Impact Factor (SJIF): 6.092 E-ISSN: 2583-1615

International Journal of Research in Academic World

Received: 08/March/2023 IJRAW: 2023; 2(4):123-127 Accepted: 04/April/2023

Developing a Model for Accurate Prediction of Soil Fertility: A Comparative Study of Machine Learning Techniques

[*1]Asha S and [2]Sony PS

[*1]Assistant Professor SCMS School of Engineering and Technology, Karukutty, Ernakulam Kerala, India.

[2]MCA Student, SCMS School of Engineering and Technology, Karukutty, Ernakulam Kerala, India.

Abstract

The prediction of soil fertility is a crucial aspect of agriculture as it helps farmers to make informed decisions about crop selection, fertilization, and irrigation. In this study, we propose use of three different machine learning techniques to predict the fertility of soil. Soil properties such as pH, nitrogen levels, and texture will be analyzed and used as input features for the models. A dataset of labeled soil samples will be collected and used to train the models. The three methods used in this study are: Random Forest, Naïve Bayes, and Support Vector Machine. The performance of the models will be evaluated using metrics such as accuracy, precision, and recall. The goal of this study is to compare the performance of these three machine learning methods and to identify the most efficient method for predicting soil fertility. This method will not only aid in agricultural decision-making but also help in improving crop yields. The results of this study will be useful for farmers, agronomists, and researchers in the field of agriculture. Additionally, this study will also explore the correlation between soil properties and fertility, which will help to understand the mechanism of soil fertility.

Keywords: Soil fertility, machine learning, support vector machines

1. Introduction

1.1. Background

Machine learning (ML) is a branch of Artificial Intelligence (AI) that enables the improvement of prediction accuracy without the need for hand-crafted software. Utilizing historical data as inputs, ML algorithms make predictions for new outputs. Its widespread adoption in various fields, including automation, is attributed to the significant advancements in data access and processing capabilities. As a result, ML has become one of the most widely adopted AI technologies among firms, institutions, and individuals seeking to obtain meaningful outcomes. The prediction of whether soil is fertile or non-fertile using machine learning (ML) involves the development of models that can predict soil fertility based on various soil and environmental factors.

Soil fertility prediction using ML has the potential to significantly improve agricultural productivity by providing more accurate and timely information about soil fertility. This can help farmers make better decisions about fertilization, crop selection, and soil management, leading to improved yields and reduced environmental impact.

1.2. Objective

The objective of soil fertility prediction using three different machine learning techniques such as Random Forest, Naive Bayes, and Support Vector Machines (SVM) is to compare the accuracy and effectiveness of each method for predicting soil fertility. The goal is to determine which technique provides the most accurate predictions and to identify the strengths and weaknesses of each method. By comparing the performance of these three ML techniques, the aim is to gain a better understanding of the best approaches for soil fertility prediction and to identify the most promising techniques for further development and implementation. The objective is to improve the accuracy of soil fertility prediction and to provide decision support for agricultural management practices, leading to improved yields and reduced environmental impact.

1.3. Scope

The scope of the study on soil fertility prediction using three machine learning techniques such as Random Forest, Naive Bayes, and Support Vector Machines (SVM) encompasses various aspects related to data collection, model training, model evaluation, and comparison of models. The study includes collecting and preparing a dataset of soil and environmental factors as inputs for the ML algorithms. The trained models, Random Forest, Naive Bayes, and SVM, are evaluated for their accuracy and effectiveness using appropriate evaluation metrics. The comparison of the performance of each model helps to determine the most accurate and effective technique for soil fertility prediction.

***Corresponding Author:** Asha S

The overall aim of the study is to improve the accuracy of soil fertility predictions and provide decision support for agricultural management practices.

1.4. Organization of Report

The report is structured into five chapters. The introductory chapter provides background information on the project, its objectives, and the scope of the study. The second chapter reviews relevant literature related to the project. The third chapter describes the design and functioning of the proposed system. The fourth chapter presents the results of experiments and provides a discussion of the findings. The final chapter summarizes the conclusions, highlights the future scope of the project, and lists references used in the report.

2. Literature Survey

The paper "Random Forest Algorithm for Soil Fertility Prediction and Grading Using Machine Learning" [1] was published in the International Journal of Innovative Technology and Exploratory Engineering in 2019 by Keerthan Kumar T.G, C. Shubha, and S. A. Sushma. The study explores the use of the Random Forest algorithm to predict soil fertility and grade soil based on various soil attributes. The authors compare the performance of the Random Forest algorithm with other machine learning methods, such as Support Vector Machines and Artificial Neural Networks, and find that the Random Forest algorithm produces more accurate results. This research highlights the potential of machine learning techniques in predicting soil fertility and grading, which can assist farmers in making informed decisions about fertilization and crop management.

The paper "Soil Analysis and Crop Fertility Prediction Using Machine Learning"[2] was published in the journal Machine Learning in 2021. It was written by Jagdeep Yadav, Shalu Chopra, and M. Vijayalakshmi. The authors explored the use of machine learning algorithms to analyze soil data and predict crop fertility. The results showed that machine learning can be an effective tool for soil analysis and crop fertility prediction, offering potential for improved agricultural productivity and food security.

"Soil data analysis using classification techniques and soil attribute prediction" [3] is a scientific paper written by Jay Gholap and others. The paper was published in 2012 on arXiv, a platform for pre-print academic articles in computer science and other fields. The paper likely explores the use of classification techniques in soil data analysis and focuses on predicting soil attributes using machine learning methods. The results of this research could be useful for agriculture and land management, as it could help in identifying soil types and predicting soil properties such as fertility, water-holding capacity, and others.

The paper "A Model for Prediction of Crop Yield" [4] was published in the International Journal of Computational Intelligence and Informatics in March 2017. It was authored by E. Manjula and S. Djodiltachoumy. The authors proposed a model that uses computational intelligence and informatics to predict crop yield. The model takes into account various factors such as weather, soil, and previous crop data to provide accurate yield predictions. The results showed that the proposed model has high accuracy in predicting crop yields, offering potential benefits for farmers, agronomists, and policy makers in improving agricultural productivity and food security.

3. Proposed System

Soil fertility prediction involves determining if soil is fertile or non-fertile based on various characteristics such as soil type, pH, organic matter content, and nutrient levels [5]. This prediction is made through a process that includes feature selection, training, testing, and prediction using machine learning techniques such as Naive Bayes, Random Forest, and Support Vector Machine. The performance of these models is compared, and the one that yields the highest accuracy is chosen for making predictions on new data.

3.1. Architecture

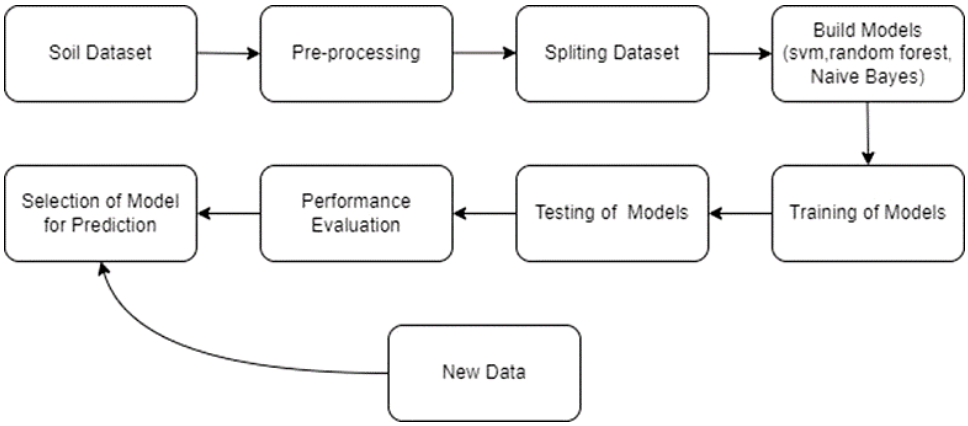

Fig 1: Architecture of proposed system

The figure above illustrates the overall design of the system. The first step involves gathering data, which can be obtained through the farmers' portal and organized by state. A sample dataset is utilized in this model. The labels were either "fertile" or "non-fertile."

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	pH	EC	OC	OM	N	P	K	Zn	Fe	Cu	Mn	Sand	Silt	Clay	CaCO3	CEC	Output
2	7.74	0.4	0.01	0.01	75	20	279	0.48	6.4	0.21	4.7	84.3	6.8	8.9	6.72	7.81	Fertile
3	9.02	0.31	0.02	0.03	85	15.7	247	0.27	6.4	0.16	5.6	90.4	3.9	5.7	4.61	7.19	Fertile
4	7.8	0.17	0.02	0.03	77	35.6	265	0.46	6.2	0.51	6.1	84.5	6.9	8.6	1.53	12.32	Fertile
5	8.36	0.02	0.03	0.05	106	6.4	127	0.5	3.1	0.28	2.3	93.9	1.7	4.4	0	1.6	Non Fertile
6	8.36	1.08	0.03	0.05	96	10.5	96	0.31	3.2	0.23	4.1	91.5	4.1	4.4	9.08	7.21	Non Fertile
7	8.36	0.73	0.03	0.05	151	10.5	230	0.38	2.5	0.37	4.2	94.2	1.5	4.3	6.23	3.34	Non Fertile
8	7.69	0.11	0.04	0.06	112	8	120	0.51	3.1	0.32	1.2	96.2	1.7	2.1	0	1.72	Non Fertile
9	8.39	0.06	0.04	0.06	125	18.5	145	0.67	2.8	0.18	1.8	87.9	4.8	7.3	0	7.34	Non Fertile
10	7.87	0.43	0.04	0.06	112	27	333	0.75	3.9	0.54	1.8	80.5	5.7	13.8	3.21	11.02	Fertile
11	8.09	0.62	0.04	0.06	89	28.2	146	0.53	3.8	0.71	3.9	80.2	9.5	10.3	8.51	8.94	Fertile
12	8.26	0.11	0.04	0.06	114	12	276	0.89	6.1	0.45	4.8	91.2	4	4.8	0	2.8	Non Fertile
13	7.9	0.91	0.04	0.06	80	29	162	0.7	5.9	0.32	5.2	80.7	5.6	13.7	9.73	6.93	Fertile
14	8.12	0.14	0.04	0.06	78	8.4	190	0.62	1	0.07	5.5	89.5	3.5	7	6.1	4.89	Fertile
15	8.64	0.18	0.05	0.08	88	13.7	164	0.5	6.1	0.48	4.5	90.3	2.9	6.8	2.34	4.59	Fertile
16	8.38	0.1	0.06	0.1	125	7.1	198	0.48	4.2	0.15	0.8	88.8	4.9	6.3	4.8	2.4	Non Fertile
17	8.37	0.12	0.06	0.1	120	3.2	288	0.41	3.3	0.35	3.1	86.9	7.1	6	0	5.6	Non Fertile
18	8.52	0.1	0.06	0.1	128	7.9	260	0.52	4	0.22	4.2	91.1	4.1	4.8	15.12	2.6	Non Fertile
19	8.67	0.23	0.07	0.12	134	10.4	410	0.34	2.5	0.11	0.4	88.6	4.2	7.2	0	2.82	Non Fertile
20	8.4	0.13	0.07	0.12	130	12.9	249	0.47	3.2	0.17	2.8	88.3	4.6	7.1	14.98	3.2	Non Fertile
21	7.53	0.15	0.07	0.12	136	12	192	0.34	2.8	0.35	3.5	94.5	2.3	3.2	0	4.91	Non Fertile
22	8.37	0.48	0.07	0.12	137	3.7	340	0.37	3.6	0.32	4.1	89.8	4.1	6.1	0	1.23	Non Fertile
23	7.87	0.11	0.07	0.12	109	26.8	149	0.04	1.8	0.01	5.3	79.8	10.7	9.5	3.31	5.37	Fertile
24	8.65	0.48	0.07	0.12	112	10	155	0.12	6.1	0.5	5.3	83.5	7.8	8.7	1.28	6.67	Fertile
25	7.8	0.53	0.07	0.12	105	7	264	0.46	6.1	0.46	6.7	88.4	3.7	7.9	7.3	3.74	Fertile

Fig 2: Dataset

Figure 2 depicts the dataset in CSV format, including the properties of the soil as features. The data set consists of sixteen unique characteristics and a table providing explanations for each feature is given below.

Table 1: Description of Dataset Attributes

Attribute	Description
PH	Soil ph Value
EC	Electronic Conductivity
OC	Organic Carbon
OM	Organic Matter
N	Nitrogen Content
P	Phosphorous Content
K	Potassium Content
Zn Fe Cu Mn Sand Slit	Zinc Content Iron Content Copper Content
Clay	Manganese Content Soil Composition
CaCo3	Sodium Bi-carbonate Content
CEC	Cati on Exchange Capacity

3.1.1. Data Preprocessing

The input features in the dataset have different value ranges, which can affect the performance of some machine learning algorithms. To handle this, a preprocessing tool called Min Max Scaler [7] from the Sklearn library can be used to scale all the attributes to the same range of [0,1]. This means that all the values for each feature will be transformed to a value between 0 and 1, where 0 represents the minimum value and 1 represents the maximum value.

The label column of the dataset represents the output and is usually a categorical variable. To use it in machine learning algorithms, it must be converted to numerical values. This can be done using the label Encoder [8] from the Sklearn library, which will convert the categorical values to numerical values, such as 0 and 1.

3.1.2. Data Spliting

The purpose of data splitting is to divide the available data into two parts: a training set and a testing set. The split ratio is that used in this system is 80:20, where 80% of the data is used for training and 20% is used for testing [9]. Data splitting helps to prevent overfitting and ensure the model's generalization ability, making it a fundamental step in the machine learning process.

3.1.3. Build Models

To predicting whether soil is fertile or non-fertile, three different techniques are used: Random Forest, Naive Bayes, and Support Vector Machines (SVM).

Random Forest is a popular machine learning algorithm that can be used for the prediction of whether the soil is fertile or non-fertile [10]. Random Forest builds multiple decision trees and combines their predictions to produce a final result. This technique is known as ensemble learning and can lead to improved accuracy and robustness compared to a single decision tree. In soil fertility prediction, Random Forest can handle complex relationships between multiple environmental factors and the soil fertility target. Random Forest can also handle noisy data and missing values, which are common in soil fertility datasets. Additionally, Random Forest can be used for feature selection, which can help to identify the most important factors for soil fertility prediction. In conclusion, using Random Forest for soil fertility prediction can lead to more accurate and reliable results.

Support Vector Machines (SVM) is a powerful machine learning algorithm that can be used for the prediction of whether the soil is fertile or non-fertile [11]. SVM is well-suited for binary classification problems, making it an ideal choice for soil fertility prediction. By finding the best boundary between the fertile and non-fertile classes, SVM can produce accurate predictions. SVM can handle noisy data, which is common in soil fertility prediction where soil samples may contain inaccuracies or outliers.

Naive Bayes is a simple but effective machine learning algorithm that can be used for the prediction of whether the soil is fertile or non-fertile [12]. Naive Bayes is based on Bayes' theorem and assumes that the input features are independent of each other. Naive Bayes can handle continuous and categorical features, which is often the case in soil fertility prediction where multiple environmental factors need to be considered.

3.1.4. Testing and Training

Training [13] and testing [13, 14] are both important for soil fertility prediction as they ensure the accuracy and reliability of the model. Training allows the model to learn from the available data, building a mapping from inputs to outputs. However, it is important to assess the model's performance on new, unseen data, which is where testing comes in. Testing allows the model to be evaluated on a different dataset than the one it was trained on, giving an indication of its ability to generalize to new data.

By performing both training and testing, it is possible to validate the model's performance and to detect any overfitting or underfitting. Overfitting occurs when a model fits the training data too well and performs poorly on new data, while underfitting occurs when the model is too simple to fit the complexity of the data. Testing helps to identify these issues and to adjust the model accordingly.

4. Results and Discussion

The soil fertility prediction project successfully created multiple models that evaluate soil fertility based on a range of soil characteristics. The models were constructed using three different machine learning algorithms. Once developed, the models were trained on a comprehensive dataset of soil samples, which had been categorized as either fertile or non-fertile. This enabled the models to make predictions about the fertility of a soil sample based on its attributes when provided with new input data. The models' performance was evaluated using accuracy scores, and the most accurate model was then selected for making predictions on previously unseen data.

	pH	EC	OC	OM	N	P	K	Zn	Fe	Cu	Mn	Sand
0	7.74	0.40	0.01	0.01	75	20.00	279	0.48	6.40	0.21	4.70	84.30
1	9.02	0.31	0.02	0.03	85	15.70	247	0.27	6.40	0.16	5.60	90.40
2	7.80	0.17	0.02	0.03	77	35.60	265	0.46	6.20	0.51	6.10	84.50
3	8.36	0.02	0.03	0.05	106	6.40	127	0.50	3.10	0.28	2.30	93.90
4	8.36	1.08	0.03	0.05	96	10.50	96	0.31	3.20	0.23	4.10	91.50
..	...	...	...	...	...	...	...	...	...	...	...	...
105	6.63	0.14	0.08	0.13	74	10.50	326	0.38	3.47	0.03	6.11	90.08
106	7.92	0.41	0.27	0.46	180	4.53	148	0.51	2.42	0.15	0.52	95.25
107	6.34	2.72	0.14	0.24	154	18.97	176	0.76	4.19	0.67	0.80	97.82
108	8.35	0.56	0.29	0.50	248	33.00	134	0.27	6.48	0.08	0.27	96.01
109	8.78	0.33	0.11	0.19	227	12.70	103	0.18	2.73	0.11	0.40	92.66

	Silt	Clay	CaCO3	CEC	Output
0	6.80	8.90	6.72	7.81	Fertile
1	3.90	5.70	4.61	7.19	Fertile
2	6.90	8.60	1.53	12.32	Fertile
3	1.70	4.40	0.00	1.60	Non Fertile
4	4.10	4.40	9.08	7.21	Non Fertile
..	...	...	...	...	...
105	7.29	2.63	1.02	3.06	Fertile
106	1.51	3.24	2.50	4.11	Non Fertile
107	1.00	1.18	3.34	2.27	Non Fertile
108	2.88	1.11	1.88	3.82	Non Fertile
109	1.67	5.67	3.03	2.73	Non Fertile

Fig 3: Loaded dataset

The dataset was visualized using the seaborn library to determine the distribution of fertile and non-fertile soil samples. Seaborn is a popular Python tool for data visualization and graphical representation of statistical information.

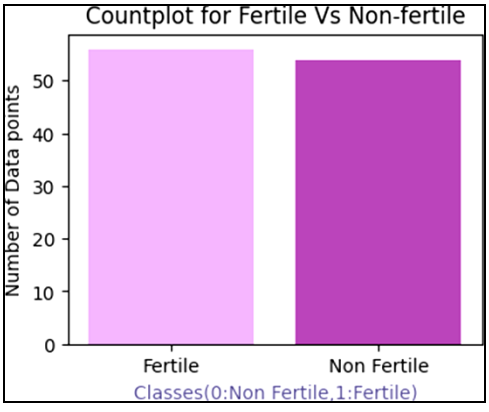

Fig 4: Statistical distribution fertile & non-fertile soil

The following table displays the accuracy score of each model.

Table 2: Accuracy Score of different classifiers

Classifier	Accuracy Score
Random Forest	0.954545
SVM	0.909091
Naïve Bayes	0.863636

The table shows that Random Forest outperforms SVM and Naive Bayes in soil fertility prediction accuracy. Thus, Random Forest is chosen to construct the interface for predicting unseen data.

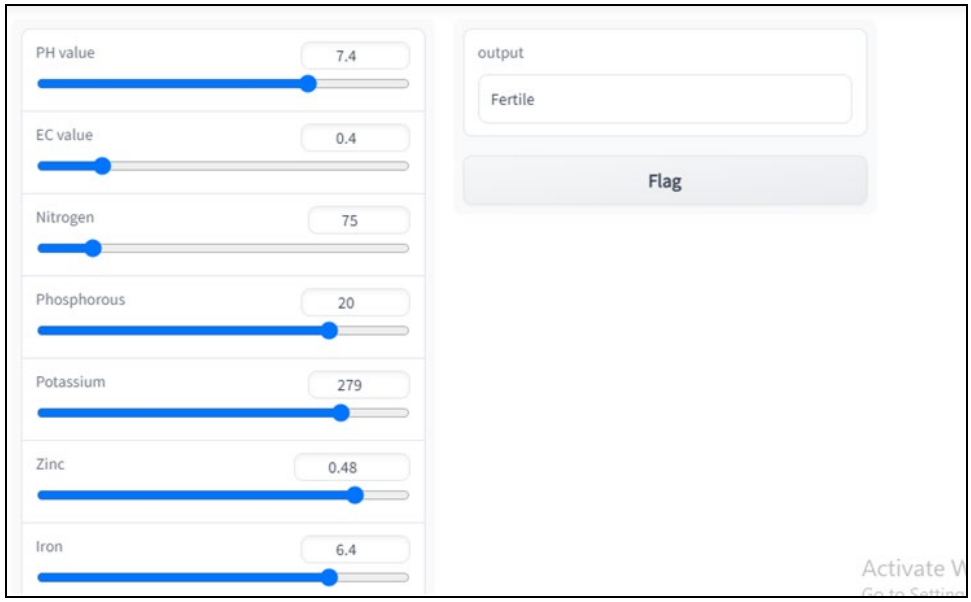

Fig 5: User interface for soil fertility prediction

5. Conclusion & Future Scope

In conclusion, the proposed system for soil fertility prediction is a promising solution that can provide accurate and efficient predictions about soil fertility. The use of multiple machine learning models, such as SVM, Random Forest, and Naive Bayes, increases the accuracy of predictions and provides a comprehensive understanding of soil fertility. The system can be integrated with other systems, such as precision agriculture systems, to provide a more comprehensive view of soil fertility and help farmers and agronomists make informed decisions about soil management practices.

The future scope of the proposed system is vast. The models used in the system can be further improved by incorporating deep learning techniques, such as neural networks, to increase their accuracy. The system can also be expanded to cover a larger area and provide a more accurate picture of soil fertility patterns across regions. Additionally, the system can be extended to include more soil properties, such as soil moisture, to provide a more comprehensive analysis of soil fertility.

Moreover, the user-interface of the proposed system can be made more user-friendly and intuitive to make it easier for farmers and agronomists to use. The system can also be made more accessible to a larger audience by developing a mobile application that can be used on smartphones and tablets. The proposed system can also be integrated with other systems, such as remote sensing and drone technologies, to provide a more complete picture of soil fertility and help farmers and agronomists make better-informed decisions about soil management practices.

In summary, the proposed system for soil fertility prediction has the potential to revolutionize the way farmers and agronomists assess soil fertility and make decisions about soil management practices. With the vast scope for improvement and integration with other systems, the proposed system has the potential to make a significant impact on agriculture and help increase crop yields and food security.

References

1. Keerthan Kumar TG, Shubha C, Sushma SA. "Random forest algorithm for soil fertility prediction and grading using machine learning." *Int J Innov Technol Explor Eng.* 2019; 9(1):1301-1304.
2. Yadav, Jagdeep, Shalu Chopra, and M. Vijayalakshmi. "Soil analysis and crop fertility prediction using machine learning." Machine Learning, 2021, 8(3).
3. Gholap, Jay *et al*. "Soil data analysis using classification techniques and soil attribute prediction." arXiv preprint arXiv. 2012; 12(6):15-57.
4. Manjula E, Djodiltachoumy S, "A Model for Prediction of Crop Yield", *International Journal of Computational Intelligence and Informatics*, 2017.
5. Wankhede DS. "Analysis and Prediction of Soil Nutrients pH,N,P,K for Crop Using Machine Learning Classifier: A Review". In: Raj, J.S. International Conference on Mobile Computing and Sustainable Informatics. ICMCSI 2020. EAI/Springer Innovations in Communication and Computing. Cham, 2021. https://doi.org/10.1007/978-3-030-49795-8_10A.
6. Olisah, Chollette C. Lyndon Smith, and Melvyn Smith. "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective." Computer Methods and Programs in Biomedicine. 2022; 220:106-773.
7. Patro S Gopal and Kishore Kumar Sahu. "Normalization: A preprocessing stage." arXiv preprint arXiv:. 2015; 1503:06462.
8. Pradhan, Madhavi, and Bamnote GR. "Efficient binary classifier for prediction of diabetes using data preprocessing and support vector machine." Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014: Volume 1. Springer International Publishing, 2015.
9. Reitermanova, Zuzana. "Data splitting." WDS. Prague: Matfyzpress, 2010, 10.
10. Li, XinHai. "Using" random forest" for classification and regression." Chinese Journal of Applied Entomology 50.4, 2013, 2019, 1190-1197.Y. A. Bachtiar.,.
11. Gunn, Steve R. "Support vector machines for classification and regression." *ISIS technical report.* 1998; 14(1):5-16.
12. McCallum, Andrew, Kamal Nigam. "A comparison of event models for naive bayes text classification." *AAAI-98 workshop on learning for text categorization*, 1998, 752(1).
13. Witten, Ian H., *et al*. "Practical machine learning tools and techniques." Data Mining, 2005, 2(4).