IBBR Early Career Symposium, Rockville, Maryland, United States, September 17, 2025

Support Vector Machine Hyperparameter Tuning for Student Addiction Prediction

Sri Kavya Penta
Information Systems
University of Maryland Baltimore County
Maryland, USA

Elaina Owens
Information Systems
University of Maryland Baltimore County
Maryland, USA

Abstract—Student addiction is a severe condition with far-reaching academic, psychological, and societal implications. Early identification of students at risk for substance use disorders (SUD) enables timely intervention and may reduce long-term impacts. While machine learning has been applied to predict addiction behaviors, few studies have systematically examined the impact of Support Vector Machine (SVM) kernel selection, regularization parameter tuning (C), and polynomial degree adjustment.

In this work, we applied SVM classifiers to predict student addiction using behavioral, academic, and social attributes. For computational efficiency, the experiments were conducted on the first 2000 rows of the dataset, as extending to the full dataset did not yield significant differences in performance. We systematically tested different kernel functions (linear, radial basis function, polynomial), values of C, and polynomial degrees, and evaluated their effect on model accuracy. Among all combinations, several configurations including the linear kernel, RBF kernel with C=1, and polynomial kernels with degrees 2–3 achieved the highest accuracy of 78.9%.

To increase the biological relevance of our study, we incorporate a neuroscience-based overview of addiction, highlighting how repeated drug use alters brain chemistry, particularly dopamine and opioid pathways, leading to impaired impulse control and increased risk-taking behavior. By integrating biological context with machine learning optimization, this study contributes a novel perspective to student health analytics and emphasizes the importance of model calibration in improving early risk prediction within education and public health domains.

I. INTRODUCTION

Student drug addiction is an emerging concern with implications that extend beyond academic performance to emotional and psychological well-being. Traditional detection methods rely heavily on human observation, which limits both consistency and scalability. Machine learning (ML) offers a scalable, data-driven alternative for identifying at-risk individuals based on behavioral and academic patterns.

Among various ML techniques, Support Vector Machines (SVM) are particularly well-suited for binary classification due to their ability to construct optimal separating hyperplanes. However, prior studies often implement SVMs without systematically analyzing the effect of kernel choice, regularization parameter (C), or polynomial degree on predictive performance [3], [4].

In this study, we perform a targeted evaluation of SVM configurations for student addiction prediction. We investigate the impact of different kernel types, hyperparameter tuning, and class balancing on model accuracy. After minimal preprocessing, we applied under-sampling to mitigate class imbalance and highlight the most effective SVM setup for this high-stakes classification task.

II. BIOLOGICAL BACKGROUND

Addiction, clinically defined as substance use disorder (SUD), is a chronic, relapsing condition that alters the brain's structure and function over time. Repeated exposure to addictive substances disrupts the brain's natural reward system, primarily by overstimulating dopamine circuits and opioid receptors [1], [2], leading to impaired impulse control, increased compulsive behavior, and diminished executive decision-making abilities. This neurochemical imbalance makes recovery difficult and relapse common, especially among adolescents and young adults.

In our study, we align this biological context with a machine learning framework to better identify at-risk students. We employ Support Vector Machines (SVM), a binary classification algorithm whose decision boundary is shaped by key hyperparameters: the kernel type, the regularization constant (C), and for polynomial kernels, the degree. These parameters govern the model's flexibility, sensitivity to noise, and generalization ability.

Given the imbalance in real-world addiction datasets where the number of addicted students is much lower than nonaddicted we applied random under-sampling to balance the classes. This approach helps ensure the model does not default to majority class predictions and improves its ability to recognize high-risk students in the minority class.

III. LITERATURE REVIEW

The application of machine learning techniques to addiction prediction has gained momentum in recent years, especially with the increased availability of behavioral and academic data. Various classification models including logistic regression, decision trees, random forests, and risk of substance use [3], [4] utilized to identify individuals at risk of substance use, particularly within university and adolescent populations. These studies commonly rely on structured data sources such as survey responses and academic performance indicators.

SVMs have shown promise in binary classification tasks due to their ability to define optimal decision boundaries in complex, high-dimensional feature spaces. They are often favored for tasks involving nonlinear relationships and are known for their flexibility when paired with kernel functions such as radial basis function (RBF), polynomial, or linear.

However, many existing implementations of SVMs in health analytics rely on default kernel selections and perform minimal hyperparameter tuning. The effects of key parameters such as kernel type, regularization strength (C), and polynomial degree on model generalization and sensitivity are often underexplored. Moreover, class imbalance, especially in datasets where positive (addicted) cases are in the minority, remains a remains a persistent challenge [4] that can undermine the reliability of model predictions.

This study addresses these gaps by conducting a focused evaluation of how different SVM configurations influence prediction outcomes for student addiction classification. Specifically, it explores the impact of kernel type, C value, and polynomial degree on model accuracy. To enhance the model's ability to detect addicted individuals, the dataset is balanced using random under-sampling, and recall is prioritized as the primary evaluation metric. This approach offers a targeted contribution by optimizing a single algorithm rather than comparing multiple models, thereby providing deeper insight into the effectiveness of parameter tuning in improving predictive sensitivity for critical health-related outcomes.

IV. METHODOLOGY

A. Dataset Description

The dataset consisted of anonymized student records, each representing an individual with a series of binary features. These features captured a range of behavioral and academic indicators, such as social isolation, academic performance decline, legal troubles, and symptoms of withdrawal. The target label," Addiction Class," indicated whether the student was classified as addicted (1) or not addicted (0).

B. Data Preprocessing

The raw dataset contained categorical features with values such as "Yes," "No," and blanks. These were converted into binary numerical values, where "Yes" was encoded as 1, "No" as 0, and missing entries as -1. All categorical features were then standardized using z-score normalization to ensure comparability across attributes.

C. Parameter Tuning

To evaluate the impact of hyperparameters on model performance, we systematically tested different configurations of the Support Vector Machine (SVM) classifier. The parameters considered were:

- Kernels: Linear, Radial Basis Function (RBF), and Polynomial
- C values: 1 and 10 (the regularization parameter controlling the trade-off between maximizing the margin and minimizing classification errors)
- Polynomial degrees: 2, 3, and 4 (tested only when the kernel was polynomial)

V. RESULTS

We evaluated the effect of different kernels, regularization values (C), and polynomial degrees on Support Vector Machine accuracy. Table I summarizes the performance across all tested configurations, using the first 2000 rows of the dataset.

Table 1. Accuracy

Kernel	C	Degree	Accuracy
Linear	1	–	0.789
Linear	10	–	0.789
RBF	1	–	0.789
RBF	10	–	0.734
Poly	1	2	0.789
Poly	1	3	0.789
Poly	1	4	0.776
Poly	10	2	0.789
Poly	10	3	0.789
Poly	10	4	0.718

We evaluated multiple SVM configurations across different kernels, (C) values, and polynomial degrees using the first 2000 rows of the dataset. Several setups including the linear kernel, the RBF kernel with (C=1), and polynomial kernels with degrees 2–3 at both (C) and (C) achieved the highest accuracy of 78.9%. In contrast, the RBF kernel with (C=10) and the polynomial kernel with degree 4 produced lower accuracies 73.4% and 71.8%, respectively. These results indicate that simpler kernel choices were just as effective as more complex configurations, highlighting the stability of SVM performance on this dataset.

VI. CONCLUSION

This study applied Support Vector Machine models with different kernel and hyperparameter settings to predict student addiction. Using the first 2000 rows of the dataset, we found that several configurations including the linear kernel, the RBF kernel with (C=1), and polynomial kernels with degrees 2–3 achieved the highest accuracy of 78.9%. More complex setups, such as the RBF kernel with (C=10) and the polynomial kernel with degree 4, yielded lower accuracies, indicating that additional complexity did not improve performance. These results suggest that simpler kernel configurations provide stable and effective predictive accuracy for this dataset, underscoring the robustness of SVM for student health analytics.

REFERENCES

[1] Substance Abuse & Mental Health Services Administration (US), & Office of the Surgeon General (US). (2016). Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health. US Department of Health and Human Services.

[2] Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995). https://doi.org/10.1007/BF00994018.

[3] Volkow, N. D., Fowler, J. S., Wang, G. J., Baler, R., Telang, F. (2009). Imaging dopamine's role in drug abuse and addiction. Neuropharmacology, 56 Suppl 1(Suppl 1), 3–8. https://doi.org/10.1016/j.neuropharm.2008.05.022.

[4] Bellaj, Mohamed & Bendahmane, Ahmed & Boudra, Said & Lamarti Sefian, Mohammed. (2024). Educational Data Mining: Employing Machine Learning Techniques and Hyperparameter Optimization to Improve Students' Academic Performance. International Journal of Online and Biomedical Engineering (iJOE). 20. 55-74. 10.3991/ijoe.v20i03.46287.