Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

11

The Triplet of Machine Learning Algorithms (Logistic Regression, SVM, Random Forest)

Rohit Shinde

School of Computer Science

MIT World Peace University, Pune, India

rohit132909@gmail.com

Pranav Rasankar

School of Computer Science

MIT World Peace University, Pune, India

rasankar.pranav14@gmail.com

Kuldeep Yadav

School of Computer Science

MIT World Peace University, Pune, India

yadavkuldeep1017@gmail.com

Prof. Dr. Shantanu Kanade

School of Computer Science

MIT World Peace University, Pune, India

shantanukanade@gmail.com

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

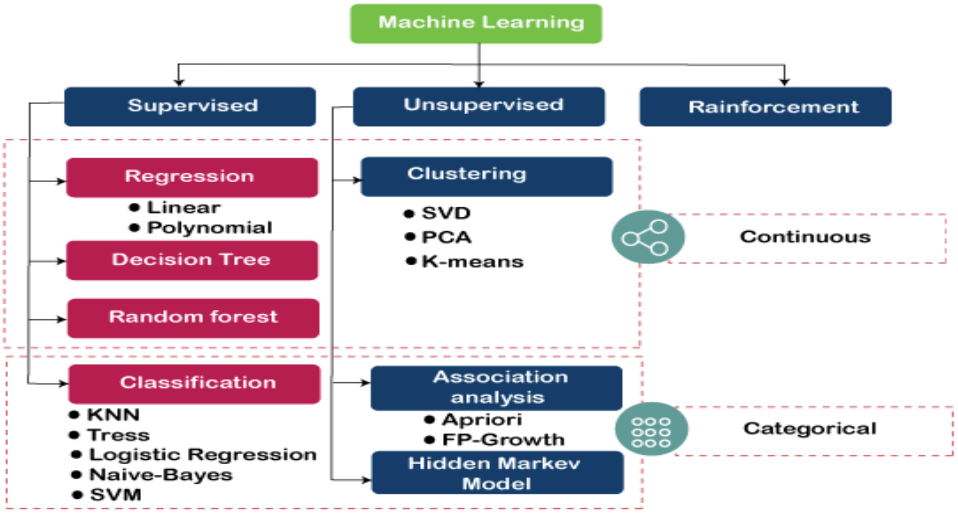
Abstract –

This article discusses machine learning algorithms, including the Support Vector Machine (SVM), Random Forest, and Logistic Regression. In our digital world, there are many different sorts of data, including Internet of Things (IoT) data, cyber security data, mobile data, corporate data, social media data, health data, and many others. It's crucial to master this data and acquire the necessary abilities and knowledge of technology, especially machine learning (ML). In order to grasp these algorithms and make the most of them, we also discuss their uses, comparisons, and applications.

Keywords - Machine learning, Logistic Regression, SVM, Random Forest

I. Introduction

American computer games and artificial intelligence researcher Arthur Samuel first used the phrase "machine learning" in 1959, stating that it "allows computers to learn without being uniquely adapted." ML investigates the evaluation and creation of algorithms that may provide information-based results and create expectations about the information. In light of more information, ML can change activities and responses to make it more efficient, versatile and adaptive. ML is the study of PC algorithms that subsequently operate based on experience. AI is a subset of machine learning. The basic goal of machine learning (ML) is to create computers that can take input data and, using factual inquiry, predict a conclusion while updating the findings as new data is learned. One of the most exciting branches of computer science, machine learning (ML), is the most recent buzzword to surface.



Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

II. Types of Learning -

1. Supervised Learning -

This learning algorithm takes a known arrangement of information (the training set) and a known response to the information (the output) and builds a model to generate intelligent predictions of responses to new information. In supervised learning, a model is trained with a specified data set and the model discovers different types of information. After completing the training cycle, the model is tested on the test data (a subset of the training set) and then the result is predicted. This means that tuned ML algorithms continue to work after deployment, finding new instances and connections as they learn new information. Three of the most popular directed AI computations are reviewed -logistic regression, SVM (Support Vector Machine), Random Forest

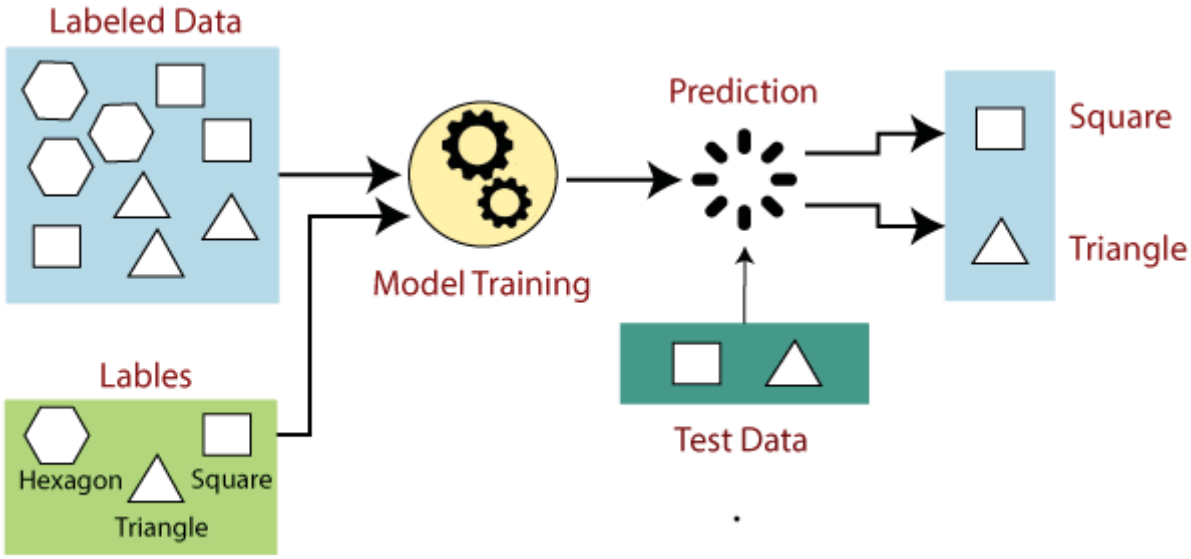


Fig 2

2. Unsupervised Learning -

The advantage is the ability to work with unmarked information.

In unaided learning, only information resources are available, and the model should search for interesting examples with information in mind. Unaided computations learn few elements from information. When new information is presented, it uses recently learned salient points to perceive a class of information.

These calculations identify hidden groups or samples of data without the assistance of a human. It is the best solution for exploratory information, strategic practises, client segmentation, and picture identification because of its capacity to identify similarities and contrasts in data.

Another name for unassisted learning is information discovery (knowledge discovery). Normal unaided learning strategies involve clustering and dimensionality reduction.

3. Reinforcement Learning -

This learning takes its motivation from how people acquire information in their lives. It is associated with taking a reasonable step to maximize the valuation in the particular circumstances. Different programmes and machines use this to notice the best behaviour or the appropriate course of action in diverse situations. Receiving support differs in some ways from administered learning that in regulated learning, the preparatory information carries with it a response key, so that the model is prepared with the correct response itself, even though there is no response in the realization of the support, except that the support specialist chooses how to play the enterprise. Without a single trace of the preparation data file, he will undoubtedly gain from his insight. Support Learning is an input-based machine learning procedure in which a specialist learns how to behave in a climate by acting out activities and seeing their consequences. For each great activity, the specialist receives a positive contribution, and for each terrible activity, he receives negative criticism or punishment. In reinforcement learning, the specialist proceeds consistently using criticism with almost no labeled information, unlike in supervised learning. RL deals with a particular kind of problem where independent control is incremental and the goal is a long journey, such as gaming, mechanical technology, etc.

Logistic Regression

This model is used for binary classification, or forecasts of the sort either, yes or no, A or B, etc. This algorithm can be utilized for multiclass order, we will zero in on its most fundamental application here. It's one of the most utilized ML techniques for twofold orders, changing the contribution over to 0 or 1

e.g: -

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

0: negative class

1: positive class

The classification issues are settled utilizing calculated relapse.

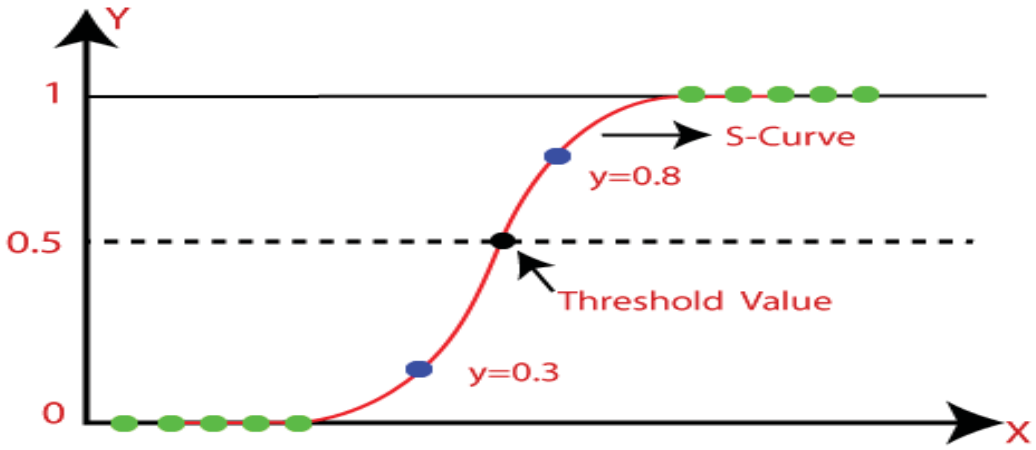
Rather than fitting a relapse line, we fit a "S" moulded strategic capacity in LR, which predicts two most extreme qualities (0 or 1).

The LR's capacity to curve likelihood of things like whether or not the cells are harmful, whether or not a mouse is corpulent in view of its weight, etc.

Since it can create probabilities and characterize new information utilizing both consistent and discrete datasets, LR is a key ML approach.

LR relapse might be utilized to order perceptions in view of many types of information and can rapidly recognize the most helpful elements for arrangement.

The logistic function is depicted in the graphic below:



Regression equation that has been calculated: The logistic regression equation can be obtained using the linear regression equation. The numerical steps to acquire the requirements for Logistic Regression are presented next:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

$$\frac{y}{1-y}$$; 0 for y= 0, and infinity for y=1

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The final Logistic Regression equation is as follows.

Steps in Logistic Regression: In order to implement logistic regression in Python, we'll use the same techniques as in preceding chapters on regression. Following are the steps:

Data Step before processing

Rational Regression Customising the Training Set

estimating a test's results

Check the result's accuracy (Creation of Confusion matrix)

Visualizing the results of the test set.

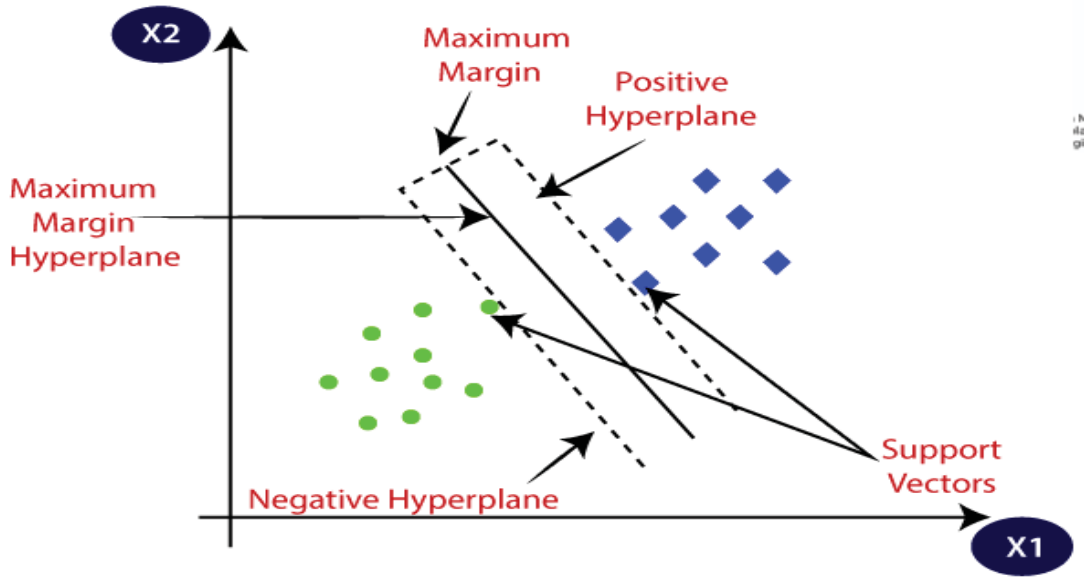
III. SVM (Support Vector Machine)

SVM is a complex administered method that performs well on both little and huge datasets. SVMs, or Support Vector Machines, can be utilized for both relapse and characterization assignments, but they perform better in order circumstances. They were very well known when they were first evolved during the 1990s, and they keep on being the go-to answer for a high-performing calculation with a little tweaking.

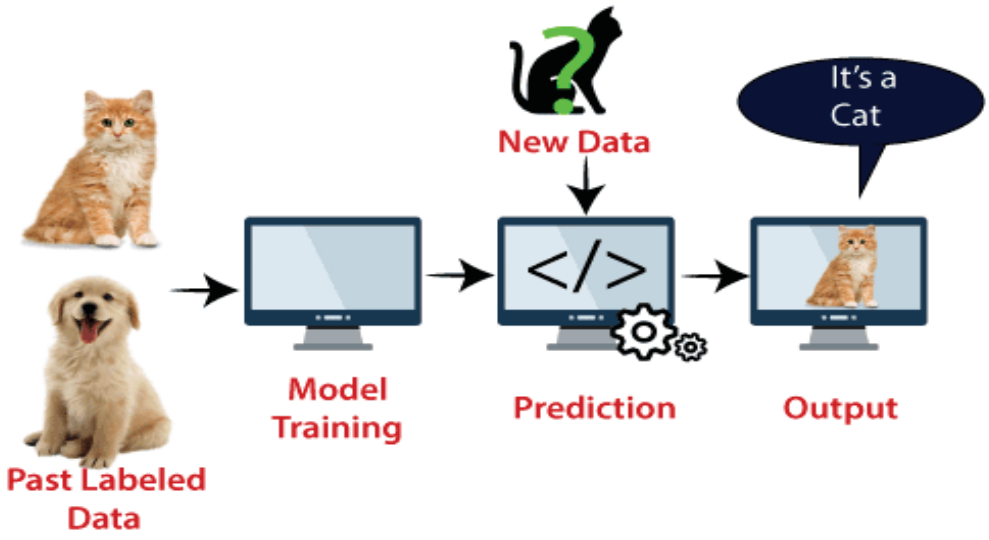
The goal of the SVM computation is to identify the best line or choice limit for categorising n-layered space so that more information foci can be quickly added and afterwards placed in the appropriate category. The best decision limit is known as a hyperplane.

SVM is used to select the outlandish focuses and vectors that contribute to the hyperplane. The calculation is referred as as a "Support Vector Machine" and involves support vectors, which are the outlandish cases. Take a look at the graphic below, which demonstrates how to organise two different categories using a choice limit or hyperplane:

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar



Using the model that we employed in the KNN classifier may help you understand SVM better. We may utilise the SVM method to create a model that can accurately determine whether a strange feline or canine is present if it has some canine-like traits as well. In order to teach our model about the many traits of cats and dogs, we will first supply it with a plethora of photos of them. this peculiar animal. In light of the fact that the aid vector frames a choice limit between these two pieces of information (feline and canine) and selects outrageous situations (support vectors), the outrageous instance of feline and canine will therefore be displayed. It will classify it as a feline based on the basis of the help vectors. Think about the graph below:



How truly does Support Vector Machine function?

SVM is only described in terms of the help vectors; we don't need to frequently consider various perceptions because the edge is determined using the points closest to the hyperplane (support vectors), but the classifier in strategic relapse is described over all locations. As a result, SVM gains from a few built-in speedups.

IV. Random Forest

Irregular woods are a regulated learning strategy that can be utilized to arrange and foresee information. Nonetheless, it is for the most part utilized to address classification issues. A woodland, obviously, is comprised of trees, and more trees approaches more solid backwoods. Also, the Random Forest technique develops choice trees from information tests, removes expectations from each, and afterward decides on the most ideal choice. It's a troupe strategy that is predominant than a solitary choice tree since its midpoints the outcomes to lessen over-fitting.

Irregular woods are a bunch of tree indicators where the upsides of an arbitrary vector gathered autonomously and with similar circulation for all trees in the backwoods are utilized to figure the conduct of each tree. As the quantity of trees in backwoods develops bigger, the speculation blunder meets a.s. as far as possible. The strength of individual trees in the backwoods and their affiliation decide the speculation blunder of a woods of tree classifiers. When a random selection of characteristics is used to divide each hub, the error rates are equivalent to Adaboost, but they are more resilient to disruption (Y. Freund and R. Schapire, Machine Learning: Proceedings of the Thirteenth International Conference, 148-156). Inner evaluations are used to show the response to increasing the number of criteria utilised in the parting by observing mistake, strength, and relationship. Internal gauges are also used to evaluate the significance of various variables. These ideas can also be applied to stop relapses.

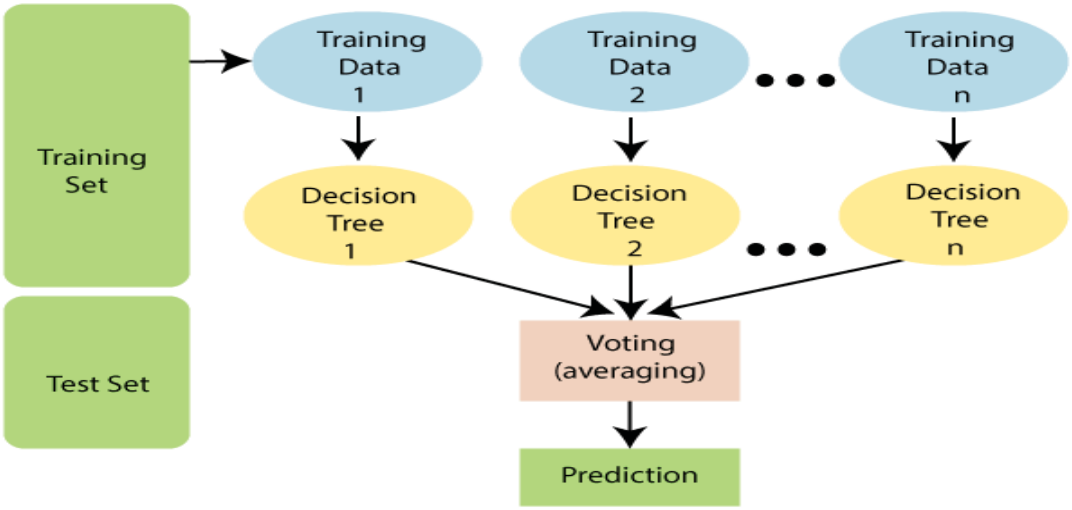
Definition 1.1. An irregular wood is a classifier that consists of a variety of tree-organized classifiers, each of which makes a unit decision for the most popular class at input x. The k variables are free, independently circulated arbitrary vectors in an irregular wood.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

According to its name, irregular forest is a classifier that uses distinct decision trees on different subsets of a given dataset and uses the normal to increase the predicted accuracy of that dataset. Instead of relying solely on one decision tree, the irregular woods collect the hypotheses from each tree and forecasts the final outcome based on the majority of votes from forecasts.

The woods are more precise and the overfitting problem is avoided the more trees there are in it.

The image below shows the Random Forest method:



How truly does Random Forest algorithm work?

Two phases make up the random forest framework: first, combine N selection trees to collect independent trees, and then wait for each tree to be produced in the main step. The following processes and diagrams can serve as an illustration of the workflow.

First, pick out a single K element from the preparation set.

Create the selection tree linked to the selected data elements (Subset) in step two.

Step 3: For the tree you need to add, choose the letter N.

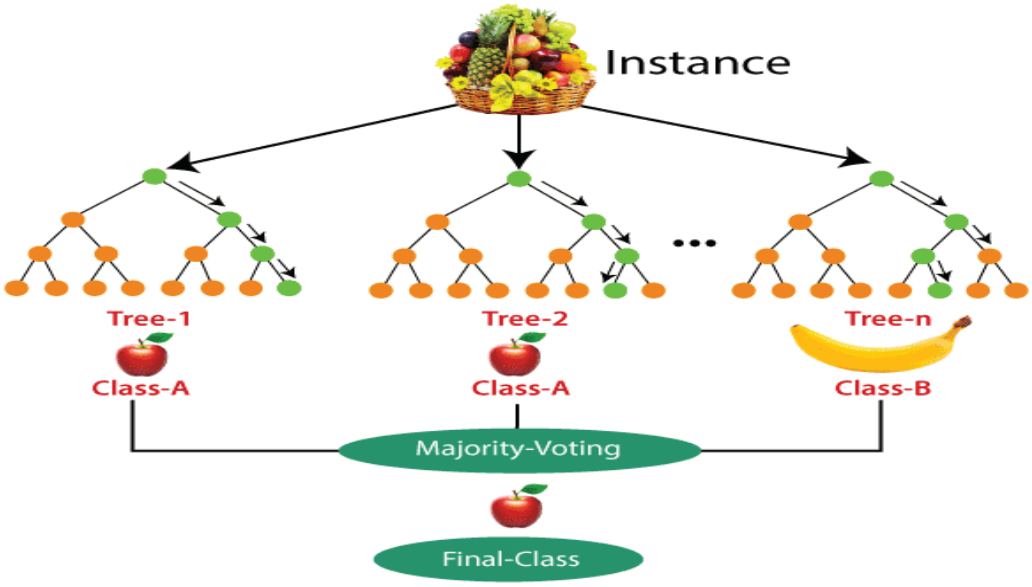
Steps 1 and 2 should be repeated.

Step 5: Execute the expectation of each option tree for new items, then transmit the updated information to the categorization by deciding on the primary component.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

The following model can help you better understand the calculation:

For instance: Imagine you have a database with pictures of different natural goods. In a same vein, a random backwoods classifier is fed the data set. Each option tree is given a portion of the data set, which is separated into pieces. Each selection tree generates a prediction result during the training phase, and if another data point appears, the random forest classifier predicts the final outcome by taking into account the majority of results. Look at the picture below:



V. Table of Comparisons: -

Sr no.	Factors	Logistic Regression	Support Vector Machine	Random Forest
1]	Definition	a statistical analysis method that makes a binary prediction, such as "yes" or "no," based on the data set's initial observations.	a method for locating a hyperplane in an N-dimensional space that can categorise data points in a specific way.	On various samples, it constructs decision trees and, in the case of regression, offers categorization and average votes.
2]	Classification	LR is used only for classification problems.	Problems involving classification and regression are solved with random forests.	Classification and regression issues are solved with SVM.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

3]	Method	It is linear method.	It is linear but kernel makes it non-linear method.	It is inherently non-linear method.
4]	Result	It gives us probability estimates which makes it easier to interpret the result.	It also gives us probability estimates but it is harder than logistic regression to interpret the results.	It lacks in interpretation because each individual decision tree is interpretable but when we take an average then we don't really know, why we get that specific prediction.
5]	Speed	It is fast, but slower than SVM.	It is the fastest algorithm among the three of them.	It is the slowest algorithm among the three of them.
6]	Data	It has some difficulty with high dimensional information	It also has some difficulty with high dimensional information.	It can manage high dimensional information.
7]	Outliers	It cannot handle the outliers.	It handles the outliers better.	It handles outliers as well as noisy data.
8]	Multi-Class handling	It can handle multi-class methods easily.	It is best suited for binary classification because we don't have a inherently multi-class method.	It best for handling the multi-class methods.
9]	Time Complexity	O(n*d) Time Complexity	O(n3) Time Complexity	O (depth of tree*k) time complexity K = number of decision trees.
10]	Space Complexity	O(d) Space Complexity = number of dimensions	O(n2) Space Complexity	O (depth of tree*k) Space Complexity
11)	Advantages	A probabilistic approach provides information about	Performer is not biased by the publisher and is not	Robust and accurate, good performance in many nonlinear

		the statistical significance of the trait.	sensitive to bias.	problems.
12)	Disadvantages	The assumptions of logistic regression.	It is not suitable for non-linear problems; it is not the best choice for a large number of functions.	No significant transplanting can easily happen, the number of trees must be chosen.

VI. Applications: -

Logistic Regression: -

AI, most clinical callings, and sociologies are generally instances of where calculated relapse is applied. Boyd et al., for instance, utilized calculated relapse to set up the Trauma and Injury Severity Score (TRISS), which is habitually used to anticipate passing in harmed patients. Numerous other clinical scales that are utilized to decide a patient's seriousness were made utilizing strategic relapse. Calculated relapse can be used to predict the possibility of nurturing a certain infection (such as diabetes or cardiac illness) based on the patient's observed highlights (age, sex, weight history, results of several blood tests, etc.). According on their age, income, sex, race, place of residence, prior political race votes, etc., another model would predict whether a Nepalese elector would vote for the Nepali Congress, the Communist Party of Nepal, or Any Other Party. The method can also be applied to design, particularly to determine the likelihood that a cycle, structure, or product would fail. It is also utilised in marketing applications like predicting a customer's likelihood to buy an item or cancel a membership, etc. It can be used to predict whether or not someone will enter the workforce in financial aspects, and it is frequently used in business to predict whether or not a property owner will default on a mortgage. In normal language handling, restrictive irregular fields are used as an addition to strategic relapse to subsequent information.

SVM (Support Vector Machine): -

1 Face recognition - SVM classifies the face and non-face part of the image and creates a square border around the face.

2 Text and Hypertext forms - SVM supports both inductive and transductive modelling in text and hypertext forms. They categorise records into several classes using data processing. The argument stated, the ensuing dispute, and the respect at-risk are the main points.

3 picture Classification - Using SVMs improves picture classification accuracy. Compared to traditional research that is focused on methods, it offers greater precision.

4 Protein characterisation and disease sequencing are both included in bioinformatics. To differentiate between patients based on traits and other organic issues, the order of the descriptors is used.

5 To calculate protein homology-distance, use the SVM method to determine protein coverage.

6 Handwriting Recognition - To identify frequently used handwritten characters, we employ SVM.

Use SVM-based Generalised Pre-Control (GPC) to manage turbulent elements with valuable boundaries.

Random Forest: -

These algorithms are used in various industries which help in designing better business strategies.

Finance: The use of algorithms makes it possible to do activities like data management more quickly. Fraud and underpricing problems can be discovered by evaluating consumers with high credit risk.

Medicine: In the field of computational biology, random forest algorithms are frequently used to address a variety of issues, including the classification of gene expression, the identification of biomarkers, and the interpretation of sequences. The doctor can then assess how the medicine will affect a particular drug in this way.

E-commerce: Used to offer machines for sale.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

Conclusions

1) Random Forest

Random Forest are viable in forecast, it produces in great outcome in arrangement, more precise however it is tedious.

This paper's primary goal was to give an audit of flow business linked to the Random Forest classifier and identify potential future research directions in that area.

Random Forests are a viable device in expectation. They give serious results with respect to helping and versatile packing, yet don't dynamically change the preparation set. Arbitrary information sources and irregular elements produce great outcomes in order less so in relapse.

The irregular forest classifier is a group strategy and therefore more accurate, yet it is tedious in contrast to other individual ordering procedures. Basically, we tried to map the result achieved to improve the accuracy and improve the performance of Random Forest.

Introduced as a Comparison Chart, this investigation will fill in as a search rule for future investigation associated with the Random Forest Classifier.

2) SVM: -

As an extremely proficient order model in AI, support vector machine enjoys benefits like great speculation, hardly any boundaries, and the capacity to produce worldwide ideal arrangements. It is an excellent decision for individuals to handle information, dissect information, and foresee information.

With regards to the present large information, support vector machines, as a conventional order strategy, are as yet appropriate because of the predominance of their design and calculations. SVM are prepared by tackling a compelled quadratic streamlining issue. SVM, executes planning of contributions onto a high layered space utilizing a bunch of nonlinear premise capacities.

In short, the advancement of SVM is a totally not quite the same as

typical calculations utilized for learning and SVM gives another understanding into this learning. Support Vector Machines goes about as one of the most amazing ways to deal with information demonstrating.

3) Logistic regression

That the vital portrayal in strategic relapse are the coefficients, very much like direct relapse. That making forecasts utilizing calculated relapse is simple that you can do it in dominate. That the information groundwork for strategic relapse is similar as direct relapse.

References

[1] Random Forests LEO BREIMAN Statistics Department, University of California, Berkeley, CA 94720

[2] Random Forest Classifiers: A Survey and Future Research Directions

[3] Improvement of Support Vector Machine Algorithm in Big Data Background Babacar Gaye, Dezheng Zhang, and Aziguli Wulamu School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China

[4] WikipediaOnline. Http://en.wikipedia.org/wiki

[5] Tutorial on Support Vector Machine (SVM) Vikramaditya Jakkula, School of EECS, Washington State University, Pullman 99164.

[6] An Introduction to Logistic Regression Analysis and Reporting CHAO-YING JOANNE PENG KUK LIDA LEE GARY M. INGERSOLL Indiana University-Bloomington

[7] WORKSHOP ON SUPPORT VECTOR MACHINES: THEORY AND APPLICATIONS Theodoros Evgeniou and Massimiliano Pontil Center for Biological and Computational Learning, and Artificial Intelligence Laboratory, MIT, E25-201, Cambridge, MA 02139, USA

[8] Abdulsalam H, Skillicorn B, Martin P, Streaming Random Forests, Proceedings of 11th International Database and Engineering Applications Symposium, Banff, Alta pp 225-232, (2007)

[9] Towards a better understanding of random forests through the study of strength and correlation Simon Bernard, Laurent Heutte, Sébastien Adam

[10] Bernard S, Heutte L, Adam S, Forest-RK: A New Random Forest Induction Method, Proceedings of 4th International Conference on Intelligent Computing: Advanced

Intelligent Computing Theories and Applications – with Aspects of Artificial Intelligence, Springer-Verlag, (2008)

[11] Grahn H, Lavesson N, Lapajne M, Slat D, A CUDA implementation of Random Forest – Early Results, Master Thesis Software Engineering, School of Computing, Blekinge Institute of Technology, Sweden

[12] Bernard S, Heutte L, Adam S, On the Selection of Decision Trees in Random Forest, Proceedings of International Joint Cobference on Neural Networks, Atlanta, Georgia, USA, June 14-19,302-307, (2009)

[13] Brieman L, Random Forests, Machine Learning, 45, 5-32, (2001)

[14] Simon Nusinovicia, Yih ChungTham, Marco YuChak Yan, Daniel Shu Wei Ting, JialiangLi, Charumathi Sabanayagam, Tien Yin Wong, Ching-Yu Cheng

[15] M. Malvoni, M. G. De Giorgi, and P. M. Congedo, "Data on support vector machines (SVM) model to forecast photovoltaic power," Data in Brief, vol. 9, no. C, pp. 13–16, 2016.

[16] Eibe Frank, Leonard Trigg, Geoffrey Holmes, Ian H. Witten. 2000. Technical Note: Naive Bayes for Regression. Machine Learning, 41, 5-25, Kluwer Academic Publishers.

[17] R. Darnag, B. Minaoui, and M. Fakir, "QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression," Arabian Journal of Chemistry, vol. 10, no. S1 pp. S600–S608, 2017.

[18] T. Singh, F. Di Troia, and C. Aaron Visaggio, "Support vector machines and malware detection," Journal of Computer Virology & Hacking Techniques, vol. 41, no. 10, pp. 1–10, 2016.

[19] Global Refinement of Random Forest Shaoqing Ren Xudong Cao Yichen Wei Jian Sun University of Science and Technology of China