Available in online @ www.iaraindia.com
SCIENCE EXPLORE-An International Journal on Science and Knowledge in General
ISSN: 3139-2997 (Online)
Volume I, Issue 1
January – June 2026

COMPARISON OF SUPPORT VECTOR MACHINE WITH NAIVE BAYES CLASSIFIER METHOD FOR CLASSIFICATION OF LECTURER PERFORMANCE

REMIANUS TUNTI
Department of Computer Science
Dili Institute of Technology, Dili, Timor-Leste

ARÃO TERNORIO ALVES SANTOS
Department of Computer Science
Dili Institute of Technology, Dili, Timor-Leste

JACOB SOARES
Department of Computer Science
Dili Institute of Technology, Dili, Timor-Leste

JAYASHREE R
Department of Computer Science
Dili Institute of Technology, Dili, Timor-Leste

ABSTRACT

One of the supervised learning techniques that extracts models from training data is classification, which identifies predefined categories for test data and provides high accuracy in predictions. The objective of this research paper is to compare two machine learning methods, such as the Support Vector Machine (SVM) and Naïve Bayes Classifier (NBC), for classifying lecturer performance. This comparative study utilizes datasets from three units, namely, data from research scorecard evaluations, teaching scorecards, and lecturer evaluations from students. Comparison results are employed in a confusion matrix table to evaluate metrics of the dataset. The results of comparing 150 Research Scorecard datasets by dividing into 70% and 30% for training and testing, respectively, using the NBC and SVM methods yielded identical accuracy, precision, and recall results, all at 100%. Meanwhile, the comparison of 150 lecturer evaluations from student datasets using the NBC method resulted in higher accuracy, precision, and recall, all at 100%, compared to using the SVM method, which achieved an accuracy: 93.18%, precision: 72.04%, and recall 75%.

KEYWORDS: classification, SVM, NBC, lecturer performance

INTRODUCTION

Supervised learning technique, classification, that extracts archetypes from training data to identify predefined categories for new test data (Morán-Fernández, 2021). It involves three phases: formation, validation, and testing (Tharwat, 2021). High accuracy is crucial for accurate predictions (Zhou, 2014), but performance bias poses a challenge (Soleymani, 2019), and imbalanced datasets are commonly encountered in real-life applications (Jedrzejowicz, 2021).

Several studies on the performance evaluation of classifiers in data mining such as analysis of teacher performance using multiple classifiers (Ahmad & Rashid, 2016: Kumar Pal & Pal, 2013), classifications on instructor performance (Agaoglu, 2016), to predict instructor performance (Ahmed, 2016), classify student-lecturer comments using sentiment analysis (Rakhmanov, 2020), on educational data perform sentiment analysis and opinion mining (Shaik, 2023) and experimental comparison of multilabel methods (García-Pedrajas, 2024).

The NBC and SVM are widely used classification techniques. The widespread choice for big data analysis is the NBC algorithm because of its efficient structure, whereas the Naive Bayes is a simple classifier based on probabilistic methods, which applies the Bayesian theorem with a strong assumption of independence (Chen et al, 2021; Perez, 2021). Bayes' theorem is crucial for inferential statistics and advanced Machine Learning (ML) models, and this theorem updates Hypothesis Probability (HP) based on new evidence (Berrar, 2018). Salmi and Rustam's (2019) research work stated that this model achieves higher classification accuracy with less complexity. Whereas SVMs are highly effective and reliable algorithms for regression and classification across various fields (Cervantes et al, 2020). As per Suthaharan's (2016) study, SVM is effective in tackling the challenges for large data categorization, especially in multidomain applications in large data environments. Additionally, the SVMs provide both linear and nonlinear decision boundaries for problems involving regression and classification (Somvanshi et al, 2016). Conventional SVMs identify a hyperplane in a higher-dimensional space that effectively divides various data classes (Amaya-Tejera et al, 2024).

In order to increase classification performance, numerous approaches are generally used, like a study conducted by Rahman (2018) on the application of feature selection with information gain (IG) for document classification. The selection of accurate and significant features in an attack detection system alerts on computer networks (Alhaj, 2016), improves classification performance in the credit scoring problem (Jadhav, 2018). Prasetiyo et al. (2021) studied the significance of feature selection in achieving classification accuracy using bank marketing datasets.

Thus, the objective of this research is to compare the SVM and NBC archetypes for classifying lecturer performance at the "Dili Institute of Technology" in Dili, Timor-Leste country. This comparative study utilizes datasets from three units, namely data from research scorecard evaluations, teaching scorecards, and lecturer evaluations from students. The comparison results employ a confusion matrix table to determine the evaluation matrix dataset.

Related Work

Classification is a supervised machine learning model with the objective of predicting categorical class labels for new instances based on initial observations (Sadiq et al., 2020). The classification process consists of two main phases: model development for training and model evaluation using testing data (Jalota & Agrawal, 2019b). In the principal three phases of the classification process, namely formation phase, validation phase, and testing phase, various steps are involved (Tharwat, 2021).

Several studies related to classification model include performance analysis of lecturers with Multiple Classifiers at Kurdistan-Iraq University (Ahmad & Rashid, 2016), evaluation of teaching quality with the Flipped Classroom model in colleges and universities as depicted in Figure 2.1 (Fu and Li, 2022), application of multiclassification models to evaluate teaching quality in the art department (Hua et al., 2022), and evaluation of English teaching quality using online analytical processing by combining classification algorithms at the college and university levels in China (Zhang et al., 2022).

Some commonly used methods in supervised machine learning include SVM and NBC. The SVM is an efficient ML technique that minimizes structural risk (Roy & Chakraborty, 2023). It serves as an algorithm for classification and regression tasks by offering advanced capabilities and parameter optimization, as stated by Cervantes et al. (2020). This method is extensively used in data mining because of its high proficiency, generalizability, and ability to find optimal solutions (Gaye et al., 2021). Ghosh et al. (2019) described the applications of SVM that include face detection, handwriting recognition, and other various real-world scenarios.

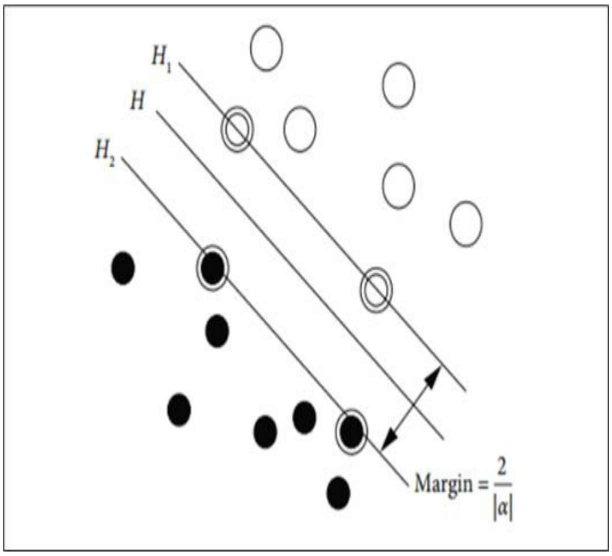


Figure 1
SVM Schematic Diagram (Fu and Li, 2022)

In the interim, the NBC method is also used in calculating the highest probability value as one of the classification processes (Atmadja et al., 2020). This NBC determines the outcomes based on the highest probability value in the classification process (Atmadja et al., 2020). Because of this simplicity and high accuracy, NBC serves as an effective classification archetype (Dangi et al, 2014). Although NBC is broadly recognized for its effectiveness in text classification, it achieves higher accuracy when trained on large sample datasets as described by Huang and Li (2011). Overall, one of the most prominent data mining algorithms for classification is Naive Bayes, and this method is the simplest and effective learning theory without the need of numerous parameters (Ramadhani et al, 2021).

One way to improve model accuracy is through feature selection techniques, and a very commonly used technique is information gain. The IG is an approach of feature evaluation which is popularly used in the field of ML (Lei, 2012). This technique is suitable for feature selection that reduces the size of a given feature by optimizing each attribute's value and providing a relative increase for that feature as analyzed by Zareapoor and Seeja (2015). Information is typically used in a variety of applications and is based on entropy metrics. The initial discussed beta can be used to determine relevance and reduce growth (Cherrington et al., 2019).

RESEARCH METHODOLOGY

The total dataset used in this research consists of 450 datasets, comprising 150 datasets of research scorecard with 69 attributes each, 150 datasets of teaching scorecard with 92 attributes each, and 150 datasets of lecturer evaluations from students with 35 attributes each. These datasets are obtained from the CARPS-CS, CEQA, and the Academic Department of DIT.

These attributes will be assigned their codes, standardized as per the standard format, and any missing values will be completed. The total attributes after these processes amount to 101 attributes ready to undergo the feature selection process, with 38 attributes from the Research Score Card, 37 from the Teaching Score Card, and 26 from the Student Evaluation.

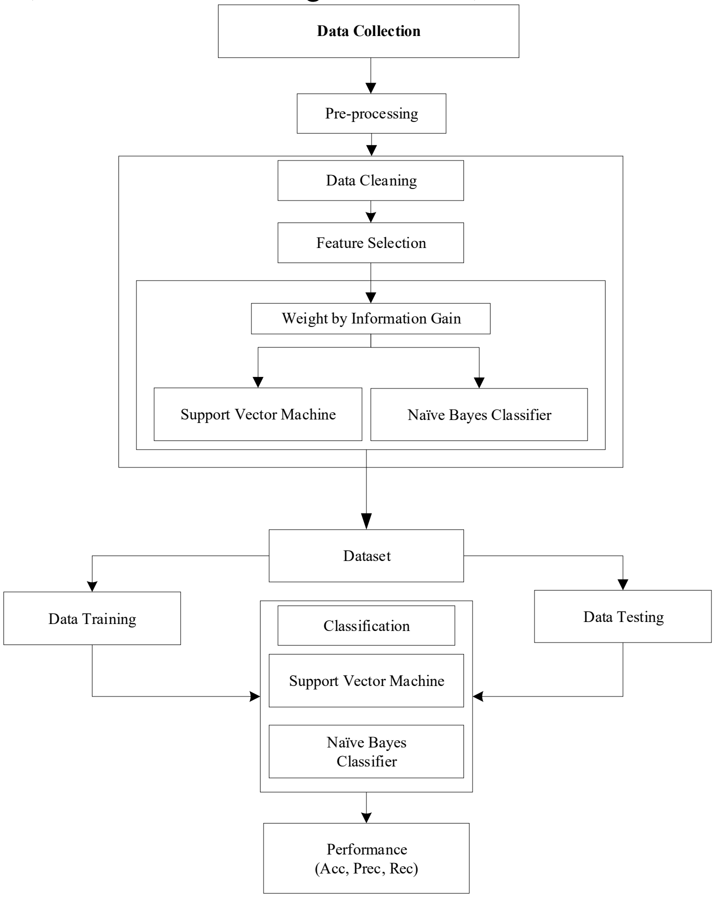


Figure 2
. Block Diagram of the Research

Figure 2. describes the stages of research to be carried out, starting with data collection, pre-processing, such as data cleaning and performing feature selection with information gain to determine the features to be used on each dataset. Then, in the next phase, divide the composition of the 70% and 30% of training and testing, respectively, to test the models built using the NBC and SVM methods. The final phase looks at the test results and performance (Accuracy, Precision, Recall) of each method using the table confusion matrix.

From the three datasets mentioned, it is necessary to conduct testing on feature selection using the IG method to determine the weight of each attribute. The use of this method is supported by research conducted by Ramanda Hasibuan (2019), where the results of the study are able to increase accuracy from 79.9% to 96.8%. The equation of the information gain method can be seen in the following equation 3.1 (Firmahsyah & Gantini, 2016):

$$\text{Entropy (S)} = \sum_{i}^{c} -\text{pi} \log_2 \text{pi} \tag{3.1}$$

Where c accumulation of values from classification classes, pi accumulation of samples from class i. After obtaining the entropy value, the calculation process of information gain can be done using the formula indicated in Equation 3.2:

$$\text{G (S, A)} = \text{Enty (S)} - \Sigma_{\text{VALUES(A)}} \frac{|\text{Sv}|}{|\text{S}|} \text{Enty } (\text{S}_\text{v}) \tag{3.2}$$

Where G (S, A) is the gain value of feature A, v possible value of feature A, values (A) possible values of set A, Sv number of examples of value v, S total number of data samples, Enty (Sv) is the entropy of value v example.

To eliminate attributes in the dataset, it is determined grounded on the weight of each attribute. If the attribute weight is zero, it will be eliminated because the attribute has no relationship and does not affect the dataset performance (Varghese & Sushmita, 2014). From the feature selection testing results on the three datasets, there are 18 attributes with zero weight, which are found in the Research Scorecard dataset. These attributes are: 21A11, 21A12, 21A21, 21A22, 21A23, 21B11, 21B21, 21B22, 22A12, 22A21, 33A11, 33A21, 33A22, 33A23, 443, 553, 554, and 555. Thus, the quantity of features used in the research scorecard dataset is 21 attributes, in the teaching scorecard dataset is 37 attributes, and in the evaluation of lecturers by students dataset is 26 attributes.

Data splitting, as a general approximation used for model validation, the dataset will split into 2 parts: training and testing data. This archetype is trained using the training data and validated (Joseph, 2022). The total estimation of datasets used for this research will be divided into 70% and 30% for training and testing data. From the total of 150 datasets for each unit, 104 will be allocated for data training and 46 will be allocated for data testing. This scaling utilization is also applied in research conducted by Abbi Nizar Muhammad et al. (2019), which combines the NBC method with SVM and demonstrates its superior accuracy level and strong performance.

In this section, we measure the performance of the NBC and SVM methods to predict the accuracy and relevancy for the Research Scorecard, Teaching Scorecard, and Student Evaluation dataset. In order to conduct this performance evaluation, we use the Confusion Matrix as a method for accurate calculation, based on the concept of data mining. The formula is to calculate evaluation metrics (Accuracy, Precision, and Recall) for various outputs. Shahi et al. (2018) conducted related research on this accuracy test to evaluate the accuracy of classifying Nepali news using the NBC, SVM, and Neural Networks methods. Additionally, research conducted by Ma et al., (2020) revealed the strength of Precision and Recall for the classification of spam emails using the NBC and SVC methods.

Confusion Matrix for Multiclass

		Predicted Number			
Actual Number		Class 1	Class 2	…	Class n
	Class 1	X_{11}	X_{12}	…	X_{1n}
	Class 2	X_{21}	X_{22}	…	X_{2n}
	. . .	. . .	. . .	. . .	. . .
	Class n	X_{n1}	X_{n2}	…	X_{nn}

RESULT

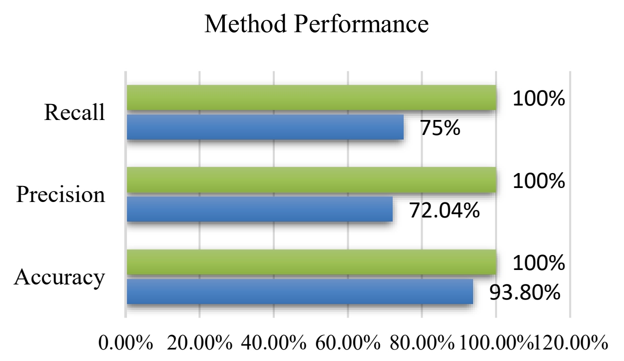
Performance testing is conducted on methods NBC and SVM using a dataset of 150 instances. The dataset is divided into 70% and 30% for training and testing, respectively. The dataset is divided into each unit as follows:

- For the Research Scorecard, 104 instances for training and the balance 46 for testing.
- For the Lecturers by Students, 106 instances for training and the remaining 44 for testing, and
- Finally, for the Teaching Scorecard, 105 instances are used for training and the remaining 45 for testing.

The three datasets are tested using the Rapid Miner Studio 102 platform to analyze the comparison results of the NBC and SVM methods. This testing aims to determine the levels of accuracy, precision, and recall on both methods. Next, in section 4.1we study the comparison results of Evaluation Metrix for the NBC and SVM methods for the Research Scorecard, Evaluation of Lecturers by Students, and Teaching Scorecard.

Comparison Results for NBC and SVM Methods for the Research Scorecard Dataset

Comparison of the Accuracy, Precision, and Recall results from the NBC Model for the Research Scorecard dataset (Figure 4.1) shows that the prediction for 46 test instances achieves a true accuracy rate of 100%, indicating that the accuracy level of this model is excellent. The precision testing results a value of 100%, meaning that the predictions for this test data align with the actual data. Similarly, the recall testing provides a value of 100%, signifying that the information obtained from the prediction results of this model is highly accurate. Also, the testing results of the SVM model reveals a true accuracy rate of 100% for the Research Scorecard dataset, which indicates that the accuracy level of this model is outstanding. The precision testing results also show a value of 100%, indicating that the predictions for this test data match the actual data. Likewise, the recall testing results show a value of 100%, indicating that the information obtained from the prediction results of this model is highly accurate.



	Accuracy	Precision	Recall
SVM	100%	100%	100%
NBC	93.80%	72.04%	75%

Figure 3
. Comparison of NBC and SVM Methods for the Research Scorecard Dataset

Comparison Results for NBC and SVM Methods for the Evaluation of Lecturers by Students Dataset

Comparison of the Accuracy, Precision, and Recall results from the NBC model for the Student Evaluation Dataset shows that the prediction for 44 test instances of Student Evaluation achieves an accuracy rate of 93.18%. The precision testing results indicate a value of 72.04%, meaning that the predictions for this test data have 27.96% correct predictions within the actual class. Meanwhile, the recall testing results show a value of 75%, indicating that 25% of the information obtained from the prediction results does not match this model. On the other hand, the testing results of the SVM model show an accuracy rate of 100%, indicating that the accuracy level of this model is excellent. The precision testing results also indicate a value of 100%, meaning that the predictions for this test data match the actual data. Similarly, the recall testing results show a value of 100%, indicating that the information obtained from the prediction results of this model is highly accurate. The results are depicted in Figure 4.2.

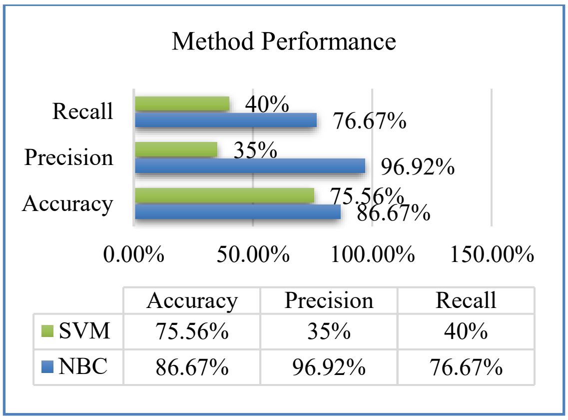


Figure 4
. Comparison Results for NBC and SVM Methods for the Evaluation of Lecturers by Students Dataset

Comparison Results for NBC and SVM Methods for the Teaching Scorecard Dataset

Comparison of the Accuracy, Precision, and Recall results from the NBC model for the Teaching Score Card dataset shows that the prediction for 45 test instances achieves an accuracy rate of 86.67%. The precision testing results indicate a value of 96.92%, meaning that the predictions for this test data have 3.08% correct predictions within the actual class. Meanwhile, the recall testing results show a value of 76.67%, indicating that 23.33% of the information obtained from the prediction results does not match this model. Meanwhile, the testing results of the SVM model show an accuracy rate of 75.56%. The precision testing results indicate a value of 35%, meaning that the predictions for this test data have 65% correct predictions within the actual class. Similarly, the recall testing results show a value of 40%, indicating that 60% of the information obtained from the prediction results does not match this model as shown in Figure 4.3.

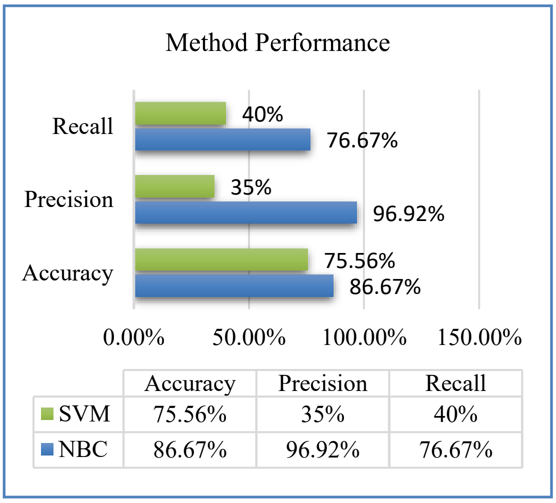


Figure 5
. Comparison Results for NBC and SVM Methods for the Teaching Scorecard Dataset

The Comparison Results of Performance Between the NBC and SVM methods

Comparison results for the test data using Naive Bayes and SVM methods to evaluate faculty performance indicate variations in prediction across three scenarios: For the first scenario, with the Research Score Card test data consisting of 46 instances, both methods demonstrate excellent prediction. Naive Bayes achieves an accuracy, precision, and recall of 100%, as does SVM. In the second scenario, with the Evaluation of Lecturers by Students test data of 44 instances, SVM outperforms Naive Bayes with perfect accuracy, precision, and recall scores of 100%, compared to Naive Bayes' accuracy of 93.18%, precision of 72.04%, and recall of 75%. In the third scenario, assessing Teaching Score Card test data with 45 instances, Naive Bayes yields superior predictions with an accuracy of 86.67%, precision of 96.92%, and recall of 76.67%, compared to Naive Bayes' accuracy of 75.56%, precision of 35%, and recall of 40%. These comparisons reveal that the choice of method can significantly impact prediction performance across different evaluation scenarios.

DISCUSSION

The use of feature selection with information gain in this research did not have a significant effect on increasing accuracy, precision and recall as in research conducted by (Omuya et al, 2021), where the results of applying feature selection with information gain using the NBC method had an effect on increasing accuracy from 94.89% rose to 97.81%, precision from 95% rose to 97.80% and recall from 94.90% rose to 97.80%. Likewise, feature selection with information gain using the SVM method, where there is an increase in accuracy from 67.77% to 100%, precision from 50% to 100% and recall percentage from 63% to 100%. Similar research is conducted by Vijayashree and Sultana (2018), where the use of feature selection with information gain using the NBC method had a significant effect on increasing the accuracy percentage from 79.35% to 82.65%. However, using the SVM method resulted in a decrease in accuracy from 75.23% to 74.12%.

Some of the researchers' findings regarding factors that influence increasing accuracy, precision, and recall, apart from feature selection, are the complexity and pattern of the dataset being tested. This is proven by increasing the number of datasets, both teaching scorecards and evaluation datasets from students, and research scorecards. The results of the Teaching Scorecard dataset test using the NBC method show that there is a significant effect on increasing the number of datasets from 150 to 1000 datasets with the data pattern in the form of a series of numbers from 0 to 4 (0,1,2,3,4). Using the number of data sets with these data patterns, the comparison results obtained in sequence, namely accuracy, precision, and recall, are initially 86.67%, 96.92%, 76.67%, increasing to 98.00%, 98.10% and 98. 00%. Likewise, the test results using the SVM method, obtained sequential comparison results, namely accuracy, precision, and recall, are 75.56%, 35.00%, 40.00%, increasing to 92.00%, 92.71% and 92. .00%. Further evidence is also found in the evaluation dataset of students using the NBC method, showing that there is a significant effect on increasing the number of datasets from 150 to 1000 datasets, with the data pattern in the form of a series of numbers from 1 to 5 (1,2,3,4,5). With the number of datasets and data patterns, the comparison results obtained sequentially, namely accuracy, precision, and recall, are 93.18%, 72.04%, and 75.00%, increasing to 100%, 100% and 100%. Test results using the SVM method, obtained sequential comparison results, namely accuracy, precision, and recall, are 100%, 100%, 100%, decreasing to 96.67%, 97.14% and 96.67%.

Specifically, for the Research Scorecard dataset, the factor that influences accuracy, precision, and recall using both the NBC and SVM methods is increasing the number and pattern of the dataset, where the data pattern used initially is 0, 2, 3, 5, 7, 8, 10, 15, 20. In the tests carried out by researchers using the NBC method, significant results are obtained with the first scenario with a dataset of 150, and the data pattern used a series of numbers 0 to 4 (0,1,2,3,4), and obtained accuracy results, precision, and recall, respectively, are 93.33%, 95.00% and 93.33%. The second scenario involves increasing 1000 datasets with better accuracy, precision, and recall results, respectively, namely 98.00%, 98.10%, and 98.88%. In testing using the SVM method on data patterns with rows of numbers 0 to 4 with a dataset of 150, the accuracy, precision, and recall test results are obtained sequentially, namely 73.33%, 76.00% and 73.33%. By increasing the number of datasets to 1000 datasets, the test results obtained are an accuracy of 92.33%, a precision of 92.97% and a recall of 92.33%.

CONCLUSION AND IMPLICATIONS

The results of comparing 150 Research Scorecard datasets by dividing 70% of the data for training and 30% for testing using the NBC and SVM methods yielded identical accuracy, precision, and recall results, all at 100%. Meanwhile, the comparison of 150 lecturer evaluations from student datasets using the NBC method resulted in higher accuracy, precision, and recall, all at 100%, compared to using the SVM method, which achieved an accuracy of 93.18%, precision of 72.04%, and recall of 75%. Subsequently, the comparison results for 150 Teaching Scorecard datasets using the NBC method showed an accuracy of 86.67%, precision of 96.92%, and recall of 76.67%, whereas using the SVM method resulted in an accuracy of 75.56%, precision of 35%, and recall of 40%.

Based on the test results and findings related to this research, it is indicated that neither method does not exhibits a significant influence on whether undergoing the feature selection process or not towards the improvement of accuracy, precision, and recall. The findings observed in the Teaching Scorecard dataset and lecturer evaluations from students are influenced by the magnitude or insignificance of the dataset size, while in the Research Scorecard dataset, it is influenced by two factors: the number of datasets and the data pattern. To achieve satisfactory accuracy, precision, and recall results in dataset testing, it is necessary to establish a standard number of datasets with an appropriate data pattern, such as the following data patterns: 0 to 4 (0, 1, 2, 3, 4) or 1 to 5 (1, 2, 3, 4, 5) with a dataset count of 1000.

ACKNOWLEDGEMENTS

The author expresses gratitude to the Dili Institute of Technology (DIT) for the research grant and extends special thanks to the Center for Applied Research and Policy Studies and Community Services (CARPS-CS), the Center for Evaluation and Quality Assurance (CEQA), and the Academic Department for facilitating the entire research process.

REFERENCE

1. Abbi Nizar Muhammad, Saiful Bukhori, & Priza Pandunata. (2019). Proceedings, 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE 2019): October 16th-17th 2019, Jember, Indonesia.
2. Amaya-Tejera, N., Gamarra, M., Vélez, J. I., & Zurek, E. (2024). A distance-based kernel for classification via Support Vector Machines. Frontiers in Artificial Intelligence, 7. https://doi.org/10.3389/frai.2024.1287875
3. Berrar, D. (2019). Bayes’ Theorem and Naive Bayes Classifier. In Encyclopedia of Bioinformatics and Computational Biology (Vols. 1–3, pp. 403–412). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20473-1
4. Cherrington, M., Airehrour, D., Lu, J., Thabtah, F., Xu, Q., & Madanian, S. (2019, October). Particle swarm optimization for feature selection: A review of filter-based classification to identify challenges and opportunities. In 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 0523-0529). IEEE.
5. Firmahsyah, F., & Gantini, T. (2016). Penerapan metode content-based filtering pada sistem rekomendasi kegiatan ekstrakulikuler (Studi Kasus di Sekolah ABC). Jurnal Teknik Informatika dan Sistem Informasi, 2(3).
6. Jalota, C., & Agrawal, R. (2019b). Analysis of Educational Data Mining using Classification. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16 thFeb 2019.
7. Joseph, V. R. (2022). Optimal ratio for data splitting. Statistical Analysis and Data Mining, 15(4), 531–538. https://doi.org/10.1002/sam.11583
8. Joseph, V. R. (2022). Optimal ratio for data splitting. Statistical Analysis and Data Mining, 15(4), 531–538. https://doi.org/10.1002/sam.11583

9. Lei, S. (2012). A feature selection method based on information gain and genetic algorithm. Proceedings - 2012 International Conference on Computer Science and Electronics Engineering, ICCSEE 2012, 2, 355–358. https://doi.org/10.1109/ICCSEE.2012.97
10. Ma, T. M., Yamamori, K., & Thida, A. (2020). A Comparative Approach to Naïve Bayes Classifier and Support Vector Machine for Email Spam Classification. 2020 IEEE 9th Global Conference on Consumer Electronics, GCCE 2020, 324–326. https://doi.org/10.1109/GCCE50665.2020.9291921
11. Mofizur Rahman, C., Afroze, L., Sultana Refath, N., & Shawon, N. (n.d.). Iterative Feature Selection Using Information Gain & Naïve Bayes for Document Classification.
12. Morán-Fernández, L., Bólon-Canedo, V., & Alonso-Betanzos, A. (2022). How important is data quality? Best classifiers vs best features. Neurocomputing, 470, 365–375. https://doi.org/10.1016/j.neucom.2021.05.107
13. Muhammad, A. N., Bukhori, S., & Pandunata, P. (2019, October). Sentiment analysis of positive and negative of youtube comments using naïve bayes–support vector machine (nbsvm) classifier. In 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE) (pp. 199-205). IEEE.
14. Odhiambo Omuya, E., Onyango Okeyo, G., & Waema Kimwele, M. (2021). Feature Selection for Classification using Principal Component Analysis and Information Gain. Expert Systems with Applications, 174. https://doi.org/10.1016/j.eswa.2021.114765
15. Omuya, E. O., Okeyo, G. O., & Kimwele, M. W. (2021). Feature selection for classification using principal component analysis and information gain. Expert Systems with Applications, 174, 114765.
16. Rakhmanov, O. (2020). A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments. Procedia Computer Science, 178, 194–204. https://doi.org/10.1016/j.procs.2020.11.021
17. Ramanda Hasibuan, M. (2019). Pemilihan Fitur dengan Information Gain untuk Klasifikasi Penyakit Gagal Ginjal menggunakan Metode Modified K-Nearest Neighbor (MKNN) (Vol. 3, Issue 11). http://j-ptiik.ub.ac.id
18. Rashid, T. A., & Ahmad, H. A. (2016). Lecturer performance system using neural network with Particle Swarm Optimization. Computer Applications in Engineering Education, 24(4), 629–638. https://doi.org/10.1002/cae.21737
19. Shahi, T. B., & Pant, A. K. (2018, February). Nepali news classification using Naive Bayes, support vector machines and neural networks. In 2018 international conference on communication information and computing technology (iccict) (pp. 1-5). IEEE.
20. Varghese, S. M., & Sushmitha, M. N. (2014). Efficient Feature Subset Selection Techniques for High Dimensional Data. International Journal of Innovative Research in Computer and Communication Engineering, 2(3).
21. Vijayashree, J., & Sultana, H. P. (2018). A Machine Learning Framework for Feature Selection in Heart Disease Classification Using Improved Particle Swarm Optimization with Support Vector Machine Classifier. Programming and Computer Software, 44(6), 388–397. https://doi.org/10.1134/S0361768818060129
22. Zareapoor, M., & K. R, S. (2015). Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection. International Journal of Information Engineering and Electronic Business, 7(2), 60–65. https://doi.org/10.5815/ijieeb.2015.02.08