International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.7 November - December – 2023
DOI:10.46647/ijetms.2023.v07i06.003 ISSN: 2581-4621

A Machine Learning Approach to Predicting Academic Performance

S Sukanya[1], Dr D William Albert[2]
[1]M.Tech Student, Dept. of CSE, Ashoka Women's Engineering College, Kurnool, A.P
[2]Professor, Dept. of CSE, Ashoka Women's Engineering College, Kurnool, A.P

Abstract— Academic performance prediction is an indispensable task for policymakers. Academic performance is frequently examined using classical statistical software, which can be used to detect logical connections between socioeconomic status and academic performance. These connections, whose accuracy depends on determine prediction accuracy. To eliminate the effects of logical relationships on such accuracy, machine learning models extended with education and socioeconomic data to predict academic performance. The decision tree, random forest, logistic regression, support vector machine, and neural network are used for testing. The neural network model can be used by policymakers to forecast academic performance, which in turn can aid in the formulation of various policies, such as those regarding funding and teacher selection. Finally, this study demonstrated the feasibility of machine learning as an auxiliary educational decision-making tool for use in the future.

Keywords—Machine Learning, Socio Economic Status, Students Performance.

I. INTRODUCTION

Educational Systems have widely utilized standardized examinations as a large-scale means of effectively sorting students. When it comes to evaluation efficiency, standardized test scores are overwhelmingly superior in identifying talent over other qualities that schools ought to place greater emphasis on, such as moral character, life adaptability, non-cognitive skills, and social responsibility. These preferences are rooted in the strengths of standardized tests, which are a product of historical and social conventions. There are considerable and obvious advantages to employing paper-and-pencil examinations that feature a series of archetypal questions, including practicality, reliability, good content validity, convenience, accessibility, and openness.

Despite the usefulness of traditional tests in assessing students' knowledge and skills, there are several other factors that can impact academic performance, often overlooked. One significant factor identified in predictive studies is socioeconomic status (SES), which plays a vital role in widening the academic performance gap between students in rural and urban institutions. In sub continent, high SES often correlates with above-average exam scores, highlighting the significant impact of SES on educational performance. Many researchers and educators continue to explore the effects of SES on academic performance using correlation and regression analysis. As such, this study aims to employ machine learning (ML) models, a novel approach in predictive studies, to investigate the impact of SES on student academic performance. Numerous studies have demonstrated the considerable accuracy of ML prediction compared to a classic statistical method such as correlation and linear. As an artificial intelligence approach, ML has had a far-reaching influence on handling the vast amounts of facts and numerical data generated by computers through simulations of the human brain. For instance, an ML algorithm is superior in analyzing considerable internet data than regular models, since it enables relatively rapid prediction with high accuracy and large datasets. Applying ML algorithms also enables researchers and teachers to recognize the key factors that strongly influence student performance and find more effective ways to improve teaching quality. The problem is that previous studies were a small-scale, incomprehensive and restricted data pool to address certain groups under limited conditions. This scope cannot ensure overall effective outcomes of ML prediction and a large representative sample has yet to be used to further verify the precision of ML results.

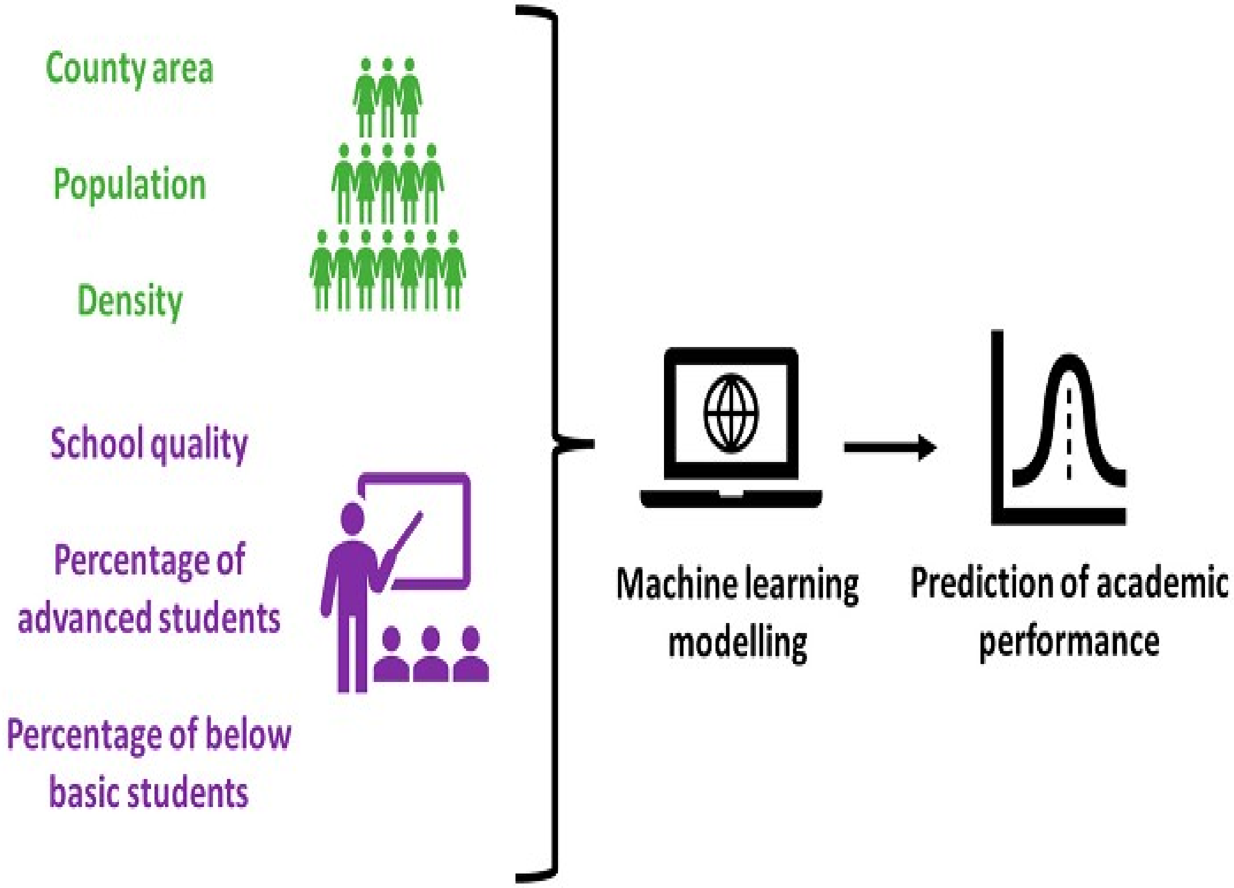

County area
Population
Density
School quality
Percentage of advanced students
Percentage of below basic students
Machine learning modelling
Prediction of academic performance

Figure. *The Diagram shows how ML used big data to predict academic performance.*

II.SYSTEM ANALYSIS

Data Description

Using population and school data, this study trained five ML models: a decision tree, random forest, logistic regression, support vector machine, and neural networks.

County Area (km^2)	Approximate Number	County Population (People)	Approximate Number
≤1000	500	≤50,000	25,000
(1000,2000]	1500	(50,000,200,000]	100,000
(2000,3000]	2500	(200,000,1,000,000]	500,000
(3000,4040]	3500	(1,000,000, 1,500,000	1,500,000

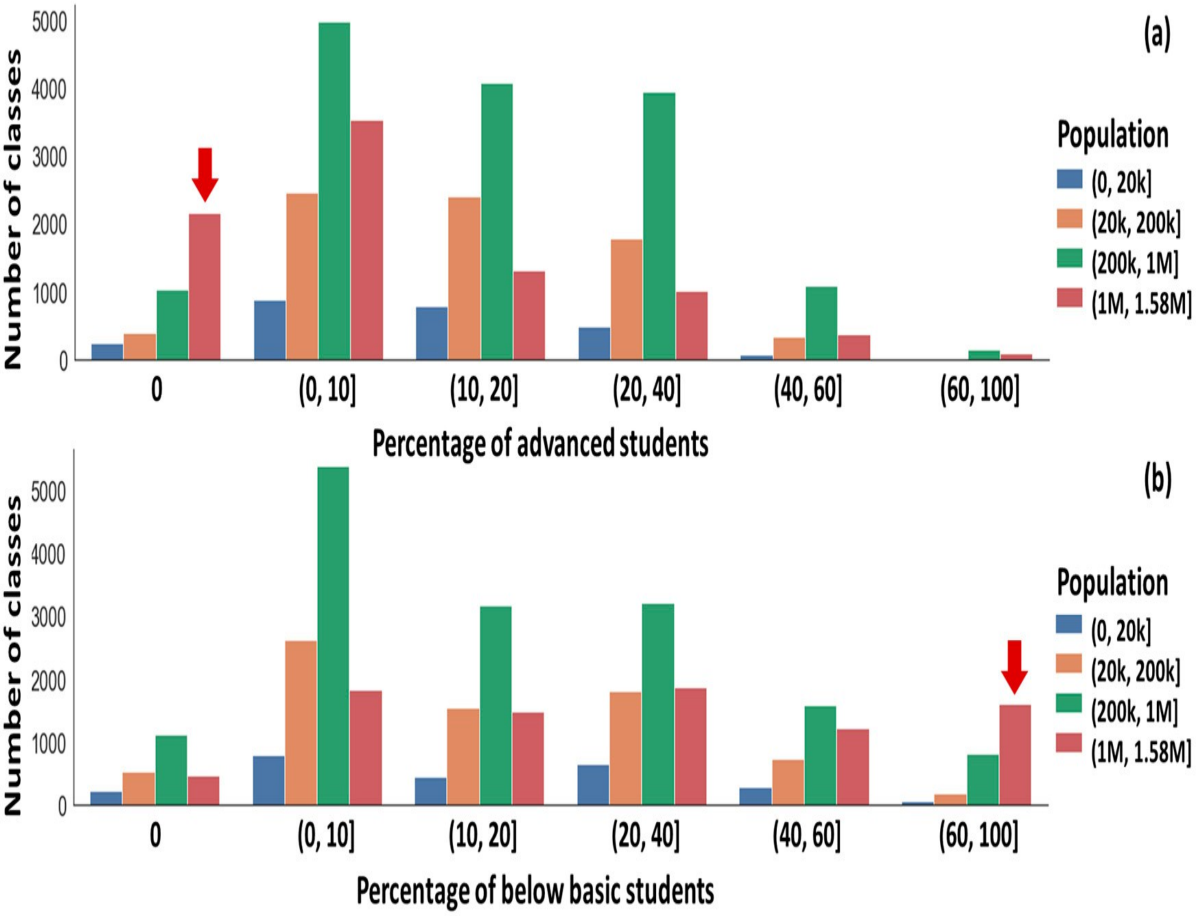
Populatio n Density (people/km2)	Approximate Number	Percentage of Advanced/Below-Basic Students	Approximate Number
≤100	50	0%	0%
(100,500]	300	(0%,10%]	5%
(500,1300]	900	(10%,20%]	15%

(1300,4564]	3000	(20%,40%]	30%

When we evaluated how the county population affects academic performance, we found the large county population (1M~1.58M, red box and red arrow in Figure) lead to a lower percentage of advanced students and a higher percentage of below-basic students. The larger county population suggested a lower academic performance. On the contrary, a smaller county population helped to improve students' overall academic performance.

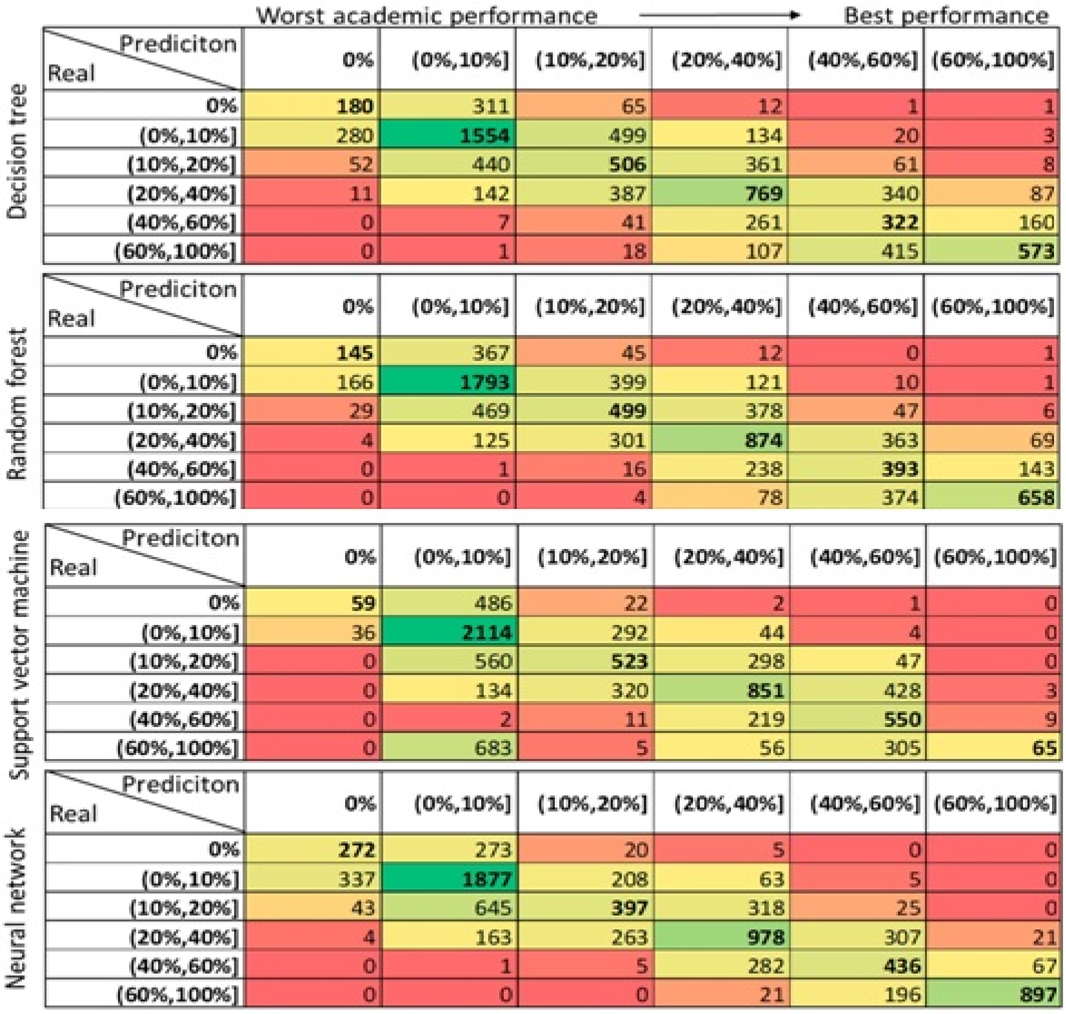
III.SYSTEM CONSTRUCTION

Simulation results- Final Students Performance Model



We evaluated four ML prediction methods, including decision tree, random forest, support vector machine, and neural network. Among all methods, the decision tree, random forest, logistic regression, and support vector machine achieved testing accuracy of 48%, 54% and 51%, respectively (Table). The neural network achieved the highest 60% testing accuracy. As a result, this paper utilized the neural network method for the next step of the analysis.
When we applied the ML models, the prediction versus reality results were shown in Figure

Method	Classifier	Training Accuracy	Testing Accuracy
Decision tree	Decision Tree Classifier	94%	48%
Random forest	Random Forest Classifier	94%	54%
Support vector machine	Support Vector Classifier	59%	51%
Neural network	MLP Classifier	61%	60%



Worst academic performance → Best performance

Decision tree

Prediciton Real	0%	(0%,10%]	(10%,20%]	(20%,40%]	(40%,60%]	(60%,100%]
0%	**180**	311	65	12	1	1
(0%,10%]	280	**1554**	499	134	20	3
(10%,20%]	52	440	**506**	361	61	8
(20%,40%]	11	142	387	**769**	340	87
(40%,60%]	0	7	41	261	**322**	160
(60%,100%]	0	1	18	107	415	**573**

Random forest

Prediciton Real	0%	(0%,10%]	(10%,20%]	(20%,40%]	(40%,60%]	(60%,100%]
0%	**145**	367	45	12	0	1
(0%,10%]	166	**1793**	399	121	10	1
(10%,20%]	29	469	**499**	378	47	6
(20%,40%]	4	125	301	**874**	363	69
(40%,60%]	0	1	16	238	**393**	143
(60%,100%]	0	0	4	78	374	**658**

Support vector machine

Prediciton Real	0%	(0%,10%]	(10%,20%]	(20%,40%]	(40%,60%]	(60%,100%]
0%	**59**	486	22	2	1	0
(0%,10%]	36	**2114**	292	44	4	0
(10%,20%]	0	560	**523**	298	47	0
(20%,40%]	0	134	320	**851**	428	3
(40%,60%]	0	2	11	219	**550**	9
(60%,100%]	0	683	5	56	305	**65**

Neural network

Prediciton Real	0%	(0%,10%]	(10%,20%]	(20%,40%]	(40%,60%]	(60%,100%]
0%	**272**	273	20	5	0	0
(0%,10%]	337	**1877**	208	63	5	0
(10%,20%]	43	645	**397**	318	25	0
(20%,40%]	4	163	263	**978**	307	21
(40%,60%]	0	1	5	282	**436**	67
(60%,100%]	0	0	0	21	196	**897**

Figure. Percentage of advanced students by ML models (Prediction versus reality).

IV.CONCLUSION

On the basis of big data this research demonstrated the feasibility of using ML models to predict class academic performance. To this we used an ML model that achieves fast and precise predictions. This study confirmed that ML models are accurate and effective instruments. With ML models as grounding, we found that well-educated people in small counties that have lower income could contribute to higher academic performance. Finally, SES exerts a significant impact on the rural–urban performance gap. The ML models are expected to provide assistance and guidance (e.g., decision making on issues that may affect performance, such as education budgets, hiring standards and practices, and teacher–student ratios) to education policymakers in the region in the future.

V.REFERENCES

1. J. Xu, K. H. Moon, and M. Van Der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," IEEE J. Sel. Top. Signal Process., vol. 11, no. 5, pp. 742–753, 2017.
2. K. P. Shaleena and S. Paul, "Data mining techniques for predicting student performance," in ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology, 2015, no. March, pp. 0–2.
3. A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," in Procedia Computer Science, 2015.
4. Y. Meier, J. Xu, O. Atan, and M. Van Der Schaar, "Predicting grades," IEEE Trans. Signal Process. vol. 64, no. 4, pp. 959–972, 2016.

5. P. Guleria, N. Thakur, and M. Sood, “Predicting student performance using decision tree classifiers and information gain,” Proc. 2014 3rd Int. Conf. Parallel, Distrib. Grid Comput. PDGC 2014, pp. 126–129, 2015.
6. P. M. Arsad, N. Buniyamin, and J. L. A. Manan, “A neural network students’ performance prediction model (NNSPPM),” 2013 IEEE Int. Conf. Smart Instrumentation, Meas. Appl. ICSIMA 2013, no. July 2006, pp. 26–27, 2013.
7. K. F. Li, D. Rusk, and F. Song, “Predicting student academic performance,” Proc. - 2013 7th Int.Conf. Complex, Intelligent, Softw. Intensive Syst. CISIS 2013, pp. 27–33, 2013.
8. G. Gray, C. McGuinness, and P. Owende, “An application of classification models to predict learner progression in tertiary education,” in Souvenir of the 2014 IEEE International Advance Computing Conference, IACC 2014, 2014.
9. N. Buniyamin, U. Bin Mat, and P. M. Arshad, “Educational data mining for prediction and classification of engineering students achievement,” 2015 IEEE 7th Int. Conf. Eng. Educ. ICEED 2015, pp. 49–53, 2016.
10. Z. Alharbi, J. . Cornford, L. . Dolder, and B. . De La Iglesia, “Using data mining techniques to predict students at risk of poor performance,” Proc. 2016 SAI Comput. Conf. SAI 2016, pp. 523–531, 2016.
11. B. Hore, s. Mehrotra, m. Canim, and m. Kantarcioglu, **“**secure multidimensional Range queries over outsourced data,**”** *vldb j.*, vol. 21, no. 3,Pp. 333–358, 2012.
12. J. Mullan, “Learning Analytics in Higher Education,” London, 2016.
13. P. and K. Al-Shabandar, R., Hussain, A.J., Liatsis, “Detecting At-Risk Students With Early Interventions Using Machine Learning Techniques,” *IEEE Access*, vol. 7, pp. 149464–149478, 2019.
14. S. Jiang, A. E. Williams, K. Schenke, M. Warschauer, and D. O. Dowd, “Predicting MOOC Performance with Week 1 Behavior,” in *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*, 2014, pp. 273–275.
15. L. Analytics and C. Exchange, “OU Analyse : Analysing at - risk students at The Open University,” in *in Conference, 5th International Learning Analytics and Knowledge (LAK) (ed.)*, 2015, no. October 2014.
16. R. Alshabandar, A. Hussain, R. Keight, A. Laws, and T. Baker, “The Application of Gaussian Mixture Models for the Identification of At-Risk Learners in Massive Open Online Courses,” in *2018 IEEE Congress on Evolutionary Computation, CEC 2018 - Proceedings*, 2018.
17. J.-L. Hung, M. C. Wang, S. Wang, M. Abdelrasoul, Y. Li, and W. He, “Identifying At-Risk Students for Early Interventions—A Time-Series Clustering Approach,” *IEEE Trans. Emerg. Top. Comput.*, vol. 5, no. 1, pp. 45–55, 2017.
18. C. Yun, D. Shin, H. Jo, J. Yang, and S. Kim, “An Experimental Study on Feature Subset Selection Methods,” *7th IEEE Int. Conf. Comput. Inf. Technol. (CIT 2007)*, pp. 77–82, 2007.
19. G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.