International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN(Online): 2347-5552, Volume-12, Issue-3, May 2024
https://doi.org/10.55524/ijircst.2024.12.3.1
Article ID IRP1448, Pages 1-5
www.ijircst.org

Enhancing Student Performance Prediction Using a Combined SVM-Radial Basis Function Approach

Yuan Anisa[1], Winda Erika[2], and *Fadhillah Azmi[3]

[1,3] Department of Electrical Engineering, Universitas Medan Area, Medan Indonesia
[2] Department of Informatics Engineering, Universitas Pembangunan Pancabudi, Medan, Indonesia

Correspondence should be addressed to Fadhillah Azmi; azmi.fadhillah007@gmail.com

Received 1 April 2024; Revised 13 April 2024; Accepted 22 April 2024

ABSTRACT- This research aims to improve student performance predictions using a combined SVM (Support Vector Machine) and radial basis function (RBF) approach. The developed model utilizes a combination of the strengths of SVM in handling class separation and the ability of RBF to capture complex patterns in data. Student assessment data, including math, reading, and writing scores, is used as a feature to predict student performance on tests. Preprocessing steps, including feature normalization and label encoding, are applied to prepare the data for model training. Next, the SVM model with the RBF kernel is initialized and optimized using GridSearchCV to find the best parameters. Model evaluation was carried out using the R^2 metric to evaluate how well the model predicts student performance. Experimental results show that the combined SVM-RBF approach can improve student performance predictions with fairly accurate prediction results of 88%. The practical implication of this research is the development of a more accurate model for predicting student performance, which can be used as a tool to improve educational interventions and decision-making in educational institutions.

KEYWORDS- Support Vector Machine, Radial Basis Function, Student Performance, Combined SVM, Student Assessment Data.

I. INTRODUCTION

Education is a vital aspect of community and individual development. In the current era of information and technology, analyzing students' learning performance has become increasingly important to understand the factors that influence their academic achievement. Student learning performance is influenced by a number of factors, including student characteristics, the learning environment, and socio-economic factors [1]. A deep understanding of the factors that influence student learning performance allows education providers to identify the challenges students face and develop appropriate intervention strategies. In the last few decades, data analysis and machine learning techniques have become important tools in understanding student learning performance patterns to get an idea of student learning levels when measuring student success or failure [2]. This approach allows educational researchers and practitioners to extract insights from available data, identify underlying patterns, and predict student behavior and performance. Additionally, predictive models can help instructors guide students to success in a course and are used to determine which activities and material are more important for course assessment [3]. Thus, the integration of data analysis and machine learning techniques opens the door to the development of more effective and evidence-based educational strategies.

The Support Vector Machine (SVM) method has become a popular machine learning technique for analyzing student learning performance [4]. SVM is a powerful learning method for classification and regression that has a good ability to handle complex and non-linear datasets [5][6]. The combination of SVM (Support Vector Machine) with other methods or with variations of SVM itself has become an important approach in dealing with a number of problems in data analysis. One problem that is often faced is class imbalance in the dataset, where SVM may not be effective in classifying minority classes [7]. In this situation, ensemble techniques such as combining SVM with oversampling or undersampling methods can help improve the performance. Additionally, SVMs sometimes have difficulty handling very complex and non-linear patterns in the data.

In this case, the combination of SVM with other non-linear classification methods can help improve its capabilities. Additionally, SVMs can experience computational challenges when applied to very large datasets. The combination of SVM with dimensionality reduction techniques or by using more structured feature subsets can help improve its efficiency and performance on large datasets. Finally, the diversity of data features and characteristics can also be an obstacle for SVM. In this situation, combining SVM with variations of SVM such as linear and non-linear SVM or using multiple kernels can help overcome this challenge [8]. Through this approach, SVM can improve its performance and increase its flexibility in handling various types of datasets and complex classification problems.

The use of Radial Basis Function (RBF) kernels in SVM allows the model to capture complex structures in data [8]. The RBFNN method implements mathematical functions that are often used in modeling and data analysis [9][10].

The RBF kernel converts the original feature space into a higher-dimensional feature space, where the data points become more linearly separated. This allows SVM to separate classes of data that may not be linearly separated in the original dimensions. Thus, the use of SVM with an RBF kernel can be very useful in analyzing students' learning performance, especially when there are complex relationships between factors that influence their performance. This study aims to analyze student learning performance using the SVM method with the RBF kernel, with the aim of identifying patterns and factors that have the most influence on student academic achievement. Thus, this research has the potential to provide valuable insights for education providers in the development of more effective strategies to improve educational quality and student learning performance.

The study will combine student test score data with various student attributes, such as gender, ethnicity, parental education level, lunch, and participation in test preparation courses. By utilizing SVM with an RBF kernel, this research aims to identify the most significant factors influencing student learning performance. This research has important relevance in the context of educational development, as it can provide valuable insight into the factors that influence student learning performance. The results of this research can be used to inform education policy and develop intervention strategies that are more effective in improving the quality of education. The proposed research has significant relevance and contribution in the fields of education, machine learning, or data analysis.

II. METHODOLOGY

The following are the research steps for studying student performance using the Support Vector Machine (SVM) method with a Radial Basis Function (RBF) kernel:

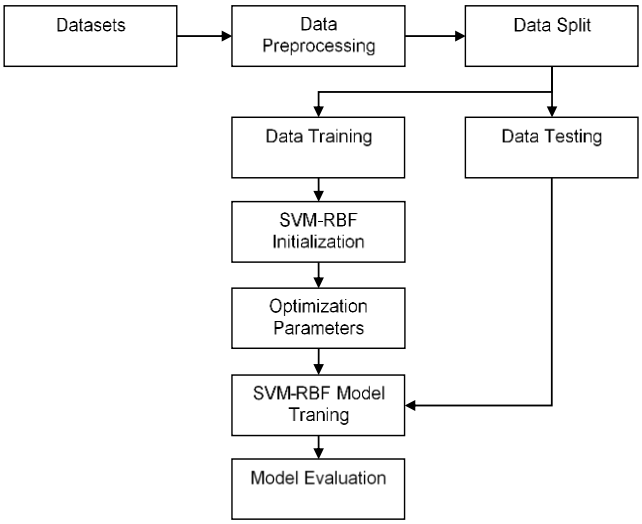

Figure 1: Proposed Methods

The following are the research steps for each step implemented in Figure 1 can be described as follows:

1. The first step is to collect data that is relevant to the problem you want to solve. In this case, data regarding student exam results is collected in a CSV file.
2. Data often requires preprocessing before being used for model training. In this program, preprocessing steps include feature normalization using XScaled to ensure uniform scaling as well as label encoding to convert categorical variables to numeric format.
3. The data is divided into two subsets, namely training data and test data. This process allows us to train a model on training data and test its performance on previously unused test data.
4. The SVM (Support Vector Machine) model with the RBF (Radial Basis Function) kernel is initialized and trained on the data used. This model will be used to predict students' math scores based on other features.
5. To improve the performance of the model, we can optimize certain parameters of the model. In this program, we use GridSearchCV to search for the best combination of parameters (C and gamma) for the SVR model.
6. The trained model is evaluated using test data to measure its performance. In this program, we use the R^2 metric to evaluate how well the model predicts a student's math score. Performance using the Support Vector Machine (SVM) method with a Radial Basis Function (RBF) kernel.

A. Datasets

Data used for analysis of student performance, such as test scores, student demographic information (gender, ethnicity, and parental education level), information about lunch, and whether the student took a test preparation course. This data was obtained from the internet page at the address: https://www.kaggle.com/datasets/bhavikjikadara/student-study-performance. The attribute information from the dataset used can be seen in Table 1 below.

Table 1: Datasets

Attributes	Description	Value
Gender	Sex of students	Male/female
Race/ethnicity	Ethnicity of students	A, B,C, D, E
Parental level of education	Parents' final education	bachelor's degree, some college, master's degree, associate's degree, high school
Lunch	Having lunch before test	standard or free/reduced
Test	Test preparation course	complete or not complete before test
Math	Math score	0 - 100
Reading	Reading score	0 - 100
Writing	Writing score	0 - 100

B. Preprocessing Data

In the data preprocessing stage, the first step is feature normalization using *Xscaled*. This process is carried out to ensure that all numerical features, such as math, reading, and writing scores, are on a uniform scale. Feature normalization can be represented by the following equation 1.

$$X_{scaled} = \frac{X-\mu}{\sigma} \quad (1)$$

Where: *Xscaled* is a normalized feature; *X* is a native feature; μ is the average of the original features; and σ is the standard deviation of the original feature.

The next step is label encoding to convert the categorical variable into a numeric format. This process is carried out so that the SVM model can process these variables as input. After that, data cleaning is carried out by deleting rows that have empty values (NaN). This process is carried out so that only complete data is used in model training.

C. Data Split

Dividing a dataset into a training set and a testing set is a crucial step in developing a machine learning model. The training set is a subset of the dataset used to train the model, where the model learns patterns from existing and technical features with the desired output. The use of this training set helps the model identify underlying patterns in the data. On the other hand, a test set is a portion that is removed from a dataset to test the performance of a drilled model.

The data in the test set is never used in the training process, so it can provide an objective evaluation of the model's ability to generalize to new data. The proportion of data sharing can vary, but generally most of the data is allocated to the training set, while the rest is allocated to the test set. This data sharing helps avoid overfitting, where the model "memorizes" the training data and fails to generalize to new data. By separating data sets into training and test sets, we can ensure that machine learning models can provide accurate and reliable predictions on never-before-seen data

D. Inisialisasi SVM-RBF

Implementing a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel involves selecting the necessary parameters, such as C and gamma parameters [11], as well as the model training process using training data.

The C parameter controls the trade-off between the margin and the number of classification errors accepted by the SVM model [12]. Larger C values lead to larger penalties for misclassification, which can result in tighter decision boundaries. Conversely, a smaller C value results in a larger margin, which can increase tolerance for misclassification.

The gamma parameter controls the model's flexibility with respect to the training data. Larger gamma values cause the RBF kernel function to be more sensitive to differences between training data. Smaller gamma values result in wider RBF kernel functions, which can result in smoother decision boundaries. The equations for the SVM model with the RBF kernel are not included in this step because this step is more focused on parameter setting and model preparation.

Training a Support Vector Machine (SVM) model with a Radial Basis Function (RBF) kernel involves optimizing parameters and adjusting weights to minimize the SVM loss function. In general, there is no single mathematical equation that covers all the steps in SVM model training. However, some concepts can be explained in the form of more general equations. The SVM Objective Function (Hinge Loss) [13]:

$$Loss(y, f(x)) = \sum_{i=1}^{n} max(0, 1 - y_i f(x_i)) \quad (2)$$

Where:

y_i is the label of the ith data sample.

$f(x_i)$ is the mapping function of the SVM model on the i-th data sample.

If $y_i f(x_i) \geq 1$, then the loss is 0, which means the sample is classified correctly and does not contribute to the loss.

If $y_i f(x_i) < 1$, then there is a classification error, and the loss will increase according to the difference between $1 - y_i f(x_i)$.

The RBF Kernel Functions [14]:

$$K(x_i, x_j) = exp(-\gamma \parallel x_i - x_j \parallel^2) \quad (3)$$

Where: x_i and x_j are the two feature vectors to be compared; and γ is an RBF kernel parameter that controls the flexibility of the model. The smaller the value of γ, the wider the radius of the kernel function. The SVM Decision Function [15]:

$$f(x) = \sum_{i=1}^{m} sign(\alpha_i y_i K(x, x_i) + b) \quad (4)$$

Where: α_i is the Lagrange coefficient associated with the feature vector x_i; y_i is the label of the feature vector *xi*; and *b* is the bias that affects the position of the hyperplane.

In essence, training an SVM model with an RBF kernel involves adjusting the parameters α and *b* such that the SVM decision function minimizes a loss function, which is usually a combination of a loss and regularization function that controls the complexity of the model.

III. RESULT AND DISCUSSION

A. Results

In this research, the SVM model is evaluated using several key parameters, namely the kernel, C parameters, and gamma parameters. The kernel is a mathematical function that is used to convert the feature space to a higher-dimensional space so that the data can be separated linearly. The C parameter determines how hard the SVM model will handle margin violations (misclassification) in the training data. The higher the C value, the harder the model will handle margin violations, which can result in overfitting. Meanwhile, the gamma parameter controls how far one data sample influences other data samples. The higher the gamma value, the smaller the influence range of one data sample.

In evaluating the performance of SVM models, relevant evaluation metrics such as the R^2 score are used. The R^2 score is a coefficient of determination that measures how well the model can explain variations in the data. The R^2 score value ranges from 0 to 1, where a higher value indicates better model performance in predicting the data. Performance evaluation is carried out on training data and testing data to ensure that the model is able to generalize patterns learned from training data to new, never-before-seen data. Thus, the use of appropriate SVM parameters and a comprehensive evaluation of model performance are key to ensuring the success of the SVM model in predicting student performance in exams.

The results of the tests carried out can be seen in Table 2 below:

Table 2. The Performance Results of Student Performance Prediction Model

Methods	R^2 Score
SVM	0.87
DT	0.73
RF	0.86
SVM-RBF	0.88

Table 2 shows the R^2 score of the four machine learning methods tested, namely SVM (Support Vector Machine), DT (Decision Tree), RF (Random Forest), and SVM-RBF (Support Vector Machine with RBF kernel). The results provided are the R^2 score (coefficient of determination) from several machine learning methods tested.

1. SVM has an R^2 score of 0.87, indicating that the SVM model can explain about 87% of the variation in student performance data in exams.
2. DT has an R^2 score of 0.73, which indicates that the decision tree model is able to explain around 73% of the variation in the data.
3. RF has an R^2 score of 0.86, indicating that the Random Forest model is able to explain around 86% of the variation in the data, almost equivalent to SVM.
4. SVM-RBF has an R^2 score of 0.88, indicating that the SVM model with the RBF (Radial Basis Function) kernel is able to explain around 88% of the variation in the data, slightly higher than the regular SVM model.

From these results, it can be seen that the SVM model with the RBF kernel has the best performance in predicting student performance in the exam, followed by the SVM, RF, and finally DT models

B. Discussion

In this research, model performance evaluation is an important aspect that is carefully analyzed. An analysis of the results of the performance evaluation of the SVM model was carried out in both the training and testing phases. Evaluation metrics such as the R2 score or accuracy are used to assess the extent to which the SVM model is able to predict student performance in exams. Additionally, a comparison of the performance of the SVM model with combined SVM approaches, such as SVM with the RBF kernel, was performed to gain a deeper understanding of the relative performance of each model. A discussion of factors influencing model performance is also warranted, including feature selection, SVM parameter optimization, and the impact of combined SVM approaches. This will help clarify how these factors contributed to the results observed in this study. Next, the advantages of combined SVM approaches, such as SVM with an RBF kernel, will be compared with those of single SVM.

This analysis will address the potential for improving the prediction performance of student study performance as well as possible improvements to the interpretability or flexibility of the model. The implications of the research findings for educational practice and future research will also be discussed in detail, including recommendations for further development of models or modeling strategies. In addition, research limitations will also be identified, such as sample size, data sources, or analysis methods used, along with a discussion of how these limitations may affect the interpretation of results and generalization of findings. By considering all these aspects, this research will provide a more holistic understanding of the performance of SVM models and their contribution to the prediction of student performance in exams

IV. CONCLUSION

The conclusion of this research shows that a combined approach between support vector machines (SVM) and radial basis functions (RBF) is able to improve predictions of student performance. Evaluation of the performance of the SVM model shows good results, with an R2 score of 0.87. However, the SVM approach combined with the RBF kernel gives slightly better results, with an R^2 score of 0.88. Factors such as feature selection, SVM parameter optimization, and combined SVM approaches play an important role in improving model performance. The advantage of the combined SVM approach lies in its ability to capture complex patterns in the data, thereby improving the prediction performance of student study performance. The implications of this research finding are highly relevant for educational practice and future research, especially in the development of more accurate and effective student performance prediction models.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] K. Liu, J. Yao, D. Tao, and T. Yang, “Influence of Individual-technology-task-environment Fit on University Student Online Learning Performance: The Mediating Role of Behavioral, Emotional, and Cognitive Engagement,” Educ. Inf. Technol., vol. 28, no. 12, pp. 15949–15968, 2023, doi: 10.1007/s10639-023-11833-2.

[2] A. Dhankhar, K. Solanki, and A. Rathee, “International Journal of Advanced Trends in Computer Science and Engineering Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse75842019.pdf Predicting Student ’ s Performance by using Classification Methods,” vol. 8, 2019.

[3] P. G. Student, “a Review of Student Performance Prediction Techniques in Virtual,” vol. 9, no. 8, pp. 183–190, 2021.

[4] F. Janan and S. K. Ghosh, “Prediction of student’s performance using support vector machine classifier,” Proc. Int. Conf. Ind. Eng. Oper. Manag., no. September, pp. 7078–7088, 2021, doi: 10.46254/an11.20211237.

[5] H. Wang, J. Xiong, Z. Yao, M. Lin, and J. Ren, “Research survey on support vector machine,” Int. Conf. Mob. Multimed. Commun., vol. 2017-July, pp. 95–103, 2017, doi: 10.475/eai.13-7-2017.2270596.

[6] S. Saikin, S. Fadli, and M. Ashari, “Optimization of Support Vector Machine Method Using Feature Selection to Improve Classification Results,” JISA(Jurnal Inform. dan Sains), vol. 4, no. 1, pp. 22–27, 2021, doi: 10.31326/jisa.v4i1.881.

[7] B. Gaye, D. Zhang, and A. Wulamu, “Improvement of Support Vector Machine Algorithm in Big Data Background,” Math. Probl. Eng., vol. 2021, 2021, doi: 10.1155/2021/5594899.

[8] H. Al Azies, D. Trishnanti, and E. Mustikawati P.H, “Comparison of Kernel Support Vector Machine (SVM) in Classification of Human Development Index (HDI),” IPTEK J. Proc. Ser., vol. 0, no. 6, p. 53, 2019, doi:

10.12962/j23546026.y2019i6.6339.
[9] M. K. Gibran and A. Saleh, "A Hybrid RBF Neural Network and FCM Clustering for Diabetes Prediction Dataset," J. Comput. Sci. Inf. Technol. Telecommun. Eng., vol. 4, no. 2, pp. 395–401, 2023, doi: 10.30596/jcositte.v4i2.15573.
[10] A. Saleh, T. Tulus, and S. Efendi, "Analysis of Accurate Learning in Radial Basis Function Neural Network Using Cosine Similarity on Leaf Recognition," 2019, doi: 10.4108/eai.20-1-2018.2281924.
[11] I. S. Al-Mejibli, J. K. Alwan, and D. H. Abd, "The effect of gamma value on support vector machine performance with different kernels," Int. J. Electr. Comput. Eng., vol. 10, no. 5, pp. 5497–5506, 2020, doi: 10.11591/IJECE.V10I5.PP5497-5506.
[12] C. L. Ma and Y. B. Yuan, "A novel support vector machine with globality-locality preserving," Sci. World J., vol. 2014, 2014, doi: 10.1155/2014/872697.
[13] A. V. Asimit, I. Kyriakou, S. Santoni, S. Scognamiglio, and R. Zhu, "Robust Classification via Support Vector Machines," Risks, vol. 10, no. 8, pp. 1–25, 2022, doi: 10.3390/risks10080154.
[14] K. Saputra, "Perbandingan Kinerja Fungsi Kernel Algoritma Support Vector Machine Pada Klasifikasi Penyakit Padi," Ijccs, vol. x, No.x, no. x, pp. 1–5, 2023.
[15] A. Z. Praghakusma and N. Charibaldi, "Komparasi Fungsi Kernel Metode Support Vector Machine untuk Analisis Sentimen Instagram dan Twitter (Studi Kasus : Komisi Pemberantasan Korupsi)," JSTIE (Jurnal Sarj. Tek. Inform., vol. 9, no. 2, p. 88, 2021, doi: 10.12928/jstie.v9i2.20181.
[16] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Comput. Sci., vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.