A Survey of Machine Learning Algorithms

Jayakumar Sadhasivam[1], Arpit Rathore[2],Indrajit Bose[3], Soumya Bhattacharjee[4], Senthil Jayavel[5]

[1234]*School of Information Technology and Engineering (SITE), Vellore Institute of Technology, Vellore, India.*
[5]*Computer Science and Engineering, Nandha Engineering College,Erode, India.*

Abstract - *Today machine-learning algorithms provide an evident way to predict the assertive outcomes of different fields of datasets like healthcare, stock exchange, population statistics etc. In this research paper, we are reviewing these three supervised type machine-learning algorithms like Support Vector Machine (SVM), Random Forest and Naïve Bayes algorithms. Give our illustrative outlook on these algorithms.*

Keywords – *Dataset, SVM, Random Forest, Naïve Bayes, Decision Tree, Training Dataset, Testing Dataset*

I. INTRODUCTION

Machine learning is a part of the data science, which do predictive analysis on the given data and provide us outcomes by making automatic learning by searching patterns in the data, which is accessing by its algorithms and programmed accordingly. The machine learning algorithms classify into two major categories are:-

A. Supervised Learning

This is the machine-learning task in which first, we make training dataset and this dataset will be applied to the testing dataset to get assertive outcomes.

It generally uses mainly two techniques like classification and regression.

B. Unsupervised Learning

Unsupervised learning is a type of machine learning algorithm in which there is no trained dataset, available algorithms are used to make their own patterns in given datasets.

This is classified into two techniques like association and clustering.

C. Support Vector Machine (SVM)

The Vladimir Vapnik, Bernhard Boser and Isabelle Guyon developed support Vector Machine (SVM) in the year 1992. [1]

It is a classification method, which applies to the linear and non-linear dataset. In this algorithm, we do mapping on the two-dimensional dataset and find outs the hyperplane in vectors using support vectors and margins in the dataset.

The Lagrangian formulation is used to find out the minimal marginal hyperplane (MMH) [2]

$$\boldsymbol{d(X^T) = \sum_{i=1}^{l} y_i \propto_i X_i X^T + b_0}$$

$\mathbf{X_i}$and $\mathbf{X^T}$is the test tuple

$\boldsymbol{y_i}$is the class label of the support vector.

$\propto_i, \boldsymbol{b_0}$are numeric parameters

D. Random Forest

Random Forest Algorithm is the most sought-after machine-learning algorithm. It is used for both Classification and Regression problems. Like other machine learning algorithms, it is used to predict the assertive outcomes based on the labeled data given. Many decision trees combine to form a forest, which is the prime focus of Random Forest algorithm. The number of trees the more powerful the forest is or in other words we can say the number of trees the more accurate the results after aggregating will be.

E. Naïve Bayes

The Naive Bayes algorithm is a probabilistic classifier that determines an arrangement of probabilities by checking the rotation and mix of qualities in a given dataset.Naive Bayes demonstrate is anything but difficult to build and especially helpful for candid information collections. Alongside inactivity, Naive Bayes is known to highly sophisticated classification methods. Bayes theorem gives a method for calculating posterior probability P (c|x) [2]

$$\boldsymbol{P(c|x) = \frac{P(x|c)P(c)}{P(x)}}$$

$$\boldsymbol{P(c|x)} = p(x_1|c) * p(x_2|c) * \dots * P(x_n|c) * p(c)$$

[3]

$\boldsymbol{P(x)}$ Predictor prior probability

$\boldsymbol{P(c)}$ Class prior probability

$\boldsymbol{P(x|c)}$ Likelihood

II. LITERATURE SURVEY

S.NO	STUDIES	LANGUAGE	ALGORITHM	DESCRIPTION	DATA SOURCE
1.	Wenying Zhang,Huaguang Zhang,Jinhai Liu, Kai Li, Dongsheng Yang and Hui Tian [3]	English	Multiclass Support Vector Machine (SVM).	According to the Zhang and his team [3], the Support Vector Machine algorithm (SVM) will help to remove the manual checking of each photovoltaic panels and reduce the cost of manufacturing of photovoltaic panels by using its proposed model with the particle-sworn optimization for more accurate results.	The weather dataset was taken for this experiment, which consists 215 samples for training set and 150 samples for testing set.
2.	Wen Wu and Hao Zhou[4]	English	Standard Support Vector Machine including Recursive Feature Elimination and Principal Component Analysis	Cervical cancer detection [4] in early stages is still very hard to find out, Wu and Zhou proposed a model using Support Vector Machine (SVM) with SVM-REF and SVM-PCA provide a way to get early detection results with minimal cost for the patients.	The dataset contains the 668 patient's medical history in 30 different tests.
3.	Radhika R Halde, Arti Deshpande and Anjali Mahajan [5]	English	Neural Network and Decision Tree algorithms.	Student's performance is very critical virtue to estimate, Halde and her team[5] suggests that using neural network and decision tree algorithms with several inputs from students will predict that student can pass the final examinations or not.	The student dataset consists data of the 150 samples which they gave answers of questionnaire which includes 98 questions.
4.	Sonia Saini, Shruti Kholi [6]	English	Statistical Metrics like Receiver Operating Curve (ROC) and Area Under Curve (AUC)	E-Health tips on social media websites is growing rapidly, Saini and her team [6] was integrating machine learning algorithms with data mining techniques with provide more efficacious predictive analysis for the any convulsion of pandemic diseases.	Pandemic disease dataset.
5.	E Deepak, G Sai Pooja, R N S Jyothi, S V Phani Kumar, K V Kishore [7]	English	Support Vector Machine (SVM) with several kernel methods.	Predictive faculty performance using several machine learning algorithms like Naïve Bayes, AdaBoost, J48 and Clojure algorithms which provide more accurate results than other set of algorithms of machine learning.[7]	The data sets for this experiment are taken from university and compared with the faculty data set of faculties of different colleges. The

					database of faculty has 487 instances.
6.	Kevin Joy Dsouza and Zahid Ahmed Ansari [8]	English	Support Vector Machine (SVM)	According to the Dsouza and Ansari [8] using, the Support Vector Machine (SVM) algorithm for the classification of the most common cancer found in the female is Breast cancer using the popular kernels, which is incorporated with the algorithms	The breast cancer dataset has been taken from UCI repository, which contains the 569 samples, which have more than 30 features used for the classification of cancer using SVM algorithm.
7.	Rui Ren, Desheng Dash Wu and Tianxiang Liu [9]	English	Sentiment Analysis used with the Support Vector Machine (SVM)	The stock market is a volatile entity, which depends on the several factors. Ren and his team members [9] proposed a model using Support Vector Machine (SVM) with the sentimental analysis provide a way to get the drift of the stock exchange market.	The stock dataset has been taken from the SSE 50 index.
8.	ROgerio Galante Negri, Erivaldo, Abtonio da Silva and Wallace Casaca [10]	English	Support Vector Machine (SVM).	Contextual Classification need of the generation to provide many artificial intelligence projects easy. Negri and his team [10] suggests that using support vector machine algorithm image classification showed tremendous results using validation and using various SVM kernels.	Image dataset used for the classification.
9.	Syed Mehedi Hasan Nirob, Md. Kazi Nayeem and Md. Saiful Islam [11]	English	Feature extraction (Cross-Validation) and Support Vector Machine (SVM)	Nirob and his research team [11] done work on the Bangla language question classification that is done using feature selection with the help of the cross-validation technique and classification on the resultant features.	Lexicon of the most frequent 300-400 words is used for the classification.
10.	Xiaofeng Ma and Zhurong Zhou [12]	English	Support Vector Machine (SVM)	Ma and Zhou [12] has done the prediction for the passing rate of the students using the cross-validation and Support Vector Machine (SVM) using the student academic dataset.	Student academic dataset from UCI.
11.	Ruhi Mahajan, RishikesanKamal	English	Random Forest	Detection of the cardiac problem from	The dataset contains 8528

	eswaran, John Andrew Howe and oguzAkbilgic [13]			electrocardiogram (ECG) manually is not always correct and time taken. Therefore, Mahajan and her team of researchers [13] worked in the dataset to produce a model using cross validation as well as random forest algorithm to achieve maximum accurate results.	samples of electrocardiogram (ECG) recording of the patients.
12.	Muhammad Mahmudun Nabi, Mohammad Tanzir Altaf, Sabir Ismail [14]	Bengali	TF-IDF (time frequency inverse document frequency)	Nabi and his team [14] performed sentimental analysis on Bengali texts to determine whether the particular texts depicts any positive negative or neutral sentence by using TF-IDF text mining method and with the use of this technique patterns are found from the sentences which classifies the sentences categorically.	The data set used in this particular paper is taken from various social sites and are Bengali text comments. There are about 1500 sentences.
13.	Dengju Yao, Jing Yang, Xiaojuan Zhan [15]	English	Random Forest Algorithm , Multivariate Adaptive Regression Splines(MARS)	Yao and his team [15] proposed that the combination of both the random forest algorithm and the Multivariate adaptive regression (MARS) technique to predict the survivability of chances of breast cancer. Random forest when combined with MARS technique proves to be much more efficient than Random forest alone and less efficient than MARS.	The data set was obtained from Irvine Machine Learning repository. The dataset consists of 569 samples with 32 attributes.
14.	PetreLamenski, EftimZdravevski, SasoKoceski, Andrea Kulakov and Vladimir Trajkovik. [16]	English	Game theory, Random Forest algorithm	The false alarms in ICU's in hospitals cause hindrance in calculating better reaction time for any medical personnel, avoids this situation, an approach where datasets are manually annotated alarms using data mining technique and Random Forest Algorithm the false alarms were considerably suppressed which helped to achieve better reaction time of any medical personnel. [16]	The data set used in this paper are the annotated alarms from the MIMIC 2 waveform database. The manual annotation describes that the readings before the alarm are taken and considered whether they are true or false.

15.	Radhika R Halde [17]	English	Random Forest	Predictive analysis in an educational institute to predict retention power of students, success rate of students that helps to determine the performance of the students and it will reduce the exam failure rates. Decision trees predicts the performance of the UG and PG students. [17]	Data collected by surveying 60 students of Thadomal Shahani Engineering college by making them fill questionnaire. The data set consists of first name, last name and the cgpa obtained by the students.
16.	NazeehGhatasheh [18]	English	Random Forest	Ghatasheh [18] suggests which machine-learning algorithm is the best one to find out the credit risk of customers who are unable to pay the loans borrowed from the banks. In the research conducted, it is observed that Random Forest algorithm is the best algorithm for prediction of the credit risk because of its accuracy and simplicity.	Data set used is the German credit dataset.
17.	Ghada Soliman, Ahmed Misbah, Ala’a El-Nabawy and SeifEldawlatly [19]	English	Random Forest	Basketball is one of the most popular game in the world and when we talk of basketball, NBA now comes as a synonym. Soliman and his team [19] worked on the prediction of the all-star players of the NBA by utilizing the random algorithm and got a 92.5 % accuracy.	NBA basketball dataset of duration of 1937 to 2011 for prediction of the all-star players in NBA.
18.	Chenguang Wang, Xueling Dong, Limin Yu and Weifen Zhuang [20]	English	Random Forest	Infant’s health is a very important concern for any country development. Therefore, Wang and his team of researchers [20] developed a prediction model of the number of hospitalization days using random forest algorithm.	Infant’s health dataset from UCI.

19.	Yi Hou, Praveen Edaraand Yohan Chang [21]	English	Random Forest	Now day's traffic is day by day increasing so that Hou and his team [21] come up with new time travel techniques which using random forest to predict time required between two places.	Data was collected from using Regional Integrated Transportation Information System and Nokia here in the St. Louis region between 2014-2016.
20.	Muhammad Asif Manzoor and Yasser Morgan [22]	English	Random Forest	Manzoor and Morgan [22] using the random forest Algorithm developed the vehicle identification system. As an output, we can know the vehicle year of production and its model.	The NTOU-MMR dataset has been used to develop this system.
21.	Selina S.Y., Yinjiao Xing, Kwok L. Tsui [23]	English	Naive Bayes	Prediction of remaining useful life of batteries. Some of the parameters that shows the conditions such as Full Charge Capacity, charging status and status of the battery's health. [23]	NASA ames li-ion battery cycle text data.
22.	Vivek Narayanan, Ishan Arora, Arjun Bhatia [24]	English	Naive Bayes	Bhatia and his team [24] done Sentiment analysis based on the factors such as Negation Handling, Laplacian moothing, Feature selection.	Data of 25000-movie review form Internet Movies Database.
23.	Jun Zhang,Chaochen, Yang Xiang, wanlen Zhou, Yong Xiang [25]	English	Naive Bayes	Xiang and her team [25] was done hypothetical investigation on why and how the proposed conspire works. Bag of Flows (BoF-NB) technique was additionally proposed to provide the total connection Naive Bayes (NB) forecasts.	Real world traffic dataset for evolution.
24.	Garima Singh, Kiran Bagwe,Shivani Shanbhag,Shraddha Singh,Sulochana Devi [26]	English	Naive Bayes	The application is proposed by Singh and her team [26] is going to evaluate various health evaluation categorises on that basis coronary heart disease will be predicated.	Some training and testing dataset they used.

25.	Claire Gallagher,Michale G. Madden, Brian D.Arcy [27]	English	Naive Bayes	Arcy and his team [27] proposed an approach has an exactness of 90.6% in foreseeing whether deals will be won or lost.	HP ES's data warehouse that supply raw sales data.
26.	Peixin Liu, Hongzhi Yu, Tao Xu and Chuanqi Lan [28]	English	Naïve Bayes	The text classification is done on the archives collected from the Gansu Province, China. Liu and her team [28] used TFIDF algorithm for the feature selection in the documents and then applied random forest algorithm for text classification.	The archives of Gansu province of the China.
27.	MykhailoGranik and Volodymyr Mesyura [29]	English	Naïve Bayes	Granik and Mesyuru [29] developed the model for the detection of the fake news using Naïve Bayes algorithm. This model is applied to the Facebook news posts which they have achieved the 74% accuracy.	The data is collected from the Facebook and used it for the detection of the fake and real news.

III. FINDINGS

Support Vector Machine (SVM) can be used for classification as well as regression methodologies. Mostly we use SVM with classification methodology. SVM performs better with a large number of evaluation points and separating planes between data points to provide clear virtue of the outcome of the testing datasets like predicting of shapes, character to face recognition etc. The major drawback with the SVM is that which type of kernel is suitable to use with SVM so that it can be applied on the particular datasets.

After reviewing, five research papers on Random Forest algorithm. After reviewing those papers, I understood the various fields in which random forest is used extensively to get the desired outcomes. Although machine-learning algorithms have many applications in health sector, education sector, robotics etc. I found out that amongst all the machine-learning algorithms Random Forest is the only algorithm, which works most efficiently and accurately. Whether it is to predict the disease survivability of a patient prediction of credit risk or prediction of the weather forecast for each of these random forest works best.

After going through a literature review on Naive Bayes, I realized that this technique is simple and accurate for prediction using a training dataset. We prefer it for multiple classes. Naive Bayes uses Bayes theorem to determine probabilistic classification on different fields like sentiment analysis, Heart disease prediction, Email spam filtering, Text classification and mainly where a big-trained dataset is given to test the probability on a testing dataset. One of the most important feature of Naive Bayes, that it can deal with missing information and requires less time to reach posterior probability.

IV. CONCLUSION

We come to this conclusion that depending on the distribution of different data sets we can determine which machine learning algorithm works best. For e.g. for any small data sets support vector machine (SVM) works best than Random Forest Algorithm. Again, when there are complex datasets or when there are large data sets Random Forest proves to be a better algorithm than the previous one.

V. REFERENCES

[1] Boser, B., Guyon, I., &Vapnik, V. (1992). "*A training algorithm for optimal margin classifiers. Proceedings Of The Fifth Annual Workshop On Computational Learning Theory - COLT '92*". doi: 10.1145/130385.130401

[2] Han, J., &Kamber, M. (2012). "*Data mining (3rd ed.)*". Haryana, India: Elsevier.

[3] Zhang, W., Zhang, H., Liu, J., Li, K., Yang, D., & Tian, H. (2017). "*Weather prediction with multiclass support vector machines in the fault detection of photovoltaic system*". IEEE/CAA Journal Of AutomaticaSinica, 4(3), 520-525. doi: 10.1109/jas.2017.7510562

[4] Wu, W., & Zhou, H. (2017). "*Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches*". IEEE Access, 5, 25189-25195. doi: 10.1109/access.2017.2763984

[5] Halde, R., Deshpande, A., & Mahajan, A. (2016). "*Psychology assisted prediction of academic performance*

using machine learning". 2016 IEEE International Conference On Recent Trends In Electronics, Information & Communication Technology (RTEICT). doi: 10.1109/rteict.2016.7807857

[6] S. Saini and S. Kohli, "*Machine learning techniques for effective text analysis of social network E-health data*," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 3783-3788.

[7] Deepak, E., Pooja, G., Jyothi, R., Kumar, S., & Kishore, K. (2016). "*SVM kernel based predictive analytics on faculty performance evaluation*". 2016 International Conference On Inventive Computation Technologies (ICICT). doi: 10.1109/inventive.2016.7830062

[8] Dsouza, K., & Ansari, Z. (2017). "*Experimental Exploration of Support Vector Machine for Cancer Cell Classification*". 2017 IEEE International Conference On Cloud Computing In Emerging Markets (CCEM). doi: 10.1109/ccem.2017.15

[9] Ren, R., Wu, D., & Liu, T. (2018). "*Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine*". IEEE Systems Journal, 1-11. doi: 10.1109/jsyst.2018.2794462

[10] Negri, R., da Silva, E., &Casaca, W. (2018). "*Inducing Contextual Classifications With Kernel Functions Into Support Vector Machines*". IEEE Geoscience And Remote Sensing Letters, 1-5. doi: 10.1109/lgrs.2018.2816460

[11] Nirob, S., Nayeem, M., & Islam, M. (2017). "*Question classification using support vector machine with hybrid feature extraction method*". 2017 20Th International Conference Of Computer And Information Technology (ICCIT). doi: 10.1109/iccitechn.2017.8281790

[12] Ma, X., & Zhou, Z. (2018). "*Student pass rates prediction using optimized support vector machine and decision tree*". 2018 IEEE 8Th Annual Computing And Communication Workshop And Conference (CCWC). doi: 10.1109/ccwc.2018.8301756

[13] Mahajan, R., Kamaleswaran, R., Howe, J., &Akbilgic, O. (2017). "*Cardiac Rhythm Classification from a Short Single Lead ECG Recording via Random Forest*". 2017 Computing In Cardiology Conference (Cinc). doi: 10.22489/cinc.2017.179-403

[14] Nabi, M., Altaf, M.T., Ismail, S., Hasan, A.S., Islam, M.S., Mashrur-E-Elahi, G.M., Izhar, M.N., Al-Mahmud, Mondal, A., Saha, A., Islam, M.A., Hasan, K.M., Rahman, M.M., Fukuhara, T., Nakagawa, H., Hatzivassiloglou, V., & Pang, B. (2016). "*Detecting Sentiment from Bangla Text using Machine Learning Technique and Feature Analysis*".

[15] Yao, D., Yang, J., & Zhan, X. (2011). "*Predicting breast cancer survivability using random forest and multivariate adaptive regression splines.*" Proceedings Of 2011 International Conference On Electronic & Mechanical Engineering And Information Technology. doi: 10.1109/emeit.2011.6023012

[16] Lameski, P., Zdravevski, E., Koceski, S., Kulakov, A., &Trajkovik, V. (2017). "*Suppression of Intensive Care Unit False Alarms based on the Arterial Blood Pressure Signal*". IEEE Access, 1-1. doi: 10.1109/access.2017.2690380

[17] Halde, R. (2016). "*Application of Machine Learning algorithms for betterment in education system*". 2016 International Conference On Automatic Control And Dynamic Optimization Techniques (ICACDOT). doi: 10.1109/icacdot.2016.7877759

[18] Ghatasheh, N. (2014). "*Business analytics using random forest trees for credit risk prediction: A comparison study*". International Journal of Advanced Science and Technology, 72(2014), 19-30.

[19] Soliman, G., El-Nabawy, A., Misbah, A., &Eldawlatly, S. (2017). "*Predicting all star player in the national basketball association using random forest"*. 2017 Intelligent Systems Conference (Intellisys). doi: 10.1109/intellisys.2017.8324371

[20] Wang, C., Dong, X., Yu, L., Ye, L., Zhuang, W., & Ma, F. (2017). "*Prediction of days in hospital for children using random forest*". 2017 10Th International Congress On Image And Signal Processing, Biomedical Engineering And Informatics (CISP-BMEI). doi: 10.1109/cisp-bmei.2017.8302287

[21] Hou, Y., Edara, P., & Chang, Y. (2017). "*Road network state estimation using random forest ensemble learning*". 2017 IEEE 20Th International Conference On Intelligent Transportation Systems (ITSC). doi: 10.1109/itsc.2017.8317743

[22] Manzoor, M., & Morgan, Y. (2018). "*Vehicle make and model recognition using random forest classification for intelligent transportation systems"*. 2018 IEEE 8Th Annual Computing And Communication Workshop And Conference (CCWC). doi: 10.1109/ccwc.2018.8301714

[23] Ng, S. S., Xing, Y., &Tsui, K. L. (2014). "*A naive Bayes model for robust remaining useful life prediction of lithium-ion battery*". Applied Energy, 118, 114-123.

[24] Narayanan, V., Arora, I., & Bhatia, A. (2013, October). "*Fast and accurate sentiment classification using an enhanced Naive Bayes model*". In International Conference on Intelligent Data Engineering and Automated Learning (pp. 194-201). Springer, Berlin, Heidelberg.

[25] Jun Zhang, Chao Chen, Yang Xiang, Wanlei Zhou, & Yong Xiang. (2013). "*Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions*". IEEE Transactions On Information Forensics And Security, 8(1), 5-15. doi: 10.1109/tifs.2012.2223675

[26] Singh, G., Bagwe, K., Shanbhag, S., Singh, S., & Devi, S. (2017). "*Heart disease prediction using Naïve Bayes*". International research Journal of Engineering and Technology, 4(03).

[27] Gallagher, C., Madden, M., & D'Arcy, B. (2015). "*A Bayesian Classification Approach to Improving Performance for a Real-World Sales Forecasting Application*". 2015 IEEE 14Th International Conference On Machine Learning And Applications (ICMLA). doi: 10.1109/icmla.2015.150.

[28] Liu, P., Yu, H., Xu, T., & Lan, C. (2017). " *on archives text classification based on Naive bayes*". 2017 IEEE 2Nd Information Technology, Networking, Electronic And Automation Control Conference (ITNEC). doi: 10.1109/itnec.2017.8284934

[29] Granik, M., &Mesyura, V. (2017). "*Fake news detection using naive Bayes classifier*". 2017 IEEE First Ukraine Conference On Electrical And Computer Engineering (UKRCON). doi: 10.1109/ukrcon.2017.8100379