.37	37	2	5	0	16	28	89	52	15	TRUE
132.33	21	0	2	0	11	10	41	17	5	TRUE
356.87	35	2	4	0	11	20	74	41	5	TRUE
644.24	41	2	2	0	12	21	95	54	11	TRUE
1836.77	62	3	2	0	16	22	156	75	23	TRUE
670.5	41	2	1	0	12	20	95	54	11	TRUE
793.24	42	2	1	0	14	22	99	56	13	TRUE
1181.2	50	2	2	0	15	25	124	69	17	TRUE
878.23	54	4	5	0	14	22	93	62	11	TRUE
2512.35	88	14	12	0	19	40	177	112	21	TRUE
219.65	20	0	0	0	11	15	43	31	7	TRUE
2024.14	61	7	7	0	18	24	153	78	37	TRUE
9.06	5	0	0	0	7	5	8	5	1	TRUE
57.83	13	0	1	0	8	14	24	19	3	TRUE
2486.52	71	4	15	0	22	35	162	95	12	TRUE
80.81	18	0	1	0	7	17	44	23	1	TRUE
15.9	18	0	1	0	5	12	5	16	1	TRUE
234	28	0	0	0	3	1	28	26	1	TRUE
382.74	30	0	3	0	12	18	69	39	5	TRUE
174.74	20	0	3	0	9	8	33	24	7	TRUE

V. DESIGN METHODOLOGY

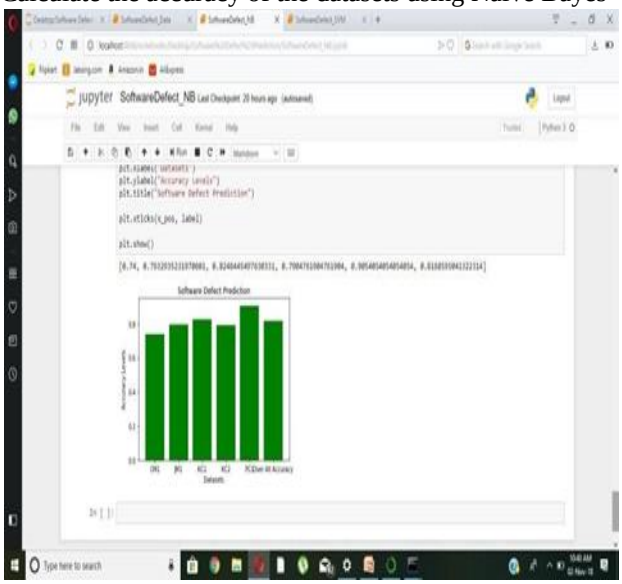
Here in this design methodology, we represent the basic design methodology for the steps involved in the process of software defect prediction analysis. These are the main important steps involved and are explained in the system architecture.

The first step we need to do is to select the datasets with software defects. The Datasets we collected are the benchmark datasets. And select the algorithms that we need to implement. Here we choose the best machine learning algorithms Naïve Bayes, SVM, and ANN. Apply to preprocess on the data so that any missing values can be replaced with other values or it can be used for the feature extraction. New apply the classification so that the data is converted into 0's and 1's by applying binary classification. And at last, find the accuracies by using Naïve Bayes and SVM and ANN algorithm to the computer the accuracies. And we predict the algorithm that gives the most accurate results.



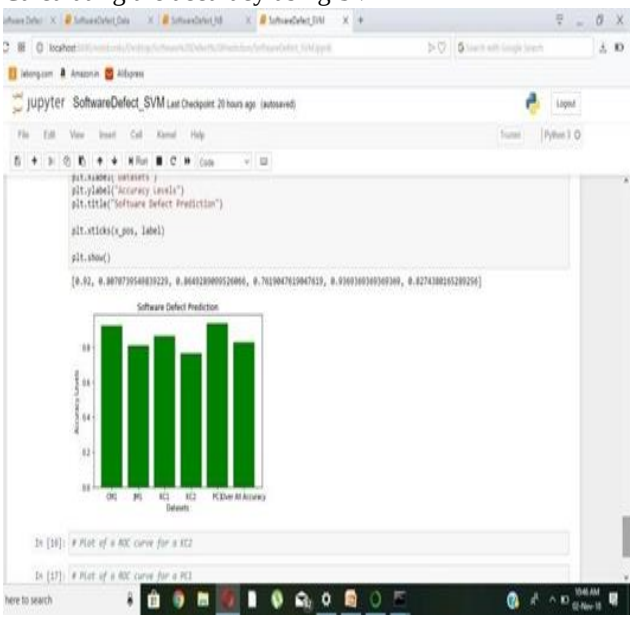
VI. EXPERIMENTAL RESULTS: NAIVE BAYES

Calculate the accuracy of the datasets using Naive Bayes



SVM:

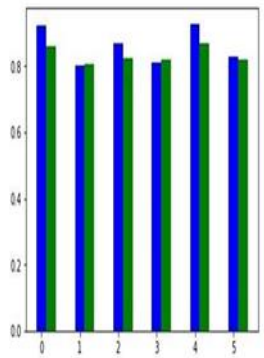
Calculating the accuracy using SVM



NAIVE BAYES AND SVM:

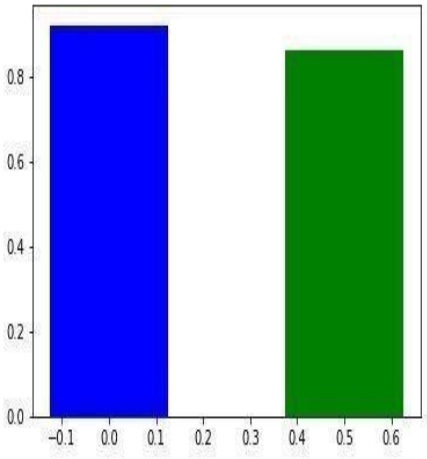
Comparison of accuracies between Naive Bayes and SVM

```
[0.92, 0.881102434542949, 0.8696682464454977, 0.8895238095238095, 0.9279279279279279, 0.827107438016529]
```



Accuracies of all the datasets of Naive Bayes and SVM

```
0.827107438016529
0.8211570247933885
```



ANN:

Two-layer perceptron based neural network for all the data

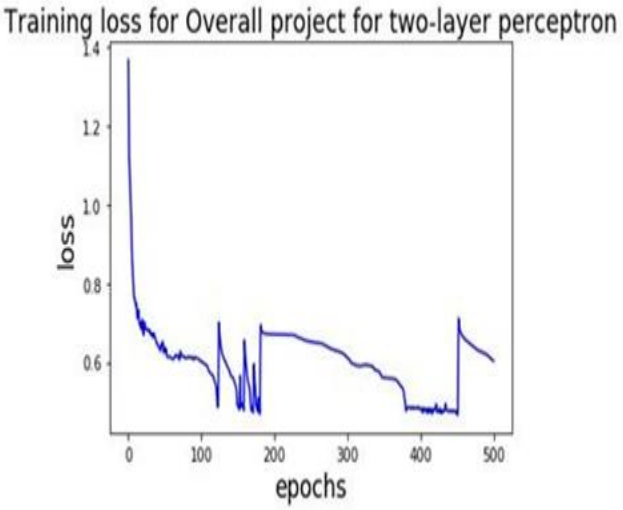
Training_Set...Overall: 12098

Testing_Set....Overall:3025

Training Done!

The accuracy for Overalltest batch is 0.817851

Plotting the training loss for overall project using two-layer Perceptron



Two-layer perceptron based neural network for CM1

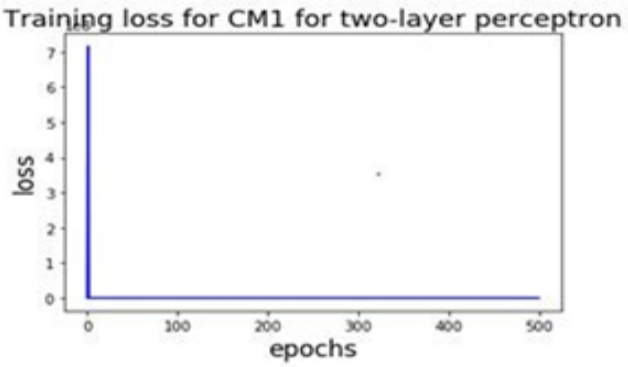
Training_Set: 398

Testing_Set : 100

Training Done!

The accuracy for CM1 test batch is 0.85

Plotting the training loss for CM1 using two-layer perceptron



Changing Hidden layer size and checking Accuracy for CM1

Training Done!

The accuracy for CM1 test batch is: 0.85 Current Hidden Layer Size: 19

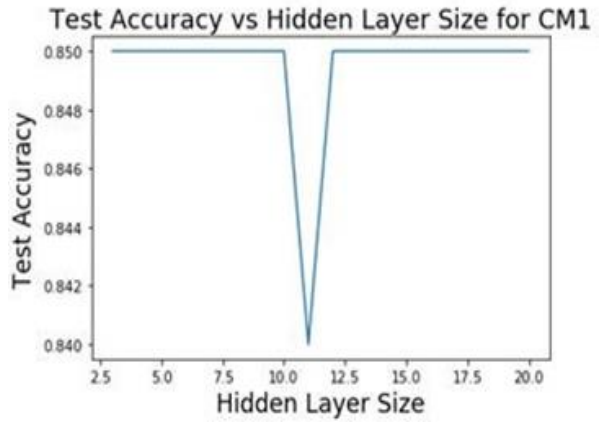
Training Done!

The accuracy for CM1 test batch is: 0.85 Current Hidden Layer Size: 20

Training Done!

The accuracy for CM1 test batch is: 0.85

Plot Accuracy-vs-Hidden Layer Increments for CM1



Changing Hidden layer size and checking Accuracy for Overall data

Current Hidden Layer Size: 3 –

Training Done!

The accuracy for our test batch is: 0.816859 Current Hidden Layer Size: 4

Training Done!

The accuracy for our test batch is: 0.815868 Current Hidden Layer Size: 5 -

…….. Training Done!

The accuracy for our test batch is: 0.817521 Current Hidden Layer Size: 19

Training Done!

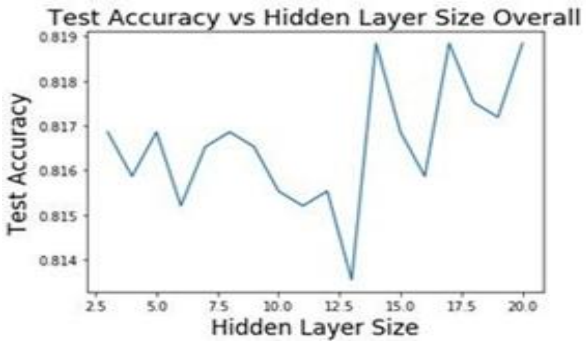
The accuracy for our test batch is: 0.81719

Current Hidden Layer Size: 20

Training Done!

The accuracy for our test batch is: 0.818843

Plot Accuracy-vs-Hidden Layer Increments for Overall Data



VII.CONCLUSION

Software Defect Prediction can directly affect the quality and has achieved significant popularity in the last few years. This software prediction analysis helps in delivering the best quality product without any defects. Therefore this helps in deploying the products that are error free. Here we performed this using machine learning algorithms Naive Bayes, Support vector machine and Artificial Neural Networks. When we observed the accuracies obtained between these algorithms SVM is more accurate than Naive Bayes and Artificial neural network(ANN) is more accurate than both of NB and SVM.

REFERENCES

1. http://www.ijcea.com/comparative-study-analysis-classification- algorithms-machine- learning/ on "COMPARATIVE STUDY AND ANALYSIS OF CLASSIFICATION ALGORITHMS THROUGH MACHINE LEARNING" by Dr. V. T. Meenatchi, Dr. M. Thangaraj , Dr. V.Gayathri , Dr. S.Gnanambal
2. https://ieeexplore.ieee.org/document/7943255/authors#authors "SOFTWARE DEFECT PREDICTION ANALYSIS USING MACHINE LEARNING ALGORITHMS" by Praman deep singh, Anuradha chug.
3. https://www.ijcaonline.org/proceedings/icccmit2014/number2/1 9773-7017 "SOFTWARE DEFECT CLASSIFICATION USING BAYESIAN CLASSIFICATION TECHNIQUES" by Sakthi Kumaresh, Baskaran R, Meenakshy Sivaguru.
4. http://thesai.org/Publications/ViewPaper?Volume=9&Issue=2& Code=IJACSA&Serial No=12"SOFTWAR BUG PREDICTION USING MACHINE LEARNING APPROACH" by Awni Hammouri, Mustafa Hammad, Mohammad Alnabhan, Fatima Al Sarayrah
5. http://image.diku.dk/imagecanon/material/cortes_vapnik95.pd f,"SUPPORT-VECTOR NETWORKS MACHINE LEARNING" by C. Cortes, V. Vapnik.
6. https://www.di.ens.fr/~mallat/papiers/svm tutorial.pdf ,"A TUTORIAL ON SUPPORT VECTOR MACHINES FOR PATTERN RECOGNITION IN DATA MINING AND KNOWLEDGE DISCOVERY" by Christopher J.C. Burges.
7. http://www.iraj.in/journal/journal_file/journal_pdf/6-222-14543_26952141-145.pdf "IMPLEMENTATION OF SVM AND NB ALGORITHMS FOR CLASSIFICATION" by ADITI ANIL GHIVE, R. PATIL
8. https://ieeexplore.ieee.org/document/749395"THE RESEARCH OF THE FAST SVM CLASSIFIER METHOD" by Yujun Yang,Jianping Li, Yimei Yang.