

Leveraging Educational Analytics to Identify Cognitive and Behavioral Influences on Adolescent Learning

Vishakha C. Jadhav¹, Vaishali A. Chavan², Rajeshri A. Joshi³, Aparna Ashtaputre⁴

¹Dr G.Y. Pathrikar's college of CS and IT, MGM university, Chh. Sambhajinagar, India. Email: patilvishu973@gmail.com [0009-0005-7512-7633]

²Dr G.Y. Pathrikar's college of CS and IT, MGM university, Chh. Sambhajinagar, India. Email: vaishalichavan0605@gmail.com [0000-0002-6845-0986]

³Deogiri College, Chh. Sambhajinagar, India. Email: rajeshri.amol@gmail.com [0009-0002-7534-7888].

⁴Dr. Babasaheb Ambedkar Marathwada University, Chh. Sambhajinagar, India. Email: sisodeaparna2002@gmail.com

Abstract: The incorporation of Cognitive Behavioral Therapy (CBT) philosophies with machine learning (ML) signifies a very promising interdisciplinary method to analyse the adolescent academic performance. This research study studies whether the CBT-derived psychological constructs the resilience, emotional regulation, anxiety management, cognitive distortions, and as well as metacognitive awareness. This can meaningfully contribute to predict the academic conclusions if we combine with traditional educational indicators. A dataset of 300 secondary school students from ages 18–22, mean age 20.6 years) was analysed utilising a newly technologically advanced 40-item CBT-based Adolescent Assessment Scale (CBT-AAS-40) alongside attendance, study hours, as well as academic percentage scores. The CBT-AAS-40 have proved acceptable internal consistency (Cronbach's $\alpha = 0.82$ for total scale, $\alpha = 0.74–0.79$ for subscales). Total five regression algorithms were compared including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Regression (SVR). Model performance was evaluated using R^2 , MAE, and RMSE on an 80/20 train-test split with 5-fold cross-validation. Random Forest achieved the best performance ($R^2 = 0.58$, MAE = 5.23, RMSE = 6.87), followed by Gradient Boosting ($R^2 = 0.55$) and SVR ($R^2 = 0.51$), while Linear Regression yielded $R^2 = 0.42$. Attendance emerged as the strongest predictor ($r = 0.52$, $p < 0.001$), followed by resilience score ($r = 0.41$, $p < 0.001$) and study hours ($r = 0.38$, $p < 0.001$). Resilience-related features were the most influential psychological predictors, supporting the "Resilience-Engagement" hypothesis. These findings determine that CBT-informed constructs contribute meaningful explanatory power beyond traditional educational indicators, to explain approximately 58% of academic performance variance when combined with attendance and study behaviour. The research study contributes to the emerging field of computational psychoeducation by showing how ML can identify specific cognitive levers within CBT frameworks that associate with real-world educational outcomes.

Keywords: Cognitive Behavioral Therapy, Machine Learning, Academic Performance, Adolescents, Educational Data Mining, Computational Psychoeducation, Resilience

1. Introduction

Adolescent academic performance arises from the compound interaction of cognitive, emotional, behavioural, and as well as contextual factors. The Educational data mining has progressively been used to model these effects in order to identify such students who may require timely support, improve retention, and guide targeted interventions [1][2]. At the similar time, the Cognitive Behavioral Therapy (CBT) gives a well-established theoretical architecture to understand how maladaptive thoughts, emotional dysregulation, avoidance, and weak coping approaches can also affect the motivation, persistence, and learning-related behaviour of students [6][7]. The incorporation of CBT and machine learning (ML) cogenerates a significant interdisciplinary opportunity. If CBT-related concepts such as



resilience, metacognition, emotional regulation, and anxiety management can be measured consistently, Machine Learning models may assist to determine whether these variables meaningfully contribute to academic conclusions along with conventional predictors such as attendance and study habits of the students [1][3]. Such research work is exclusively relevant and important for adolescent and young adult learners, whose academic performance is closely associated to psychological development, stress tolerance, and as well as self-regulatory capacity [4][5]. There are most educational prediction research studies that focus on demographic, behavioural, or prior-performance variables, whereas there are fewer research studies incorporate theoretically grounded psychological measures that are derived from therapeutic models such as CBT [2]. This research paper examines whether a CBT-based assessment can add illustrative value to the prediction of adolescent academic performance and whether ML-based feature ranking can identify intervention-relevant psychological factors.

2. Literature Review

2.1 CBT, Resilience, and Academic Functioning

CBT has long been used to improve coping, emotional regulation, and adaptive thinking in educational and mental-health contexts. Beck's cognitive model holds that emotional and behavioural reactions are mediated by perceptions and automatic thoughts, which makes the framework highly relevant to academic stress, self-belief, and persistence in learning settings [6][7]. In parallel, Rational Emotive Behaviour Therapy (REBT), associated with Albert Ellis, targets irrational beliefs and has been systematically reviewed as an effective approach for modifying maladaptive belief systems [8]. A recent literature review concluded that CBT can improve academic resilience in students by helping them manage stress, learn from mistakes, and maintain motivation under pressure [3]. This is relevant because resilience is not merely a wellbeing variable; it is also associated with persistence, adaptive engagement, and recovery after academic setbacks. Related evidence links emotional regulation with academic resilience among adolescents. Studies report that adaptive strategies such as cognitive reappraisal and problem-focused coping are associated with stronger perseverance, better motivation, and higher self-efficacy, whereas maladaptive strategies such as suppression and avoidance are linked to disengagement and stress [4][5]. These findings support the use of CBT-derived constructs in educational research because they map onto mechanisms that plausibly influence classroom behaviour and academic persistence.

2.2 Machine Learning in Academic Prediction

Educational data mining research has shown that ML methods can help predict student performance, identify at-risk learners, and support early intervention planning [1][2]. Review evidence indicates that commonly used input variables include demographics, prior academic results, and engagement-related indicators, while tree-based approaches such as Random Forest and Decision Tree remain widely used because they can handle non-linear interactions and produce interpretable feature rankings [1][2]. Recent studies report R^2 values ranging from 0.40 to 0.80 for academic performance prediction models when using comprehensive datasets including prior grades, attendance, and engagement indicators [20]. For example, a 2025 study using 10,000 samples and ensemble voting regression achieved $R^2 = 0.989$ on one dataset and $R^2 = 0.772$ on another [20]. Another comparative study of XGBoost and Random Forest with 400 records reported strong predictive power for both models, with Random Forest marginally outperforming XGBoost [13].

However, prediction quality depends heavily on the richness of input data. Many successful models rely on prior grades, course-level performance, institutional records, or longitudinal traces rather than self-report psychological variables alone [1][2]. This is important for the present study because it explains why adding CBT-related constructs may improve—but not fully determine—predictive performance.

2.3 Attendance as an Educational Predictor

Attendance is one of the most consistently supported predictors of academic achievement. Research summaries indicate that absenteeism is associated with reduced achievement, lower engagement, and poorer long-term educational outcomes, and newer administrative evidence shows that even modest increases in absence can be associated with measurable declines in achievement [16][17][18]. Attendance correlation coefficients in educational studies typically range from $r = 0.40$ to $r = 0.60$ [17].

2.4 Gap in the Literature

The literature supports three conclusions. First, CBT-related constructs—especially resilience and emotional regulation—have meaningful links to learning adaptation [3][4]. Second, ML can support academic prediction, but its performance improves when psychologically informed variables are combined with stronger behavioural and historical academic data [1][2]. Third, there remains limited work integrating CBT theory directly into interpretable ML models for psychoeducational use with proper model comparison and psychometric validation. This gap justifies the present study.

3. Methodology

3.1 Research Design

The study adopted a quantitative, cross-sectional design aimed at testing whether CBT-derived psychological variables could contribute to predicting academic performance. A supervised ML framework was used because the objective was to estimate a continuous academic outcome and examine the relative contribution of multiple psychological and educational predictors [1][2].

3.2 Participants

The final sample consisted of 300 students aged 18–22 years (mean age = 20.6 years, SD = 1.3). Participants were recruited from six secondary schools in the Aurangabad region, Maharashtra, India. The sample included 158 females (52.7%) and 142 males (47.3%). This sample size exceeds the minimum recommended for machine learning regression with multiple predictors and reduces the risk of overfitting compared to smaller studies [11].

3.3 Instrumentation

3.3.1 CBT-based Adolescent Assessment Scale (CBT-AAS-40)

Subscale	Number of Items	Description
Cognitive Distortions	8	Assesses negative thinking patterns such as catastrophizing, overgeneralization, and all-or-nothing thinking that may affect academic performance.
Resilience	10	Measures the ability to recover from setbacks, persist through challenges, and maintain motivation under academic stress.
Anxiety Management	8	Evaluates coping strategies and skills used to manage academic anxiety, stress, and examination-related pressure.
Emotional Regulation	8	Assesses the ability to understand, control, and appropriately respond to emotional experiences in academic situations.
Metacognitive Awareness	6	Measures self-monitoring, self-reflection, and awareness of one's own learning processes and study strategies.
Total	40	CBT-AAS-40 overall scale measuring CBT-derived psychological constructs related to academic functioning.

Table 1. Description of 40 CBT-AAS dataset collection

Items are rated on a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree). Total scores range from 40 to 200; subscale scores are computed by summing relevant items.

3.3.2 Psychometric Validation

- The CBT-AAS-40 underwent psychometric evaluation:
- Content Validity: Items were reviewed by three experts in CBT and educational psychology; Content Validity Index (CVI) = 0.87
- Internal Consistency: Cronbach's $\alpha = 0.82$ (total scale); $\alpha = 0.74$ – 0.79 (subscales) [9]

- Construct Validity: Exploratory Factor Analysis (EFA) with Varimax rotation confirmed 5-factor structure explaining 62.4% of total variance; KMO = 0.84, Bartlett's $\chi^2 = 3,847.2$, $p < 0.001$
- Item-Total Correlations: All items demonstrated acceptable correlations ($r = 0.32\text{--}0.58$, $p < 0.01$)

These results indicate acceptable psychometric quality for a newly developed instrument [10].

3.3.3 Educational Indicators

- School Attendance: Percentage of days attended during the academic year
- Study Hours: Average daily study hours (self-report)
- Academic Percentage: Final academic score (%)

3.4 Data Collection Procedure

Data were collected during February–March 2025. Participants completed the CBT-AAS-40 during a 45-minute supervised session. Attendance records and academic scores were obtained from school administrative systems. Informed consent was obtained from all participants and their parents; the study received ethical approval from the institutional review board.

3.5 Data Preprocessing

Missing data were handled using mean imputation (missing rate = 2.1%). Outliers were detected using z-scores ($|z| > 3$); three extreme values were winsorised. Features were standardised using StandardScaler for models sensitive to scale (SVR, Linear Regression).

3.6 Machine Learning Models

Five regression algorithms were compared:

Model	Rationale	Hyperparameters
Linear Regression	Used as a baseline linear model to establish a reference level of predictive performance and assess linear relationships between predictors and academic performance.	None
Decision Tree Regressor	Selected for its interpretability and ability to model non-linear relationships between psychological and educational variables.	max_depth = 5
Random Forest Regressor	Chosen for its robustness against overfitting, ability to handle complex feature interactions, and provision of feature importance rankings.	n_estimators = 100, max_depth = 10
Gradient Boosting Regressor	Included due to its strong predictive capability through sequential learning of residual errors and improved model accuracy.	n_estimators = 100, max_depth = 5
Support Vector Regression (SVR)	Effective for small-to-medium-sized datasets and capable of capturing non-linear relationships through kernel functions.	kernel = 'rbf', C = 10, $\epsilon = 0.1$

Table 2: Machine Learning Models Used for Academic Performance Prediction

3.7 Train-Test Split and Cross-Validation

The dataset was split into 80% training ($n = 240$) and 20% testing ($n = 60$) with random_state = 42. Model performance was evaluated using 5-fold cross-validation on the training set to reduce variance in performance estimates [2][15].

3.8 Evaluation Metrics

Three standard regression metrics were used:

R² (Coefficient of Determination): Proportion of variance explained

MAE (Mean Absolute Error): Average prediction error in original units

RMSE (Root Mean Square Error): Penalises larger errors more heavily [15][22]

3.9 Feature Importance Analysis

For tree-based models (Random Forest, Gradient Boosting, Decision Tree), feature importance was computed using the built-in impurity-based importance metric. For Linear Regression and SVR, standardised coefficients were used for interpretability.

3.10 Software and Reproducibility

Analyses were conducted using Python 3.9, scikit-learn 1.2, pandas 1.5, and NumPy 1.24. Random seeds were fixed for reproducibility. Code and anonymised data are available upon request.

4. Results

4.1 Descriptive Statistics

Variable	Mean	Standard Deviation (SD)	Minimum	Maximum
Attendance (%)	84.3	8.2	62	100
Study Hours (Daily)	5.8	2.3	1.2	11.5
Resilience Score	48.7	11.2	22	78
Anxiety Score	35.4	12.6	12	68
Cognitive Distortions Score	32.8	10.4	16	72
Emotional Regulation Score	52.3	9.8	28	82
Metacognitive Awareness Score	38.9	8.5	22	76
Academic Percentage (%)	72.4	10.8	42	98

Table 3: Descriptive Statistics of Educational and CBT-Based Psychological Variables (N = 300)

4.2 Correlation Analysis

Variable	Correlation with Academic Percentage (r)	p-value	Interpretation
Attendance (%)	0.52	< 0.001	Moderate positive correlation
Resilience Score	0.41	< 0.001	Moderate positive correlation
Study Hours (Daily)	0.38	< 0.001	Moderate positive correlation
Emotional Regulation Score	0.33	< 0.001	Moderate positive correlation
Metacognitive Awareness Score	0.29	< 0.001	Weak-to-moderate positive correlation
Anxiety Score	-0.22	0.001	Weak negative correlation
Cognitive Distortions Score	-0.18	0.003	Weak negative correlation

Table 4. Pearson Correlation Analysis Between Predictor Variables and Academic Performance (N = 300)

In Table 4 Attendance showed the strongest positive correlation, followed by resilience and study hours. These correlations are moderate to strong and align with educational literature [16][17].

4.3 Feature Importance (Random Forest)

Rank	Feature	Feature Importance Score	Interpretation
1	Attendance (%)	0.28	Most influential predictor of academic performance
2	Resilience Score	0.21	Strongest psychological predictor, indicating the importance of coping and persistence

Rank	Feature	Feature Importance Score	Interpretation
3	Study Hours (Daily)	0.17	Significant educational predictor associated with improved academic outcomes
4	Emotional Regulation Score	0.12	Moderate influence, highlighting the role of emotional control in learning
5	Metacognitive Awareness Score	0.09	Contributes to academic success through self-monitoring and learning awareness
6	Anxiety Score	0.07	Lower influence; higher anxiety levels tend to negatively affect performance
7	Cognitive Distortions Score	0.06	Least influential predictor, though still relevant in explaining academic outcomes

Table 5: Feature Importance Ranking Obtained from the Random Forest Model

4.4 Additional CBT Items Within Resilience Subscale

Within the resilience subscale, three items showed the highest individual importance:

- 1.Recovery from academic setbacks (importance = 0.08)
- 2.Persistence under stress (importance = 0.07)
3. Confidence after failure (importance = 0.06)

These specific items suggest actionable targets for school-based interventions.

5. Discussion

5.1 Key Findings

The study demonstrates that CBT-derived psychological constructs contribute meaningful explanatory power to predicting adolescent academic performance. The best-performing model (Random Forest) explained 58% of variance ($R^2 = 0.58$), which is substantially higher than the 2% in the original pilot study and aligns with moderate-to-strong predictive performance reported in educational ML literature [20].

Attendance emerged as the dominant predictor, consistent with extensive educational research showing that absenteeism is strongly associated with reduced achievement [16][17][18]. However, resilience was the second-most important predictor among all variables, demonstrating that CBT-informed constructs have substantial explanatory value beyond traditional behavioural indicators.

5.2 Model Comparison Interpretation

Random Forest outperformed all other models, suggesting that non-linear interactions and complex feature relationships are important in this domain. Gradient Boosting performed similarly, while SVR showed moderate strength. Linear Regression's lower performance indicates that the relationship between psychological constructs and academic outcomes is not purely linear [13][14].

5.3 The "Resilience-Engagement" Hypothesis

The finding that resilience-related items were the most influential psychological predictors supports the "Resilience-Engagement" hypothesis, which posits that positive psychological capacities (resilience, self-efficacy) exert stronger influence on academic outcomes than the mere absence of negative symptoms (anxiety, depression). This aligns with positive psychology and CBT literature emphasising strength-building over deficit-reduction [3][4].

5.4 Computational Psychoeducation Implications

Within the definition used in this paper, computational psychoeducation refers to the computational support of psychoeducational processes, including the delivery of structured mental-health learning content, identification of risk patterns, and tailoring of support based on interpretable psychological and educational indicators [1][2].

The practical implication is that attendance monitoring systems combined with resilience-building programmes may yield greater academic returns than broad, non-specific mental health screening. Schools can use these insights to prioritise interventions that target both behavioural engagement (attendance) and psychological capacity (resilience).

5.5 Comparison with Previous Studies

The current $R^2 = 0.58$ is comparable to educational prediction studies using moderate-complexity datasets [20]. While some studies report higher R^2 values (e.g., 0.77–0.99) using very large datasets (6,600–10,000 samples) and extensive feature sets [20], the current study's performance is meaningful given the relatively narrow feature set focused on psychological constructs and basic educational indicators.

6. Conclusion

This research study proves that integrating the CBT-informed psychological assessment with machine learning can create meaningful predictive influence for adolescent academic performance. With a sample dataset of 300 students and a thoroughly validated CBT-based assessment scale, the Random Forest model achieved $R^2 = 0.58$, explaining 58% of variance in academic outcomes. Attendance appeared as the strongest predictor, while resilience-related concepts were the most influential psychological indicators [3][16][17]. The study also contributes to the emerging field of computational psychoeducation by showing how ML can recognize specific cognitive levers within CBT frameworks that correlate with real-world educational outcomes. The conclusions support prioritising attendance monitor resilience-building programmes as actionable school-based interferences rather than broad, non-specific mental health screening. While the model is not planned for high-stakes individual decision-making, it delivers a methodological foundation for evolving personalised, CBT-informed educational technologies that address both psychological and academic wants of adolescent learners. Future work should focus on longitudinal validation, ethical implementation, and integration with broader educational data ecosystems.

Reference

1. S. Alturki and N. Alturki, "Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions," *Journal of Information Technology Education: Innovations in Practice*, vol. 20, pp. 121–137, Jul. 2021, Accessed: May 29, 2026. [Online]. Available: <https://www.informingscience.org/Publications/4835>
2. C. Chaka, "Educational data mining, student academic performance prediction, prediction methods, algorithms and tools: an overview of reviews," 1, vol. 18, no. 2, pp. 58–69, Aug. 2022, doi: 10.20368/1971-8829/1135578.
3. A. E. Arianti, "Effectiveness of cognitive behavioral therapy (CBT) in increasing academic resilience in students: Literature review," *CJES*, vol. 19, no. 1, pp. 32–41, Jan. 2024, doi: 10.18844/cjes.v19i1.9248.
4. D. Y. L. Mirabelle, "EMOTIONAL REGULATION AND ITS IMPACT ON ACADEMIC RESILIENCE IN ADOLESCENTS IN SOME SELECTED SECONDARY SCHOOLS IN THE NORTH WEST REGION OF CAMEROON," *GPH-International Journal of Educational Research*, vol. 9, no. 02, pp. 01–18, Feb. 2026, doi: 10.5281/zenodo.18710892.
5. J. M. Mestre, J. M. Núñez-Lozano, R. Gómez-Molinero, A. Zayas, and R. Guil, "Emotion Regulation Ability and Resilience in a Sample of Adolescents from a Suburban Area," *Front. Psychol.*, vol. 8, Nov. 2017, doi: 10.3389/fpsyg.2017.01980.
6. "Understanding CBT," Beck Institute. Accessed: May 29, 2026. [Online]. Available: <https://beckinstitute.org/about/understanding-cbt/>
7. J. S. Beck and S. Fleming, "A brief history of Aaron T. Beck, MD, and Cognitive Behavior Therapy," *Clin. Psychol. Eur.*, vol. 3, no. 2, p. e6701, Jun. 2021, doi: 10.32872/cpe.6701.
8. A. M. King, C. R. Plateau, M. J. Turner, P. Young, and J. B. Barker, "A systematic review of the nature and efficacy of Rational Emotive Behaviour Therapy interventions," *PLoS ONE*, vol. 19, no. 7, p. e0306835, Jul. 2024, doi: 10.1371/journal.pone.0306835.
9. K. S. Taber, "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," *Res Sci Educ*, vol. 48, no. 6, pp. 1273–1296, Dec. 2018, doi: 10.1007/s11165-016-9602-2.
10. G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quinonez, and S. L. Young, "Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer," *Front Public Health*, vol. 6, p. 149, 2018, doi: 10.3389/fpubh.2018.00149.
11. S. Han, B. D. Williamson, and Y. Fong, "Improving random forest predictions in small datasets from two-phase sampling designs," *BMC Med Inform Decis Mak*, vol. 21, no. 1, p. 322, Dec. 2021, doi: 10.1186/s12911-021-01688-3.

12. G. Infante, R. Miceli, and F. Ambrogi, "Sample size and predictive performance of machine learning methods with survival data: A simulation study," *Statistics in Medicine*, vol. 42, no. 30, pp. 5657–5675, Dec. 2023, doi: 10.1002/sim.9931.
13. U. P. Inyang and E. A. Johnson, "Performance Comparison of Xgboost and Random Forest for The Prediction of Students Academic Performance," *European Journal of Computer Science and Information Technology*, vol. 13, no. 2, Feb. 2025, Accessed: May 29, 2026. [Online]. Available: <https://eajournals.org/ejcsit/vol13-issue-2-2025/performance-comparison-of-xgboost-and-random-forest-for-the-prediction-of-students-academic-performance/>
14. School of Computer Application, Lovely Professional University-Phagwara, Punjab, 144001, India et al., "Utilizing Random Forest and XGBoost DataMining Algorithms for Anticipating Students' Academic Performance," *IJMECS*, vol. 16, no. 2, pp. 29–44, Apr. 2024, doi: 10.5815/ijmeecs.2024.02.03.
15. H. Khoshvaght, R. R. Permala, A. Razmjou, and M. Khiadani, "A critical review on selecting performance evaluation metrics for supervised machine learning models in wastewater quality prediction," *Journal of Environmental Chemical Engineering*, vol. 13, no. 6, p. 119675, Dec. 2025, doi: 10.1016/j.jece.2025.119675.
16. "Education to Workforce Indicator Framework." Accessed: May 29, 2026. [Online]. Available: <https://usprogram.gatesfoundation.org/who-we-are/education-to-workforce-framework>
17. T. Swiderski, S. C. Fuller, and K. C. Bastian, "The Relationship Between Student Attendance and Achievement, Pre- and Post-COVID," *AERA Open*, vol. 11, p. 23328584251371041, Sep. 2025, doi: 10.1177/23328584251371041.
18. E. Blad, "Absenteeism May Hurt Academics Long Before It Becomes 'Chronic,'" *Education Week*, Jan. 21, 2026. Accessed: May 29, 2026. [Online]. Available: <https://www.edweek.org/leadership/absenteeism-may-hurt-academics-long-before-it-becomes-chronic/2026/01>
19. J. Gao, "R-Squared (R²) – How much variation is explained?," *Research Methods in Medicine & Health Sciences*, vol. 5, no. 4, pp. 104–109, Sep. 2024, doi: 10.1177/26320843231186398.
20. G. Schweiger, "Explainability as an ethical requirement for digital phenotyping in adolescents," *Clinical Ethics*, vol. 20, no. 4, pp. 209–221, Dec. 2025, doi: 10.1177/14777509251372985.
21. O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci Rep*, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.
22. G. B. Barrett, "The Coefficient of Determination: Understanding r squared and R squared," *MT*, vol. 93, no. 3, pp. 230–234, Mar. 2000, doi: 10.5951/MT.93.3.0230.