



Data Analytics on a Higher Education Dataset Using Predictive Analytics

D. Gandhimathi¹ and Jagadeeswari J. M²

¹Department of CSE – Data Science, Dr. ACS College of Engineering, Bengaluru, India

²M.Sc. Data Science, Annamalai University, Chidambaram, India

Abstract— Predictive analytics, a branch of data analytics, combines historical data with statistical algorithms and modelling techniques to anticipate future outcomes, while data analytics more broadly turns raw data into insight that supports decisions. This paper applies predictive analytics to higher education in order to identify learners' interests, behaviour patterns and cognitive level across different aspects of the curriculum. Today's learners acquire knowledge not only in the classroom but also through tutorials, online courses and other digital sources. Drawing on institutional records and a survey of undergraduate learners at a reputed institution, the study examines how classroom learning and digital learning together contribute to academic outcomes. Academic performance and additional skills gained from other sources are summarized, a predictive-analytics workflow comprising model training, validation and comparison with related machine-learning models is set out, and correlation analysis is used to examine the relationship between academic performance and additional skill acquisition. Internal marks, attendance, Learning Management System (LMS) activity and a Skill Development Index all correlate positively with GPA ($r = 0.58-0.74$). On this basis the paper argues that combining academic and skill-development data allows institutions to identify proactive learners and to direct support where it is most needed.

Keywords— Predictive analytics, data analytics, correlation analysis, higher education, learning analytics, skill development

I. INTRODUCTION

A higher education system that aims to support learners proactively must monitor each learner's performance on several fronts and distinguish between different kinds of learner. Classroom performance alone cannot reveal how far a learner is progressing, because learners in the digital era draw on many other sources to extend their knowledge. Quantitative analysis grounded in descriptive statistics offers a structured way to evaluate and summarize such data and so to gain insight into both the current performance and the potential of learners.

Descriptive statistics, with its emphasis on central tendency and variability, helps to reveal patterns, trends and inconsistencies in educational datasets, and it can be used to characterize the academic performance of individual learners. In well-established institutions, academic performance is also analysed against Bloom's taxonomy to determine the cognitive level that learners have reached, and statistical methods have been used to set and track learning-outcome targets within outcome-based education [6]. Predictive analytics goes a step further. Rather than describing the past, it applies statistical

modelling, data mining and machine learning to historical data in order to detect patterns and trends and to forecast outcomes. Its aim is to move from knowing what has happened to the best available assessment of what is likely to happen in the area of interest, and organizations and institutions increasingly adopt it to cope with heightened competition in the digital era.

Predictive analytics is most familiar from business settings. In fraud detection, combining several analytic methods sharpens pattern recognition and helps to expose criminal behaviour; in operations, companies use it to forecast inventory and manage resources; and in marketing it models customer purchases and responses to identify cross-selling opportunities. Its use in education is less developed. This paper examines how predictive analytics can serve the education sector in the Industry 5.0 era by identifying proactive learners in higher education institutions. The study extracts learners' academic performance together with the additional skills they have acquired, analyses these data in R, and correlates the two in order to characterize different levels of learner improvement and to predict which learners are likely to perform proactively. The underlying argument is that academic records and evidence of self-directed skill development, read together, describe a learner more completely than either does alone.

II. LITERATURE REVIEW

This section reviews recent research on the use of predictive analytics and artificial intelligence (AI) in higher education, with particular attention to the prediction of student performance and to retention strategies.

A. AI and Machine Learning in Higher Education

Deri et al. [1] provide a comprehensive overview of AI in higher education institutions and emphasize its considerable potential. They show that AI can adapt learning materials to the needs and styles of individual learners and thereby personalize education for a diverse student body. On the administrative side, they note that automating grading, admissions, scheduling and resource allocation reduces costs and improves institutional decision-making. They also identify serious ethical concerns, namely data privacy, algorithmic bias and digital equity, which institutions must address when adopting AI.

Vieriu and Petrea [2] weigh the benefits of AI in education against its challenges. They describe how intelligent tutoring systems, adaptive learning platforms and AI-powered virtual assistants are reshaping teaching, but warn of risks that include



weaker critical thinking, cognitive disengagement among students and a range of ethical problems. Importantly, they observe that established learning theories such as constructivism and socio-cultural approaches have received little attention in studies of AI integration, leaving open the question of how AI tools fit within existing educational frameworks.

B. Cognitive Impact and Learning Performance

Liu et al. [3] studied how AI-based learning systems affect students' cognitive abilities and academic achievement, asking whether AI-supported training improves learning performance by strengthening cognitive function. They present evidence of such cognitive benefits and regard AI tools, including machine learning, natural language processing and computer vision, as a transformative force capable of personalizing learning. Their work builds on earlier findings that cognitive skills such as memory, attention and reasoning bear directly on the academic performance of students in higher education.

C. Predictive Analytics and Student Success

Predictive analytics in higher education spans a range of methods for forecasting student success and identifying at-risk groups. Predictive models have been shown to uncover patterns in historical student data that are associated with academic success or failure, and early-warning systems built on such models have been reported to improve retention and academic performance substantially. Models that combine several data sources, such as academic performance indicators, engagement metrics and demographic factors, predict more accurately than models based on a single variable, and they allow institutions to design interventions tailored to specific student needs. The present study adopts this multi-source view, but adds a dimension that the reviewed work largely leaves aside: the skills learners acquire on their own initiative outside the formal curriculum.

III. DATA ANALYTICS AND PREDICTIVE ANALYTICS

A. Data Analysis and Data Analytics

Data analytics is the overall process of collecting and transforming data and organizing it in order to draw conclusions, make predictions and extract information for better decisions. Data analysis, the extraction of meaning from data, is one component of this process; data analytics goes further, turning raw data into detailed, actionable insight and shaping it to support decisions about future growth. Its principal forms include statistical analysis, text analysis, diagnostic analytics, predictive analytics and prescriptive analytics. This paper is concerned with predictive analytics and its application in education.

B. Predictive Analytics

Predictive analytics is a branch of advanced data analytics that combines historical data with statistical modelling to estimate future outcomes, using data-mining techniques to

prepare machine-learning models that support decisions. Advances in predictive modelling have made it a powerful tool in education for forecasting student outcomes and identifying opportunities for intervention. Machine-learning algorithms such as decision trees, random forests and neural networks have been reported to outperform traditional statistical methods in predicting student success and identifying at-risk populations. Chen and Zhang (2023) showed that ensemble methods, which combine several algorithms, predicted student retention with more than 85% accuracy and consistently outperformed individual algorithms. Deep-learning models applied to educational data have further revealed complex non-linear relationships between student characteristics and as many as three academic outcomes, relationships that conventional correlation analysis would not have detected.

A predictive-analytics project typically proceeds through five stages:

- 1) Problem definition.
- 2) Data identification.
- 3) Data preprocessing.
- 4) Predictive model development.
- 5) Validation and recommendation for decision-making.

Applied to educational data, this process can reveal each learner's pattern of engagement in the first weeks of a semester, so that support can be directed to weaker learners in time to improve their performance in the following semester.

C. Correlation Analysis

Correlation analysis is a statistical technique for identifying relationships between variables, which helps to reveal patterns and trends in data and to anticipate future needs. The correlation coefficient quantifies the strength of the relationship between two variables: the larger its magnitude, the stronger the association. A positive correlation means that the two variables increase together; a negative correlation means that one increases as the other decreases; and a correlation of zero indicates no linear relationship, so that a change in one variable is not accompanied by a change in the other. In this study, correlation analysis is used to test the central hypothesis that academic performance is linked to engagement in skill development.

IV. PREDICTIVE ANALYTICS IN HIGHER EDUCATION: APPROACH

The study aims to identify proactive learners from their cognitive level across various aspects of the curriculum of the course they are taking. In practice, this is examined through learners' performance in a particular course. The dataset covers one batch of learners and their academic performance in one course over a semester. In addition, a questionnaire was circulated to capture their additional skills, defined here as knowledge enrichment beyond the curriculum, such as extra online or offline courses completed and internships attended.

The extracted data were cleaned and processed using descriptive statistics. Exploratory data analysis (EDA) is



central to this step: by investigating and summarizing the attributes of a dataset, supported by visualization, it exposes the structure, trends and anomalies in the data, highlights the features most relevant for modelling and indicates how they should be prepared. All analyses were carried out in R, an open-source statistical programming environment widely used across research fields for data-science tasks and visualization. The process comprises four stages:

- 1) Characterize learner behaviour, in academic work and in additional skill acquisition, with a view to improving performance.
- 2) Build a machine-learning model that categorizes patterns of learner behaviour.
- 3) Compare the model's performance with that of alternative models.
- 4) Identify the most suitable model for evaluating educational data in higher education.

Fig. 1 shows the resulting workflow, from raw data through preprocessing, model training, validation and evaluation to intervention, which an institution can implement in R to evaluate its own educational data.



Workflow for predictive analysis of the educational dataset.

V. METHODOLOGY

The analysis was designed to move from exploratory analysis through predictive-model evaluation to fairness checks and actionable interventions. Statistically, it progressed from description to inference: data cleaning and preprocessing, descriptive statistics, correlation analysis and visualization. The aim of this section is not only to report numerical results but to draw educational meaning from them, showing what predictive analytics can reveal about student behaviour, performance and learning paths.

A. Data Collection and Sample

Data were drawn from three sources. Academic records, comprising internal assessment results, external examination results, attendance rate and cumulative GPA, were obtained from the institution's Enterprise Resource Planning (ERP) system and Learning Management System (LMS). A behavioural engagement dataset was compiled from LMS interaction logs, including login frequency, resource access and on-time submission. Skill development was measured with a standard Likert-scale survey of learners' participation in online courses, internships, professional development programmes and certifications; the questionnaire was developed from a review of existing instruments and pilot-tested with a small group of students for clarity and completeness.

The target population comprised undergraduate learners enrolled in a Data Science programme at a reputed institution in southern Tamil Nadu during the 2024–2025 academic year, and one course of the programme was analysed. The accessible population consisted of 650 learners from various programme

levels with a wide range of academic backgrounds and skill-development experience. Purposive sampling selected 560 of them, ensuring representation across academic levels and categories of skill-development engagement. A power analysis assuming a medium effect size ($r = 0.30$), $\alpha = 0.05$ and power of 0.80 gave a minimum sample of 84; this was deliberately exceeded to allow for data-quality problems and to strengthen the analyses. Inclusion required enrolment in a target programme, completion of at least one full semester and complete academic records. Students with incomplete records, part-time enrolment or significant interruptions during the study period were excluded. The final sample was suitably diverse in demographic characteristics and levels of skill-development engagement.

B. Data Preprocessing

Before modelling, the data were preprocessed systematically to ensure quality and consistency. Missing values amounted to less than 5% for every variable and were handled with Multiple Imputation by Chained Equations (MICE); the proportion missing and the imputation method for each variable are listed in Table I.

MISSING DATA AND IMPUTATION STRATEGY

Variable	% Missing	Imputation Method
Attendance	2.0%	MICE (regression)
LMS Logins	1.5%	MICE (predictive mean)
Skill Survey	3.2%	MICE (ordinal regression)
Internal Marks	0.8%	Mean substitution

Outliers were screened with both the ± 3 SD rule and an Isolation Forest; genuine extreme performers were retained to preserve the shape of the distribution. Continuous features were standardized as z-scores, categorical variables were one-hot encoded and ordinal predictors such as skill-development scores were rank-encoded. Behavioural features derived from LMS logs were reduced by Principal Component Analysis (PCA) to remove noise and multicollinearity. Fig. 2 shows a sample of the dataset after cleaning.

How do you develop your additional mode of learning (Any online course)	Any other online course	Internship attended - yes / no
My working on the skill I'm trying to Youtube	Text Book (Physical Book/Swayam, Udemy, C No)	Yes
Great	E-Book, Tutorial, Youtube/Udemy, Coursera	None
Yes	E-Book, Tutorial, Youtube/Udemy	Infantry Springboard
I develop my skills by spending much	E-Book, Tutorial, Youtube/Udemy	Yes
Architecture help develop everything	Udemy	No
Reading researches	E-Book, Tutorial, Web/Udemy	NPTEL
Through college lab and online test	Text Book (Physical Book/Swayam, Udemy, C No)	Yes, working in an architect
Has	E-Book, Tutorial, Youtube/Udemy	Yes, art restoration at local
In my sites	Text Book (Physical Book/Swayam, Udemy, C No)	Visual rejection in manufa
Developing additional skills, where	E-Book, Tutorial, Youtube/Udemy, C No	Yes
Through online sources	Text Book (Physical Book/Swayam, Udemy, C No)	Yes
Online learning	Text Book (Physical Book/Swayam, Udemy, C No)	Yes
Let learning technical courses and try	E-Book, Tutorial, Web/Swayam, Udemy, C No	Yes - R&D
I learn in my free time and engage	E-Book, Tutorial, Youtube/Swayam, Coursera	Yes
By using online platforms like You	Text Book (Physical Book/Microsoft, IBM, TCI RA	Data visualization conduct

Sample of the dataset after cleaning.

C. Descriptive Statistics

Measures of central tendency and dispersion were computed for a sample of 100 drawn from the population of 650, using the marks of the 560 learners in the selected course imported into R. The results are summarized in Table II.



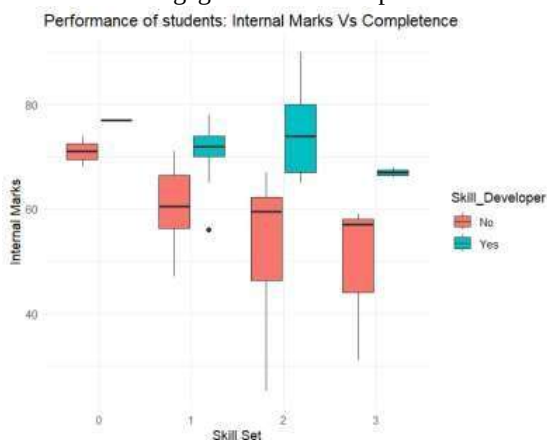
MEASURES OF CENTRAL TENDENCY AND DISPERSION

Measure	Value
Mean	85.4
Median	78.0
Mode	76.0
Range	60–100
Interquartile range (IQR)	25
Variance	65.5
Standard deviation	8.05

The mean of 85.4 lies above the median of 78.0, and the most frequent score is 76. Marks span the range 60–100, and the interquartile range of 25 indicates a moderate spread around the middle of the distribution. The standard deviation of 8.05 gives the typical deviation of a score from the mean. Taken together, the measures of central tendency describe the overall level of performance, while the standard deviation and variance describe how widely scores differ between students.

D. Exploratory Visualization

The preprocessed data were then visualized to examine their distribution. Fig. 3 is a grouped boxplot of internal marks by Skill Score (0–3), with separate boxes for skill developers and non-developers. It was produced with ggplot2, mapping the categorical skill score to the x-axis, internal marks to the y-axis and skill-developer status (Yes/No) to the fill colour; geom_boxplot() draws each group and theme_minimal() gives a clean layout. This design allows comparison at two levels at once: across degrees of skill acquisition, and between learners who do and do not engage in skill development.

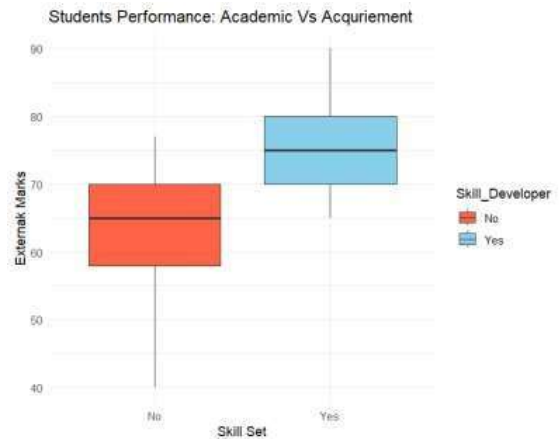


Internal marks by skill score and skill-developer status.

At every skill-score level, learners classified as skill developers have higher median internal marks than non-developers. The figure therefore supports the hypothesis that engagement in skill development is associated with better internal academic performance.

Fig. 4 turns to external marks, using a boxplot because it shows the median, spread and outliers of each group clearly. Learners who took part in skill-development programmes (“Yes”) achieved consistently higher external marks than those who did not; their median is markedly higher and their scores

are less dispersed. This again points to a positive relationship between skill acquisition and academic achievement, here in external evaluation.



External marks by skill-developer status.

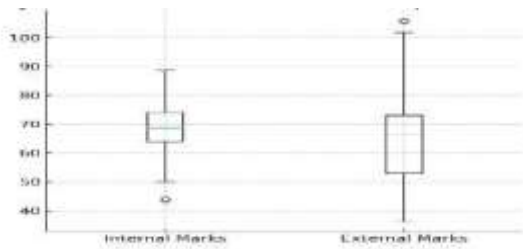
A pie chart (Fig. 5) shows the proportion of learners who participate in skill development. Across the dataset, 58% reported active involvement in skill-development programmes and 42% did not. A majority of the cohort has therefore undertaken at least one skill-building activity beyond the curriculum, in line with the growing emphasis on the holistic development of learners in higher education.

Skill Developer Participation



Distribution of learners by skill-development participation.

Fig. 6 compares the distributions of internal and external marks. Both show a moderate spread, with a few genuinely high performers appearing as mild outliers. The interquartile ranges of 15.2 for internal and 13.8 for external marks indicate a fairly uniform distribution of academic ability. GPA was close to normally distributed, centred at 3.1 on a 4-point scale, which justified the use of parametric procedures for inferential analysis.



Distribution of internal and external marks with outlier detection.

VI. CORRELATION ANALYSIS OF EDUCATIONAL DATA

Correlation analysis can be applied to learners' data at mid-semester to show educators which components of assessment learners are struggling with. A coefficient of +1 indicates a perfect positive linear relationship, -1 a perfect negative linear relationship and 0 no linear relationship. As a first step, a subset of the data was used to correlate performance in internal assessment (x2) with LMS activity (x3). The R code and output are shown in Fig. 7: the coefficient of 0.1557538 indicates only a weak positive association, which tells educators that LMS activity alone says little about internal assessment performance and that attention should go to the specific assessment components where learners lag.

```
corrval<-cor(x2,x3)
> print(corrval)
[1] 0.1557538
```

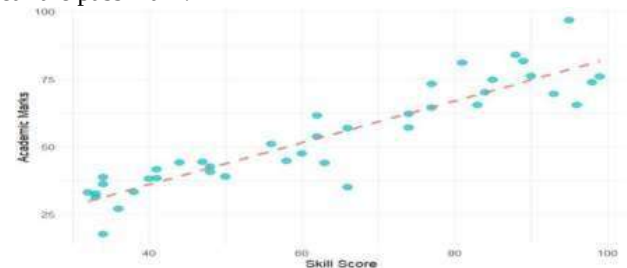
Correlation between internal assessment and LMS activity (R code and output).

A full correlation matrix was then computed on the preprocessed data, using Pearson's r for continuous and Spearman's ρ for ordinal indicators. The strongest positive correlations were:

- internal marks and GPA ($r = 0.74, p < 0.001$);
- attendance and GPA ($r = 0.68, p < 0.001$);
- LMS Activity Index and GPA ($r = 0.61, p < 0.01$);
- Skill Development Index (SDI) and GPA ($r = 0.58, p < 0.01$).

These associations identify regular attendance, strong continuous-assessment results and sustained participation in skill-building activity as determinants of academic performance. Late assignment submission showed a moderate negative correlation with final grades ($r = -0.42, p < 0.05$), underlining the importance of time management. Cross-correlation plots in R showed attendance, LMS participation and SDI clustering as a group of mutually reinforcing variables, an integrated "engagement network". Correlation does not establish causation, but the statistical significance of these relationships justified retaining the variables as predictors in subsequent models. Variance Inflation Factor (VIF) tests confirmed that all VIF values were below 3, so the predictors were not excessively collinear and regression coefficients would remain stable.

Fig. 8 plots academic marks against skill score and shows a clear positive trend. In the same vein, students who accessed the LMS more than five times a week consistently achieved a GPA above 3.2, while those who interacted less often clustered near the pass mark.



Relationship between academic performance and skill score.

By Cohen's conventions, the correlation coefficients represent large effect sizes and thus relationships of considerable explanatory power. Subgroup analysis showed the same pattern at different academic levels: for subgroup 1 learners the coefficients were $r = 0.71$ for internal assessment and $r = 0.65$ for external examinations, while for subgroup 2 learners they were slightly stronger ($r = 0.76$ and $r = 0.72$). This suggests that the benefits of skill development may become more pronounced at higher academic levels.

Analysing learners' performance through measures of central tendency and dispersion also helps institutions forecast enrolment trends and allocate resources. Adding skill-development data shows how learners are extending their knowledge beyond the curriculum and how this relates to their academic results, so that institutions can concentrate on the areas that most need attention and improve both academic and operational outcomes.

VII. DISCUSSION

The results show that skill development and academic performance are correlated in this cohort. The strong positive relationships between skill-development activity and both internal and external measures support the view that a broad educational approach is associated with better results than a purely academic one. In allocating resources, institutions should therefore weigh the benefits associated with skill-development programmes, particularly structured online learning and internships. That 58% of students already take part in such activities indicates substantial interest and readiness for further provision. Because skill engagement helps to predict academic performance, it also gives institutions a view of the new ways in which individual learners acquire knowledge.

Two qualifications temper these conclusions. First, the evidence is correlational: learners who are already strong academically may also be those most inclined to seek additional skills, so the direction of influence cannot be established from these data. Second, the data come from one course in one programme at a single institution. Within these



limits, the findings make a case for investment in experiential learning, online-learning partnerships and professional-development programmes, and for assessing learners not only on traditional academic results but also on their engagement in skill development. The case study, carried out entirely in R, illustrates how predictive analytics that combines academic and skill-development data can help institutions maximize student achievement.

VIII. CONCLUSION AND FUTURE WORK

This paper has shown the value of combining statistical and machine-learning techniques to understand and anticipate learner performance, bringing exploratory and predictive methods together in a single framework. It argues that predictive analytics should support proactive, data-informed intervention rather than serve only to record learners' success. Learner performance was examined along two dimensions, academic results and additional skills acquired through online courses and other activities, and learners who had acquired such skills performed better academically.

In future work, an AI model could be developed to classify learners into types, so that learning activities can be designed

for each type and institutions can better prepare learners for the challenges of their careers. Applying a wider range of predictive models to educational data in its various forms would make digital transformation integral to learner-driven support and inclusive excellence across the higher education system.

REFERENCES

- [1] Deri, M.N., Singh, A., Zaazie, P., Anandene, D.: Leveraging Artificial Intelligence in Higher Educational Institutions: A Comprehensive Overview. *Revista de Educación y Derecho* (Education and Law Review), Issue No. 30 (April–September 2024)
- [2] Vieri, A.M., Petrea, G.: The Impact of Artificial Intelligence (AI) on Learners' Academic Development. *Education Sciences* (2025)
- [3] Liu, Y., Qi, Y., Sun, Y.: AI Learning System Research on Cognitive Ability and Academic Achievement. Presented at ICAIE 2024
- [4] Warter, L.: Impact of Organizational Culture in Higher Education: Case Study. *Journal of Intercultural Management and Ethics*, Issue No. 2 (2019)
- [5] Tierney, W.G., Lanford, M.: *Institutional Culture in Higher Education*. Springer Science (2018)
- [6] Sheik Abdullah, A., Gayathri, N., Abirami, A.M., Parkavi, R.: A Statistical Approach for Setting SLO Targets over Outcome Based Education – A Case Study. *Journal of Engineering Education Transformations*, Special Issue (Jan. 2018)