Science in Information Technology Letters
Vol. 1., No. 2, November 2020, pp. 52-59
ISSN 2722-4139
http://pubs2.ascee.org/index.php/sitech

A study on the forecasting bigmart sales using optimized data mining techniques

N. M. Saravana Kumar [a,1,*], K. Hariprasath [b,2], N.Kaviyavarshini [c,3], A.Kavinya [c,4]

[a] Professor and Head, Department of Artificial Intelligence and Data Science, M. Kumarasamy College of Engineering, Karur, India
[b] Assistant Professsor, Departemnt of Information Technology, Vivekanandha College of Engineering for Women, Namakkal, India
[c] PG Scholar, Departemnt of Information Technology, Vivekanandha College of Engineering for Women, Namakkal, India
[1] saravanakumaar2008@gmail.com; [2] khariprasathit@gmail.com; [3] kaviyanaagarrajan97@gmail.com;
[4] kavinyasre21@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received August 02, 2020
Revised October 15, 2020
Accepted November 20, 2020

Keywords
Exploration
Naive bayes with PSO
Predictive modeling
Decision Tree
Data Analysis

ABSTRACT

Data mining is an in-depth study of enormous amounts of data present in an organization or institution's repository. Business experts mostly utilize data analytics approaches to confirm their opinion. It will rapidly boost the global interest of the organization. In this scenario, the information and conclusion are gathered from Data analysis by data analytics. The experts also use it to validate, diagnose, or authenticate speculate layouts suppositions and completion of the analysis. In this paper, the prediction is based on grocery data sets by inspecting and analyzing the big mart sales data set. Among several predictive algorithms, data mining algorithms are considered for prediction. It includes Decision Tree, Naïve Bayes, Adaboost with Particle Swarm optimization, and Random forest. The proposed method of this research is a novel Naïve Bayes with a PSO algorithm. This algorithm optimizes the model iteratively. Exploration of the data must be done before prediction. The root means squared error (RMSE) is used as evaluation metrics for comparing the data mining algorithms. The proposed algorithm performs well and gives a lower RMSE value. So, the proposed algorithm fits the best model when compared with the existing algorithms. This paper describes the prediction of high-quality data analysis data and determines the efficiency of data mining algorithms.

1. Introduction

In the education data, differentiation and classification have a significant role in the data mining approach, and this approach is utilized in an organized way. The data mining approach is used in modeling the user, user grouping, modeling the domain, profiling the user, and developing analysis [1]. Company sales and predicting sales have significant importance in any business for the guidances.

http://doi.org/10.31763/sitech.v1i2.167

Prediction is more critical for the development and improvement of a business. For the prediction, the process is to look into the condition previously followed and relates the interpretation of customer purchase, finds the shortage and strong point before the budgets are planned for the following year. Prediction is also made from past resources. Based on the result gained from the past resources business needs and circumstance of the business is increased. When a business takes the sales prediction as the first step, it would be a successful business [2].

The most significant business role is to estimate future sales, so the prediction of the past must be accurate for the company's development and improvement. Predictions help companies interpret past events, identify budget errors, and plan everything. By making the plan, the success rate is increased [3].

To create a better prediction model is impossible till the exploration of the dataset is unknown. So this point evaluates a strong base for the manipulation of data. After, the usage of various data mining algorithms comes for the prediction of big mart sales.

This paper is structured as follows: In Section 2, several topics such as literature study, methods, dataset descriptions, preprocessing steps, such as exploring data, data cleaning, and feature engineering, are discussed. The discussion also explains building models such as Decision Tree, Random Forest, Naïve Bayes, Random forest with particle swarm optimization (PSO), and Naïve Bayes with PSO. In Section 3, the results and discussion on the evaluation matrix of Root Mean Squared Error (RMSE) are discussed and followed by the discussion results. Finally, Section 4 concludes the study of forecasting big mart sales.

2.Method

Sadia et al. predict the rate of the stock market by evaluating the best form. In the investigative process, various presentations and objects must be taken as details and processes must not be split completely. So, all the data was re-processed then adjusted for analysis. There are two algorithms have used in this paper, namely, support vector machine and random forest. By evaluating both algorithms' accuracy, it is found that the random forest algorithm performed well suits for the prediction of the market price [4].

Nitin et al. focus on the repeated disruption in the exchange of stock. Despite it being impossible to calculate the stock exchange development with greater veracity, lost items during trading and the trading issues can be trimmed to rapid boost [5], [6] using the prediction of stock exchange calculation from the study of past data items. The authors have used three algorithms for the analysis purpose. They are support vector regression, linear regression, and polynomial regression. Finally, the support vector regression (SVR) algorithm performs well compared with the other two algorithms. Also, this algorithm minimizes the complexity of the performance [7].

Chandel et al. inspected the forecasting issues in commercial transactions on the internet and proposed a stacked generalization method consists of sub-level regressors. It then tested the results of single classifiers individually combined with the general model. This approach will predict much better when more data is used. Because the variation is not statistically important between the proposed model and random forest, the proposed method can forecast the demand due to its accuracy with lesser data. The XGBoost regressor and gradient-based algorithm are used to predict the sales of the big mart companies [8].

Kadam et al. proposes a software tool for forecasting future sales based on past sales data. Firstly, the raw data is analyzed and then preprocess the data before training. Two algorithms are compared to finalize the results. The algorithms are random forest and multiple linear regression. This system is used to predict big mart companies' future sales with multiple linear regression and random forest models [9].

Jain et al. analyze the various algorithms that are iteratively used for the association rule mining field. They are éclat and apriori algorithms. These algorithms are primarily used to extract data in the dataset and evaluate the various data's relationship by using the R programming language. R is a domain-based language [10]. It is mainly used for analytics purposes, especially for data exploration. The authors have examined the performance of the two algorithms on various data in the dataset. Also, analyze the time taken during the execution of the algorithms [11].

Punam et al. introduces the prediction of big mart sales products through two-level techniques. It is a statistical model that uses MAE values. It means the squared error value is evaluated by using various algorithms. Those are KNN, support vector regression, linear regression, and regression tree [2].

Hilda et al. interprets and compares the performance of the data mining tools. They are R, weka, and rapid miner. They analyze the performance in terms of their time-series analysis for visualization, handle statistical models, filtering, and others [12].

2.1. Proposed Method

Prediction of sales while working is simple, which means that the industry faces many problems without predictive data. Success can be made if the predictions are accurate. This paper focuses on data visualization and exploration as well as performing preprocessing operations. Preprocessing steps, such as data exploration, data, and feature engineering, have been carried out, as well as and find that approach. Then the data mining algorithm [13] is applied for accurate predictive analysis. Finally, all algorithms are compared and served with RMSE values and find the most suitable algorithm for predictive analysis for the big mart sales.

2.2. Dataset Interpretation

The dataset used is Big Mart sales data. The dataset consists of 12 attributes. The attributes are described in Table 1. The dataset has multiple data from a few towns. The ratio of train and test data is 80 and 20.

Table 1. Description of dataset

Attribute	Description
item identifier	Product Unique id
item weight	Product Weight
item fat content	Fat content of the product
item visibility	Total percentage of are allocated to this store
item type category	Product belong to which category
item MRP	Price of the product
outlet identifier	Store unique id
outlet Establishment year	Established year of the store
outlet size	Store size
outlet location type	Type of located cities
outlet type	Supermarket sort
item outlet sales	Product sales in particular area

2.3. Preprocessing

Preprocessing in this study used three stages: Exploration of data, data preprocessing and cleaning, and Feature Engineering. There are two ways of analyzing the data; they are Univariate Analysis and Bivariate Analysis. Univariate analysis is used for the exploration of data. The histograms are used to plot continuous variables to give out—similarly, the bar plots for finite variables.

In Outlet Size's plot, there are 4016 observations, and some values are blank or missing and check in the bivariate analysis. As per the analysis, it has missing values. Imputation of Weight with mean according to Identifier variable in the item. It results in zero missing value. It means the missing data have been imputed in the feature.

Preprocessing of data include the conversation of raw data into the articulate form. Raw data commonly consists of incomplete and error data. This step also includes finding the missing values, categorical values, splitting of data as train and test set, and at last of feature scaling [4].

In this process, the attribute, namely item weight, and outlet size are missing values. The missing values of item weight are filled with average values, and outlet size is filled with the mode values based on the particular outlet type [2].

Feature Engineering is a process of conversion of data which is cleaned into predictive models. This process is done for presenting the problems in a better understandable way. The noise of the data that is identified in the exploration phase is resolved. Here the new features are created to bring the model effectively. This form covert the data in a way in which the algorithm can be understood [3].

2.4. Building the model

10-fold cross-validation and 5-fold cross-validation is used in all the models build basedupon the above algorithms. Basically cross-validation givesan idea of how well a model generalizes to unseen data. This paper has concentrated in building the model of various data mining algorithms. They are Decision Tree, Random Forest, Naive Bayes, Random Forest with PSO, Naive Bayes with PSO.

The decision tree is the most commonly used method of data mining, which obtains attribute values as input and produces the decision in the form of boolean as output. The decision is one of the approaches of tree-like structure in which every path starts to form the root node to the leaf node that represents data sequence by splitting up the data till the result is reaches boolean form. Each node in the decision tree has a rule that is decoded to programming and human languages. The complete decision tree is an expression of boolean that includes union and intersection to change the boolean decision. Tree structure theory does not consist of any restriction like braches count of tree node and the different values the leaf node consists. The decision tree result is a decision in the boolean form that has the power of splitting an attribute and information gain, which results in a reduction in entropy [14].

A decision tree is a tool used for classification with maximum accuracy for the dataset, which varies from nominal and numerical and small to medium. It also has good performance for small to medium size datasets used for learning the model. A decision tree is one of the commonly used data mining models to predict and classify systems [15].

Random forest is one of the algorithms used to predict a single model of decision from the multiple decision trees. The algorithm utilizes the bagging concept for the creation of a forest with decision trees. As it produces a result from multiple decision trees, it gives the most accurate prediction of output [3]. Random is also a tree-based algorithm with a bootstrapping method in which few counts of weak learners are accomplished for producing more accurate prediction models. A set of rows and a few set of variables are utilized to develop a model by the learners. Finally, the prediction is the task for every prediction model developed by the individual learners. For the regression model, the prediction is made based on the mean value of all predictions [16].

Random forest is one of the algorithms used in the prediction of the stock market [17]. So it is referred to as one of the easy algorithms for flexible use in data mining algorithms. It also provides the most accurate value for prediction. This algorithm is most commonly utilized in the classification functions [4]. The random forest workflow is to construct several decision trees in the training course and develop a group of classes that include classifications. This process is referred to as class and regression mode or as tree mean prediction, constructed individually [9].

Naive Bayes is one of the methods commonly used for text classification. However, it produces the most accurate result for only vast samples of the training set. The huge samples carry dense work for classification done annually and give a request for high storage and resource of computing is high for the process done later in computer [18].

When Naive Bayes is applied to a vast dataset, it produces more accurate and adequate, but it produces poor results for a few samples. However, in reality, it is not easy to work with a vast dataset. Though a large dataset produces accurate results, it shows high running time and occupies more space. Naive Bayes' performance can be similar to a neural network and decision tree learning [18].

Naive Bayes is a set of classification rules that is related to the Bayes theorem. Naive give high accuracy result, especially for data analysis of text like natural language processing. Naive Bayes is also used for the classifier, which has probabilistic characteristics. For performing classification, it utilizes the concept of mixture models [19].

Particle swarm optimization is one of the metaheuristic algorithms used to find the possible space of training subsets of the dataset. The feature selection of the subset is made by the inductive algorithm, namely random forest. Random forest is used as a function for evaluation. This inductive algorithm is trained using cross-validation and then tested. The next stage is to train a model with a subset of features based on the selection from PSO [20] and test it [21].

Naive Bayes is one of the classifier families that use basic probabilistic and Bayes theorem with the expectation of better individuality between the features. This classifier has maximum scalable value by demanding a small count of linear parameters and the number of predictions in the problem of learning [22]. For the training phase, this classifier is more simple and fast for probabilistic classification. Naive Bayes is combining with particle swarm optimization (PSO) for the optimal solution. PSO has the advantage of easy, simple to implement, and always produce an optimization model continuously. So it is used for the selected subset for the feature model [23]. The proposed Algorithm for Naïve bayes with PSO shown in the steps below.

Step 1: Initialize the data set.Step 2: For feature selection, PSO is applied.
Step 3: Features with minimum PSO values are omitted.
Step 4: Naive Bayes classifier is applied for significant features.
Step 5: Performance of Naive bayes+ PSO is calculated.

2.5. Evaluation metrics

The process of model building is not complete without an evaluation of the model's performance. So an evaluation metric is needed to evaluate the model. The RMSE as an evaluation metric is used. The regression problem is evaluated by (1).

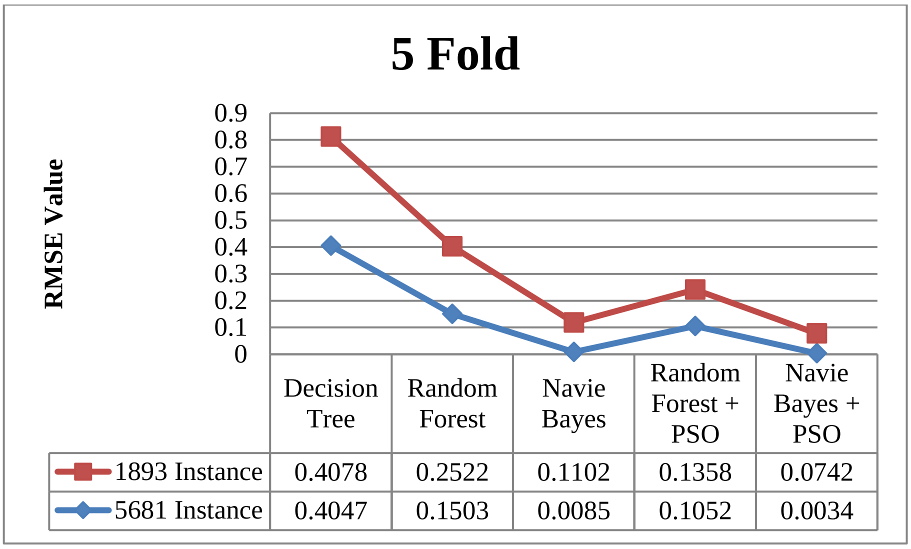
$$\sqrt{\frac{1}{m}\sum_{k=1}^{m}(\mathrm{z_k}-\hat{\mathrm{z}}_\mathrm{k})^2} \tag{1}$$

3. Results and Discussion

Table 2. Root Mean Squared Error (RMSE) values for 5 algorithms

Algorithm	5 Fold		10 Fold	
	5681Instances	*1893 instances*	*5681instances*	*1893 instances*
Decision tree	0.4047	0.4078	0.3026	0.3081
Random Forest	0.1503	0.2522	0.1732	0.2530
Naïve Bayes	0.0085	0.1102	0.0028	0.0732
Random forest +PSO	0.1052	0.1358	0.1077	0.1432
Naïve Bayes + PSO	0.0034	0.0742	0.0017	0.0062

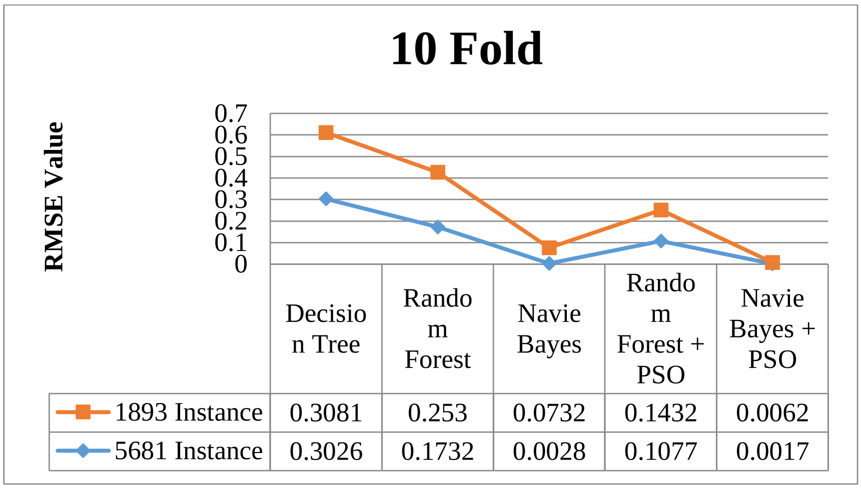
From Table 2, we can draw two different perceptions concerning the experimental results. In this study, the wholesome experiment is carried out in two phases. The first phase of the experiment includes training the classifiers in 5 folds and then studying the classifiers' performance with half of the remaining test cases initially and is then analyzed with the rest of the test cases on the whole. This strategy intends to find the classifier's performance in two different situations, such as having more training and lesser test case and lesser training and more test case and finally how the same classifier behaves. At the same time, it was more trained and especially on increasing test cases.



	Decision Tree	Random Forest	Navie Bayes	Random Forest + PSO	Navie Bayes + PSO
1893 Instance	0.4078	0.2522	0.1102	0.1358	0.0742
5681 Instance	0.4047	0.1503	0.0085	0.1052	0.0034

Fig. 1. RMSE for various technique for 5 Fold

As a result suggested (Fig. 1), initially, Naïve Bayes outclassed the others with lesser trained samples, and the stand is the same. Even though RF provides better RMSE after PSO optimized feature selection, it cannot outperform the PSO-based Bayesian strategy. Whereas in the intensive training phase of the 10-fold model (Fig. 2), every classifier's corresponding performance seems better. Henceforth, intensive training of any classifier will significantly boost its performance over poorly trained same classifier. Another important perception is that, by appropriately optimizing the feature selection, the performance can be further fine-tuned.



	Decisio n Tree	Rando m Forest	Navie Bayes	Rando m Forest + PSO	Navie Bayes + PSO
1893 Instance	0.3081	0.253	0.0732	0.1432	0.0062
5681 Instance	0.3026	0.1732	0.0028	0.1077	0.0017

Fig. 2. RMSE for various technique for 10 Fold.

4. Conclusion

In this paper, sales in the future are predicted from the data mining techniques. This prediction is utilized for maximizing the profit and for finding the strategy in the big mart. The algorithms compared in this paper are Decision Tree, Naive Bayes, Adaboost with Particle Swarm Optimization (PSO), Random Forest, and Naive Bayes with Particle Swarm optimization PSO for the sales prediction. The dataset is preprocessed with the removal of noise, filling missing values. Feature engineering is performed to convert the data understandably. Then the data mining algorithms are applied. Each algorithm is compared concerning root mean squared error value. After analyzing, Naïve Bayes with particle swarm optimization algorithm has the lowest RMSE value. The algorithm which has the lowest RMSE value is considered the best algorithm. By the result of the algorithm, Naive Bayes with PSO is concluded as the best model. With the help of this prediction, big mart sale is increased by improving their strategy.

References

[1] Y. Meier, J. Xu, O. Atan, and M. Van Der Schaar, "Personalized grade prediction: a data mining approach," in *2015 IEEE International Conference on Data Mining*, Nov. 2015, pp. 907–912, doi: 10.1109/ICDM.2015.54.

[2] K. Punam, R. Pamula, and P. K. Jain, "A two-level statistical model for big mart sales prediction," in *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, Sep. 2018, pp. 617–620, doi: 10.1109/GUCON.2018.8675060.

[3] M. N, P. Chatradi, A. C. V, S. M. Kalavala, and N. K. S, "Improvizing big market sales prediction," *J. Xi'an Univ. Archit. Technol.*, vol. XII, no. IV, pp. 4307–4313, 2020, doi: 10.37896/JXAT12.04/1172.

[4] K. H. Sadia, A. Sharma, A. Paul, Sarmistha Padhi, and S. Sanyal, "Stock market prediction using machine learning algorithms," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 4, pp. 25–31, 2019, [Online]. Available: https://www.ijeat.org/wp-content/uploads/papers/v8i4/D6321048419.pdf.

[5] K. Chen, Y. Li, and X. Xu, "Rotating target classification base on micro-Doppler features using a modified adaptive boosting algorithm," in *2015 International Conference on Computers, Communications, and Systems (ICCCS)*, Nov. 2015, pp. 236–240, doi: 10.1109/CCOMS.2015.7562907.

[6] B. X. Chen, R. Sahdev, and J. K. Tsotsos, "Person following robot using selected online ada-boosting with stereo camera," in *2017 14th Conference on Computer and Robot Vision (CRV)*, May 2017, pp. 48–55, doi: 10.1109/CRV.2017.55.

[7] N. N. Sakhare and S. S. Imambi, "Performance analysis of regression based machine learning techniques for prediction of stock market movement," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6S4, pp. 206–213, 2019, [Online]. Available: https://www.ijedr.org/papers/IJEDR1804010.pdf .

[8] A. Chandel, A. Dubey, S. Dhawale, and M. Ghuge, "Sales prediction system using machine learning," *Int. J. Sci. Res. Eng. Dev.*, vol. 2, no. 2, pp. 667–670, 2019, [Online]. Available: http://www.ijsred.com/volume2/issue2/IJSRED-V2I2P83.pdf.

[9] H. Kadam, R. Shevade, D. Ketkar, and S. Rajguru, "A forecast for big mart sales based on random forests and multiple linear regression," *Int. J. Eng. Dev. Res.*, vol. 6, no. 4, pp. 41–42, 2018, [Online]. Available: https://www.ijedr.org/papers/IJEDR1804010.pdf.

[10] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and As. Perera, "Opinion mining and sentiment analysis on a Twitter data stream," in *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*, Dec. 2012, pp. 182–188, doi: 10.1109/ICTer.2012.6423033.

[11] T. Jain, A. . Dua, and V. Sharma, "Quantitative analysis of apriori and eclat algorithm for association rule mining," *Int. J. Eng. Comput. Sci.*, vol. 4, no. 10, pp. 14649–14652, 2015, [Online]. Available: http://103.53.42.157/index.php/ijecs/article/view/2973/2752.

[12] P. P. Shinde, K. S. Oza, and R. K. Kamat, "Big data predictive analysis: Using R analytical tool," in *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Feb. 2017, pp. 839–842, doi: 10.1109/I-SMAC.2017.8058297.

[13] P.-Y. Zhou, K. C. C. Chan, and C. X. Ou, "Corporate communication network and stock price movements: insights from data mining," *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 2, pp. 391–402, Jun. 2018, doi: 10.1109/TCSS.2018.2812703.

[14] F.-J. Yang, "An extended idea about decision trees," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, Dec. 2019, pp. 349–354, doi: 10.1109/CSCI49370.2019.00068.

[15] S. Patil and U. Kulkarni, "Accuracy prediction for distributed decision tree using machine learning approach," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Apr. 2019, pp. 1365–1371, doi: 10.1109/ICOEI.2019.8862580.

[16] A. Narkhede, M. Awari, S. Gawali, and A. Mhaisgawali, "Big mart sales prediction using machine

learning techniques," *Int. J. Sci. Res. Eng. Dev.*, vol. 3, no. 4, pp. 693–697, 2020. Available: Google Scholar.

[17] D. Shah, H. Isah, and F. Zulkernine, "Stock market analysis: a review and taxonomy of prediction techniques," *Int. J. Financ. Stud.*, vol. 7, no. 2, p. 26, May 2019, doi: 10.3390/ijfs7020026.

[18] Y. Huang and L. Li, "Naive bayes classification algorithm based on small sample set," in *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*, Sep. 2011, pp. 34–39, doi: 10.1109/CCIS.2011.6045027.

[19] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," in *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, Nov. 2019, pp. 266–270, doi: 10.1109/SMART46866.2019.9117512.

[20] H. Pan, Y. Zhu, and L. Z. Xia, "Hierarchical PSO-adaboost based classifiers for fast and robust face detection," *Int. J. Information, Control Comput. Sci.*, vol. 4, no. 11, 2011, doi: 10.5281/zenodo.1327696.

[21] H. Faris, I. Aljarah, and B. Al-Shboul, "A hybrid approach based on particle swarm optimization and random forests for e-mail spam filtering," in *Nguyen NT., Iliadis L., Manolopoulos Y., Trawiński B. (eds) Computational Collective Intelligence. ICCCI 2016. Lecture Notes in Computer Science*, Cham: Springer, 2016, pp. 498–508, doi: 10.1007/978-3-319-45243-2_46.

[22] A. Tripathi, S. Yadav, and R. Rajan, "Naive Bayes Classification Model for the Student Performance Prediction," in *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, Jul. 2019, pp. 1548–1553, doi: 10.1109/ICICICT46008.2019.8993237.

[23] U. N. Dulhare, "Prediction system for heart disease using Naive Bayes and particle swarm optimization," *Biomed. Res.*, vol. 29, no. 12, 2018, doi: 10.4066/biomedicalresearch.29-18-620.