# PERFORMANCE ANALYSIS OF SUPERVISED LEARNING ALGORITHMS ON DIFFERENT APPLICATIONS

Vijayalakshmi Sarraju[1], Jaya Pal[1] and Supreeti Kamilya[2]

[1]Department of Computer Science and Engineering,
BIT Extension Centre Lalpur, India
[2]Department of Computer Science and Engineering, BIT Mesra, India

***ABSTRACT***

*In the current era of computation, machine learning is the most commonly used technique to find out a pattern of highly complex datasets. The present paper shows some existing applications, such as stock data mining, undergraduate admission, and breast lesion detection, where different supervised machine learning algorithms are used to classify various patterns. A performance analysis, in terms of accuracy, precision, sensitivity, and specificity is given for all three applications. It is observed that a support vector machine (SVM) is the commonly used supervised learning method that shows good performance in terms of performance metrics. A comparative analysis of SVM classifiers on the above-mentioned applications is shown in the paper.*

***KEYWORDS***

*The supervised learning algorithm, stock data mining, undergraduate admission scheme, breast lesion detection, and performance analysis.*

## 1. INTRODUCTION

Machine learning is popular today; finding patterns and relationships within highly complex datasets is used in today's era. Several machine learning techniques are making real-world applications possible with recent storage and computational capabilities developments. Machine learning requirements are increasing day by day due to the ever-growing data sets. In 1959, machine learning was first coined by Arthur Samuel, who is an eminent pioneer in the field of artificial intelligence and gaming technology [1]. Machine learning is the branch of artificial intelligence that uses data and algorithms to imitate intelligent human behaviour. In general, machine learning is categorised into three types of learning methods: supervised learning, unsupervised learning and reinforcement learning. Supervised learning needs a trained data set with labelled data [2] with a known output value. Unsupervised learning does not use the training data set. The most common example of an unsupervised learning algorithm is clustering [3]. In Reinforcement learning, the input data from the environment is used as a stimulus to determine how the model should react [2]. In brief, supervised learning requires the training of labelled data with inputs and desired outputs. The labelled input data set of supervised learning is used to train the algorithm. The algorithm improves its estimates by making exact predictions in this training process and re-iterates the algorithm until it achieves the desired level of accuracy. Classification and regression problems are solved through supervised learning [2]. In a regression problem, we try to predict the results within a continuous output, which means that we try to plot input

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 29-35, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121903

variables to some continuous function. In a classification problem, we try to predict the results in discrete output. Supervised learning is used successfully in different fields such as data mining [4], pattern recognition [3], the internet of things [3], health monitoring [3] and market analysis [5]. Three significant and different types of works on supervised learning techniques are based on stock market analysis [5], admission schemes in educational institutes [6] and breast lesion detection [7], where the supervised learning algorithms are used to predict future outcomes based on historical data. Analysing different supervised learning algorithms on applications from different fields is the main motivation of the current work.

This paper is focused on the performance analysis measures of various supervised learning algorithms, such as linear discriminant (LD), logistic regression (LR), naïve Bays (NB), support vector machine (SVM), K-nearest neighbour (KNN), and random forest (RF) methods in the applications of stock market analysis, undergraduate admission data and breast lesion detection. The performance analysis of the said algorithms is first done on the established works of [5]–[7]. Among the various supervised algorithms, SVM is shown to be a good classifier that can be commonly applied to all the aforementioned applications. Therefore, a comparative study of SVM statistics on the said applications is presented in the paper. Section II gives a brief description of machine learning and different supervised learning algorithms. Three applications of supervised learning techniques along with the performance measurements are provided in Section III. Section IV provides a discussion about the performance analysis of the algorithms and a comparative analysis of SVM on three different applications. Finally, the paper is concluded in Section V.

## 2. SUPERVISED MACHINE LEARNING ALGORITHMS

Machine learning (ML) adds a new dimension to technology where a computer performs any task without human intervention on the basis of constantly learning from past experiences. Therefore, ML has three features: (1) Task, (2) Performance measure, and (3) Experience [8]. Supervised learning (SL) is a type of machine learning in which a function which converts input to output is learned using examples of input-output pairs. Based on labelled training data that comprises training examples, it infers a process. Supervised learning uses labelled datasets to train algorithms to recognise data or properly predict outcomes. The two different tasks in machine learning are classification and regression. In classification, the labelled data is discrete, but in the regression, it is continuous. In the case of unsupervised learning, the training data is not labelled. A classifier is designed by deducing existing clusters in the training data set. The supervised learning algorithms, used in the paper for performance analysis, are discussed as follows. Naive Bayes (NB) is a statistical classification approach used in supervised algorithms. The Bayesian classification's key benefit is that it can handle predicted difficulties. The Bayes theorem is used to underpin this categorization method. The moniker "Naive" stems from the algorithm's strong assumption that all input features are independent of one another and have no correlation in simple probabilistic classifiers like Naive Bayes. Because it is a probabilistic model, Naive Bayes provides a posterior probability of belonging to a class given input features.

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}$$

Where A and B are two independent events, P(A) and P(B) are the probabilities of A and B. The conditional probability, P(A/B), is the probability of an occurrence A given an event B. P(B/A) denotes the chance of seeing an event B if A is true.

Logistic regression (LR) is used to determine how much possibility is there for cases where the event is successful, and the identical event is unsuccessful. When the dependent variable is in binary form (having just one of two values), logistic regression is used. That means, it can only have two possible values [9]. Because maximum likelihood calculations are less accurate in small sample sizes than simple least squares and only large sample sizes are needed in logistic regression. The relationship between the dependent and independent variables does not have to be linear.

Because of the ease of interpretation and short calculation time, the K-nearest neighbour (KNN) method is well-known for its simplicity. It saves available cases and categorises new instances based on homogeneity, similar to the distance function [10]. A majority vote of the object's neighbours determines its classification and this process is known as class integration. Following that, the object is assigned to the class with the highest similarity among the K nearest neighbours [10].

A decision tree (DT) acts like a flowchart that categorises instances depending on their characteristics. Each internal node represents a test case; branches show the results of the tests, and leaf nodes show class labels. This strategy will perform better when there are discrete characteristics [11]. In the simplest situation, each test considers a single attribute, and the instance space is partitioned based on the attribute's value. The condition refers to a range in the case of numeric properties.

A support vector machine (SVM) is an ML algorithm that solves problems of classification and regression. The SVM model is supported by the margin calculation idea. This method can be applied to both linear and nonlinear data. Each and every data observation is plotted as a point in n-dimensional space by this algorithm where n represents the number of features. Each feature's value is the matching coordinate's value. It divides the training datasets into classes by finding a line (hyperplane) that divides them. It maximises the margin between the nearest data point (in the classes) and the hyperplane.

Random forest (RF) is a special kind of ensemble learning algorithm used in classification and regression problems. A random forest is a forest of trees, each based on a different bootstrap sample from the training data. When a tree is fitted, some predictor variables are censored at each node. The optimal split is then determined using random forests based on the predictor variables chosen. After the trees have been voted on, they are grown to their full depth and an agreement prediction is made. It creates complex models with high predicted accuracy and highlights the significance of each variable in the categorization model.

A perceptron is a single-layer neural network with weights and biases that may be trained to produce the correct target vector when an acceptable input vector is provided. The training method employed is the perceptron learning rule. Perceptron is mostly used to solve simple pattern classification problems.

Multi-layer perceptron (MLP) networks contain, in general, three layers. Each layer, with the exception of the input layer, operates like a neuron and uses a non-linear activation function. Back-propagation, a supervised learning approach, is used to train the MLP network. The MLP is distinguished from a linear-perceptron by its several layers and non-linear activation function, which allows it to recognise non-linear data. Wide applications of MLPs include speech recognition and speech enhancement.

## 3. APPLICATIONS

This section discusses three different applications of various supervised learning algorithms. For the applications, we consider the works of [5]–[7] for stock data mining, undergraduate admission scheme and breast lesion detection, respectively.

In [5], the authors have proposed the work on the performance analysis of twelve years of renowned bank stock data. The performance of 4 different SL approaches, that is, SVM, RF, NB, and ANN (artificial neural network), are compared for the particular study. The authors use the values of correctly and incorrectly classified instances, which are used to calculate the performance of predictive classification models. Various statistical metrics, such as accuracy, precision, sensitivity, and specificity of different supervised classifiers are analysed by a renowned Bank's stochastic data set.

In [6], the authors study undergraduate admissions applications. The work focuses on how machine learning techniques can assist admission counsellors in concentrating their efforts on applicants who are more seemingly to join. The admission process is carried out in three stages: stage 0, 1 and 2. In stage 0, the applicants fill up their application forms. In stage 1, admissions counsellors evaluate and select some of the applicants. The applicants accept the requests of counsellors in stage 2. SL techniques are employed to classify stages 1 and 2. Predictive modelling consists of three phases: 1. data processing, 2. classification, 3. feature selection. In data processing, raw data provided by the admission cell is converted into a numerical form and accepted by a classifier. The classifiers are trained to make valuable predictions about future applicants in the classification phase. In the third phase, classifiers are used to determine the features in an application that are dominant in stage 1 and 2 predictions. Classifiers are used to predict stages 1 and Stage 2 relevant to the admission scheme, as discussed in [6]. The first stage indicates whether or not an applicant is accepted, whereas the second indicates whether or not an approved applicant attends the university. In the training process, five classifiers, such as MLP, Perceptron, linear SVM (LSVM), polynomial SVM (SVM POLY) and quadratic SVM (SVM RBF) are trained, and different statistics have been measured to find out the classifier with the highest performance rate. The performance rate is considered in the validation set and the same classifier is used in testing the current data.

In [7], the authors have presented an automatic computer-aided detection and diagnosis system of breast lesions. The developed model is dependent on supervised machine learning algorithms. These techniques are used to classify benign and malignant localised breast lesions. Four different machine learning algorithms are examined for classification: (1) support vector, (2) k nearest neighbours, (3) random forest, and (4) naive Bayes classifiers. The evaluation metrics used to demonstrate the study's success are accuracy, sensitivity, specificity, and precision.

## 4. DISCUSSION

In the presented work, we discuss the performance analysis of supervised learning techniques in different applications.

From Table I, performance analysis of supervised learning algorithms is observed on various statistics such as accuracy, precision, sensitivity, and specificity. In our comparative analysis, we observe the accuracies of different algorithms. In the case of a stock marketing application, the accuracy for LD is 53.6%, LR is 99.8%, NB is 48.3%, RF is 45.9% and LSVM is 99.1% respectively. Here, we found that the LR technique is the outstanding one out of all the above-mentioned techniques. In the case of the undergraduate admissions application, MLP, perceptron,

LSVM, polynomial support vector machine (SVM POLY), and Quadratic support vector machine (SVM RBF) algorithms are implemented. Overall, we observe that MLP is showing better accuracy than the other algorithms. The third application mentioned in Table I is breast lesion detection. Comparing the accuracies of SVM, ANN, NB and RF algorithms, the RF accuracy is the most accurate that decides tumour aggressiveness. The best precision value classifier in the stock data mining application is LR with 99.8%. Almost all the classifiers show

Table 1. Performance Analysis of Classifier in Three Different Applications

| Sl. no. | Reference Applications | Supervised Learning Algorithms | Accuracy (%) | Precision (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|---|---|
| 1 | Stock Data Mining | Linear Discriminant | 53.61 | 74.58 | 40.13 | 14.04 |
| | | Logistic Regression | 99.82 | 99.86 | 53.24 | 0.18 |
| | | Naive Bays | 48.33 | 31.68 | 16.99 | 37.84 |
| | | Random Forest | 45.97 | 97.93 | 52.51 | 1.77 |
| | | LSVM | 99.12 | 98.76 | 52.95 | 1.03 |
| 2 | Undergraduate Admission | MLP | 94.57 | 95.62 | 94.36 | 94.82 |
| | | Perceptron | 94.32 | 95.69 | 93.71 | 94.04 |
| | | SVM POLY | 94.36 | 95.67 | 93.87 | 94.94 |
| | | SVM RBF | 94.44 | 95.66 | 94.01 | 94.95 |
| | | LSVM | 94.45 | 95.70 | 94.05 | 94.93 |
| 3 | Breast Lesion Detection | SVM | 82.15 | 81.0 | 100 | 56.66 |
| | | KNN | 79.36 | 80.0 | 96.8 | 20 |
| | | RF | 90.36 | 92.0 | 96.25 | 83.33 |
| | | NB | 87.82 | 86.0 | 100 | 66.67 |

the same precision value in undergraduate data. However, in the case of breast lesion detection application, the highest precision value classifier is RF with 92%. In comparing maximum and minimum values of sensitivity and specificity in a stock data mining application classifiers, LR is the highly sensitive classifier and Naïve Bayes is less sensitive. Whereas, in the case of specificity, the lowest one is LR and the highest one is NB with 18% and 37.8%, respectively. Let us now compare the values of sensitivity and specificity in the undergraduate application classifiers.

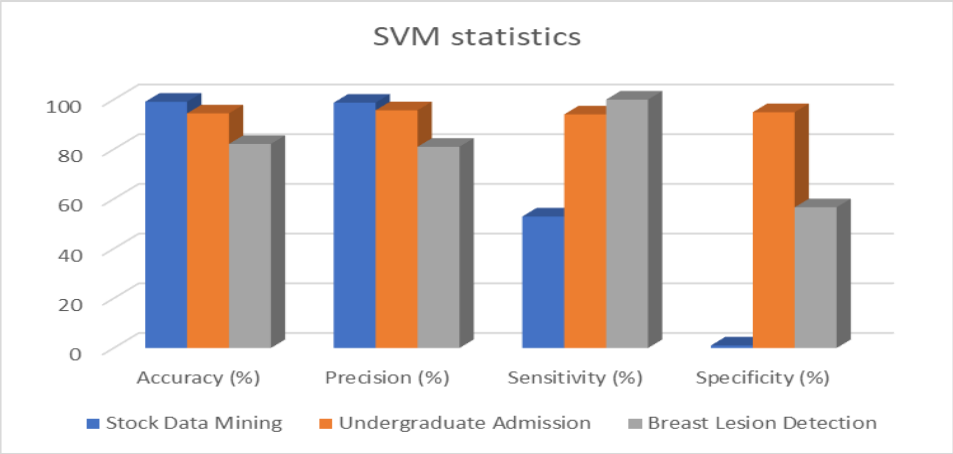


Fig. 1. Comparative analysis of the performance measures of SVM algorithm for the three applications

The highest sensitive rate classifier is MLP with 94.36%, however, specificity is almost identical in all classifiers. In the breast lesion detection application, the SVM classifier has a sensitivity of 100%, and the highest value of the specificity classifier is RF with 83.3%. From the table, we can make one comparative analysis of the three mentioned applications using the SVM classifier (as shown in Fig. 1). The SVM classifier shows better accuracy in the stock data mining application, compared to the other two applications. In terms of other statistics, such as precision, sensitivity and specificity of SVM analysis from the above graph, we found that there is not much variation in the precision value of stock data mining application and undergraduate data application but in comparison to these two applications, moderate interpretation of precision is found in breast lesion detection. Hence, we can say that the precision value is better in the stock data mining application on the overall analysis. Sensitivity is higher in breast lesion detection, lesser in stock data mining and moderately high in undergraduate admission data. Specificity is less in stock data mining applications than in the other two, but quite good in breast lesion detection.

## 5. CONCLUSION

Performance analysis of different supervised learning algorithms on three existing applications, such as stock data mining, undergraduate admission data and breast lesion detection, have been discussed in the work. It has been observed that, the support vector machine is a commonly used algorithm among all the supervised algorithms and shows good results in terms of different performance metrics. In this paper, we have compared SVM for all three applications. However, some more applications are needed to be explored for detailed analysis of the algorithms. As a future scope of the study, more applications can be used to analyse the best-suited supervised algorithm for classification.

## REFERENCES

[1] A. L. Samuel, “Some studies in machine learning using the game of checkers. ii—recent progress,” Computer Games I, pp. 366–400, 1988.

[2] S. B. Kotsiantis, I. Zaharakis, P. Pintelas et al., “Supervised machine learning: A review of classification techniques,” Emerging artificial intelligence applications in computer engineering, vol. 160, no. 1, pp. 3–24, 2007.

[3] U. S. Shanthamallu, A. Spanias, C. Tepedelenlioglu, and M. Stanley, “A brief survey of machine learning methods and their sensor and iot applications,” in 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA). IEEE, 2017, pp. 1–8

[4] N. Saleem and M. I. Khattak, “A review of supervised learning algorithms for single channel speech enhancement,” International Journal of Speech Technology, vol. 22, no. 4, pp. 1051–1075, 2019.

[5] M. Sharma, S. Sharma, and G. Singh, “Performance analysis of statistical and supervised learning techniques in stock data mining,” Data, vol. 3, no. 4, p. 54, 2018.

[6] T. Lux, R. Pittman, M. Shende, and A. Shende, “Applications of supervised learning techniques on undergraduate admissions data,” in Proceedings of the ACM International Conference on Computing Frontiers, 2016, pp. 412–417.

[7] F. Mutlu, G. Cetinel, and S. Gul, “A fully-automated ¨ computer-aided breast lesion detection and classification system,” Biomedical Signal Processing and Control, vol. 62, p. 102157, 2020.

[8] J. Alzubi, A. Nayyar, and A. Kumar, “Machine learning from theory to algorithms: an overview,” in Journal of physics: conference series, vol. 1142, no. 1. IOP Publishing, 2018, p. 012012.

[9] N. Pahwa, N. Khalfay, V. Soni, and D. Vora, “Stock prediction using machine learning a review paper,” International Journal of Computer Applications, vol. 163, no. 5, pp. 36–43, 2017.

[10] R. Choudhary and H. K. Gianey, “Comprehensive review on supervised machine learning algorithms,” in 2017 International Conference on Machine Learning and Data Science (MLDS). IEEE, 2017, pp. 37–43.

[11] D. Dhall, R. Kaur, and M. Juneja, “Machine learning: a review of the algorithms and its applications,” Proceedings of ICRIC 2019, pp. 47–63, 2020.

## AUTHORS

**Mrs. Vijayalakshmi Sarraju** is pursuing her PhD from Birla |Institute of Technology Extension Centre Lalpur. She completed her M. Sc in Mathematics from Nagarjuna University, Andhra Pradesh. She has 22 years of teaching experience. She has expertise in Statistics, Quantitative Techniques, discrete mathematical structures, and Numerical methods. Her research interest is inferential statistics and machine learning.

**Dr. Jaya Pal** is an assistant professor in the Department of Computer Science and Engineering of Birla |Institute of Technology Extension Centre Lalpur. She completed her M.Sc. and MCA in Mathematics and PhD in Technology. She has 16 years of teaching experience and 10 years of research experience. Her research interests are Fuzzy Logic & its Applications, Soft Computing, Machine Learning, Software Quality Prediction, and Data Mining.

**Dr. Supreeti Kamilya** is an assistant professor in the Department of Computer Science and Engineering of Birla |Institute of Technology Mesra. She completed her M.Tech and PhD in Computer Science from IIEST, Shibpur. She has 8 months of teaching experience and 7 years of research experience. Her research interests are Theoretical Computer Science, Chaos Theory, Machine Learning and Deep Learning.