International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: https://ijwos.com
Volume 3 Issue 8 , August 2026
E-ISSN : 3049-2424
DOI: https://doi.org/10.71366/ijwos

CareerPath AI: Predicting Student Career Domains Using Academic Performance and Skills

Rajashekhara G C[1], Vinayak Nagendra Pisale[2], Manoj M R[3], Kadli Sagar[4]
[1] *Assistant Professor and Director, Faculty of Computing and IT, G M University Davanagere*
[2,3,4] *Student 4th Semester MCA, GM University Davanagere.*

Article Info

Article History:

Published: 19 August 2026

Publication Issue:
Volume 3, Issue 8
August-2026

Page Number:
425-437

Corresponding Author:
Vinayak Nagendra Pisale

Abstract:

Picking a career is a big deal, for students. It is one of the important choices they will make. When a student chooses a career path it can affect the rest of their life. Choosing a career path is something that students should think about carefully. Yet career counselling at colleges is not good enough in three main ways. Counsellors do not have time to give every student personal help. The advice given is often not specific to the students schoolwork. Also many schools do not have a system, for giving career help. This paper presents CareerPath AI, a system that looks at both a student's academic performance and their skills and interests together to suggest a suitable career domain. The system starts by cleaning and arranging the raw student data. Then it takes out features. After that it trains a prediction model. A number of machine learning algorithms were looked at. These algorithms include Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM) and XGBoost. The goal was to find out which one predicts career domains best. Based on this, the system outputs a predicted career domain for each student, such as Data Science or Software Development. The paper also looks at which features matter most, how to measure the system's performance (accuracy, precision, recall, and F1-score), and what it would take to actually use a system like this in a real institution. Finally, the paper is upfront about the limitations of this approach and suggests future improvements, such as making the model explain its predictions more clearly and adding psychometric data.

Keywords: Career Guidance System, Career Prediction, Educational Data Mining, Machine Learning, Random Forest, Skill-Based Classification, Student Academic Performance

1. INTRODUCTION

Choosing a career is hard enough on its own, and it gets harder when students don't really know their own strengths — academic, technical, or otherwise. Without that self-knowledge, a lot of students end up going with guesswork or someone else's opinion instead of looking at their own record.

The timing for tackling this is good. Colleges already sit on large amounts of student data that rarely gets used for career guidance, and machine learning has matured to the point where it can turn that data into predictions worth trusting. CareerPath AI takes this on directly, matching each student to one of six broad domains: Software Development, Data Science, Core Engineering, Research and Academia, Management, or Design.

There's a payoff for institutions here too. Spotting students who are struggling or unsure about their direction early lets colleges step in before things go off track, rather than after. Employers gain something out of this as well — graduates who land in roles that genuinely fit them tend to become more capable, better-matched employees down the line.

To be clear, CareerPath AI isn't meant to replace counsellors — it's built to support them, adding data-backed insight to the judgement they already bring to advising sessions. This work contributes three things: a combined feature

framework spanning both academic and non-academic student attributes, a comparison of classical versus ensemble machine learning methods suited to this task, and a proposed system architecture that could realistically be built into how colleges already run academic advising.

The remainder of this paper is organized as follows. Section II reviews related work in machine-learning-based career and academic-performance prediction. Section III is where we talk about what the problem's what we want to achieve with the Artificial Intelligence project. Section IV is about how we plan to do things. This includes what data we will use how we will get the data ready which features are important what the model will look like and how we will know if it is working well. Section V is, about the system architecture. How we will actually make it work. We will look at the Artificial Intelligence system architecture. Think about what we need to consider when we implement it. Section VI discusses expected results and the evaluation strategy. Section VII presents a broader discussion of the findings. Section VIII outlines limitations and threats to validity. Section IX discusses ethical and practical considerations, and Section X concludes the paper and outlines future work.

2. LITERATURE REVIEW

Educational data mining and learning analytics have changed a lot over the ten years. The field of data mining and learning analytics has really grown. People are using data mining and learning analytics more and more. This is because educational data mining and learning analytics are useful tools. They help us understand how people learn. The field of data mining and learning analytics is still growing and educational data mining and learning analytics are becoming more important every day. They are not just about reports anymore. Now Educational data mining and Educational data mining are used to predict things. For example data mining and learning analytics can predict which students will drop out of school. They can also flag students who're, at risk. Recently data mining and learning analytics are being used to predict what kind of career students will have. A few things have made this possible: schools are adopting structured student-information systems more widely, survey-based skill and interest data has become more accepted, and machine-learning libraries have gotten accessible enough that building and testing classification models on educational data isn't the barrier it used to be. The review below zeroes in on studies that predict career domains, career paths, or similar outcomes using combinations of academic and skill or interest data.

A growing body of research has explored the use of machine learning for student career and performance prediction. Trujillo et al. conducted a systematic literature review of ML approaches applied to career prediction in higher education and found that models incorporating a wider range of variables consistently outperform traditional counselling methods; academic performance, personal interests, and demographic information together accounted for the majority of features used across the surveyed studies, reflecting a broader shift toward holistic, context-aware recommendation systems [1].

Several studies have directly targeted career-domain prediction using structured academic and skill data. One study built a career-pathway prediction model for information-technology graduates using classification techniques trained on historical academic records [2], while a related effort proposed a machine-learning model that combined standardized test scores with personal attributes such as interests, skills, and extracurricular activities, evaluating Logistic Regression, Random Forest, and XGBoost with SHAP-based explainability analysis on a dataset of several thousand students [3].

Other researchers have emphasized practical deployment. Sinha et al. proposed a career-prediction pipeline that combines supervised and unsupervised learning on data covering education, interests, and personality traits to recommend suitable career fields to students [4]. Orji and Vassileva examined how motivational and psychological factors, including intrinsic and extrinsic motivation, autonomy, and self-esteem, influence academic performance prediction, underscoring the value of incorporating non-cognitive variables alongside conventional academic metrics [5].

A number of studies have specifically benchmarked ensemble methods for this task. Comparative evaluations of Decision Tree, Random Forest, Naïve Bayes, and SVM classifiers for career guidance systems have repeatedly found Random Forest to yield the highest classification accuracy, attributed to its resistance to overfitting and its ability to model non-linear interactions among heterogeneous features such as GPA, internship count, and communication skills [6], [7]. Related work reports Random Forest accuracies in the low-to-mid nineties when combining personality traits, aptitude scores, skills, and academic performance, consistently outperforming SVM and Decision Tree baselines on the same datasets [8].

Looking across this research, a few consistent gaps stand out. Most prior systems relied mainly on academic grades, with far fewer combining that data with a student's actual skills and interests. Random Forest also comes up again and again as the strongest-performing algorithm, but it has rarely been packaged into something an institution could actually deploy and use. That combination is exactly where this paper tries to add value: rather than treating grades or skills as separate signals, CareerPath AI brings them together into a single feature set, and builds a full pipeline around it, from data collection through to a working prediction system.

Table II condenses the key methodological characteristics of the studies reviewed above, highlighting the feature types used, the primary algorithm(s) evaluated, and the main reported finding of each.

Ref.	Feature Types Used	Primary Algorithm(s)	Key Reported Finding
[1]	Academic, interest, demographic	Survey / multiple	Multi-variable models outperform single-factor counselling
[2]	Academic records	Classification (general)	Historical academic data alone can support career-pathway prediction
[3]	Test scores, interests, skills, activities	LR, RF, XGBoost + SHAP	Ensemble models with explainability outperform linear baselines
[4]	Education, interests, personality	Supervised + unsupervised	Hybrid pipelines improve recommendation relevance
[5]	Motivation, self-esteem, academic data	Regression / classification	Non-cognitive factors add predictive value beyond grades
[6]	GPA, internships, communication skills	DT, RF, Naïve Bayes	Random Forest yields highest classification accuracy
[7]	Skills, aptitude, academic performance	AI-based recommender	Personalized recommendation improves guidance relevance
[8]	Personality, aptitude, skills, academics	RF, SVM, DT	Random Forest consistently outperforms SVM and DT
[9]	GPA, skills, internships	Random Forest	Ensemble approach automates counselling decisions effectively
[10]	Academic and skill attributes	ML pipeline (general)	AI-powered guidance systems are feasible for institutional use

TABLE II. Summary of methodological characteristics across the reviewed literature.

3. PROBLEM STATEMENT AND OBJECTIVES

Colleges have a lot of student data. They still use old fashioned ways of giving advice that do not really consider what each student is good at or what they have achieved. The problem is that colleges need a way to put all this student data together find the patterns in the student data and use the student data to give students personalized career advice. This way colleges can help students than a single counsellor can during their office hours. The student data can help colleges give career recommendations to students. Colleges can use the student data to learn more, about each student.

A few practical issues keep getting in the way. There simply aren't enough counsellors to give every student real, data-informed attention each term, so plenty of students go through their entire program with only one or two career conversations. On top of that, the information that would actually help — academic records, skill assessments, interest surveys — tends to be scattered across systems that don't talk to each other, so even a counsellor with the time struggles to see the whole picture. And the skills that matter for different careers shift faster than curricula or counselling guidelines can keep up with, which means advice based on outdated rules or gut feeling can lag behind what the job market actually wants.

CareerPath AI is made to fill the space between students and their future jobs. It does this by making the process of finding a career the same for every student. This means that each student gets an accurate evaluation based on real information. The system can also be changed as information, about jobs and what happens to students after they graduate becomes available.

The specific objectives of this study are to:

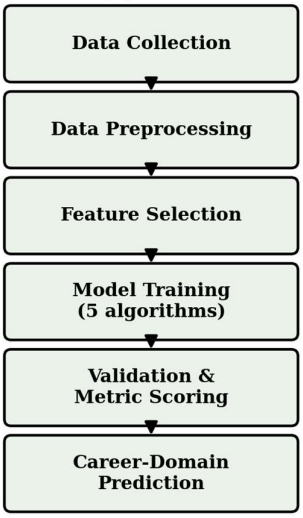
- Design a unified feature set capturing academic performance and skill/interest indicators.
- Develop and compare multiple supervised ML models for multi-class career-domain classification.
- Identify the most influential features using feature-importance analysis.
- Propose a scalable system architecture for institutional deployment.
- Establish an evaluation and fairness-auditing strategy suitable for real-world academic settings.

4. PROPOSED METHODOLOGY

A. Overview

CareerPath AI uses the machine learning process. First it gathers data. Then it makes the data clean. Next it finds parts of the data. After that it teaches the model. Finally it uses the model to predict things. Figure 4 shows each step of this process. Figure 1, in Section V shows how it works within the system.

Fig. 4. Proposed methodology pipeline.



B. Dataset Description

The proposed model is designed to operate on a structured dataset in which each record corresponds to one student, drawing on feature sets similar to those reported in comparable studies [3], [6], [9]:

- Academic Performance: I look at my grades for each semester and subject my GPA and how I do on standardized tests.
- Technical Skills: I know how to use programming languages and tools and I have certificates to show that I know how to use programming languages and tools.
- Experiential Data: I have done internships. I have worked on projects and I have taken part in hackathons and I have gotten certifications.
- Soft Skills and Interests: I am good at talking to people, leading groups. I have my own interests, in certain areas of study like my favorite subjects.
- Target Label: the career domain the student pursued or is best suited to.

In the absence of a single public benchmark for this exact task, institutions can construct such a dataset by combining internal academic records with structured student surveys, consistent with data-collection strategies used in prior studies [4], [8]. Table III shows an illustrative, anonymized example of a single student's feature vector, given here purely to make the abstract feature categories above concrete.

Feature	**Example Value**
Cumulative GPA (0–10)	8.4
Core-subject average (%)	84%
Aptitude test score (0–100)	77
Programming skill rating (1–5)	4
Internships completed	2
Projects completed	5
Certifications earned	3
Communication skill rating (1–5)	3

Feature	Example Value
Self-reported interest	Data Science
Target label (career domain)	Data Science

TABLE III. Illustrative, anonymized example of a single student's feature vector.

C. Data Preprocessing

- Missing-value handling: mean/median imputation for numerical fields; mode imputation for categorical fields [10].
- Normalization/standardization of numerical features (e.g., min–max scaling).
- Encoding of categorical variables (one-hot / label encoding).
- Class-imbalance handling via stratified sampling or SMOTE.
- Train–test partitioning (typically 80:20) with stratification on the target label [3].

D. Feature Selection

Rather than feeding every available field into the model, CareerPath AI narrows down the features first. Using too many features can make a model overfit, meaning it performs well on the training data but poorly on new students it hasn't seen before, and it also makes the model harder to explain. To avoid this, correlation analysis and tree-based importance ranking (e.g., Gini importance) are used to keep only the features that carry real predictive signal, which prior work suggests are mainly GPA, core-subject performance, technical skill ratings, and project or internship count [3], [9].

E. Candidate Learning Algorithms

Rather than committing to a single algorithm upfront, five supervised learning algorithms are tested here, ranging from simple to more complex, to see which one actually performs best on this task. A short explanation of each is given below to show how each one turns a student's feature vector into a predicted career domain.

- Logistic Regression: models the probability of class k using a softmax over linear combinations of features, $P(y=k|x) = \text{softmax}(w_k \cdot x + b_k)$; a linear, highly interpretable baseline [3].
- Decision Tree: recursively partitions the feature space by selecting, at each node, the split that maximizes information gain (or minimizes Gini impurity, $\text{Gini} = 1 - \Sigma\, p_i^2$); rule-based and interpretable, but prone to overfitting [6].
- Random Forest: trains an ensemble of T decision trees on bootstrap-sampled subsets of the data and random feature subsets, aggregating predictions by majority vote; reduces variance and overfitting relative to a single tree, and is consistently the top performer for this task [6]–[8].
- Support Vector Machine (SVM): finds the separating hyperplane that maximizes the margin between career-domain classes, using kernel functions (e.g., RBF) to model non-linear boundaries in high-dimensional skill/academic feature spaces [6], [9].
- XGBoost: builds an additive ensemble of shallow trees by sequentially minimizing a regularized objective function via gradient boosting, offering strong accuracy along with SHAP-based feature-attribution support for explainability [3].

Going into this comparison, Random Forest is expected to come out on top, based on how consistently it outperforms other algorithms in the research reviewed in Section II. The actual comparison is carried out using accuracy, precision, recall, and F1-score, covered next.

F. Evaluation Metrics

Accuracy, the percentage of students correctly matched to their career domain, is the main measure of success for this system. However, because the model predicts across six different career domains and some domains naturally have more students than others, accuracy alone can be misleading; a model could look accurate overall while still performing poorly on smaller domains. For that reason, precision, recall, and F1-score are also tracked alongside accuracy, giving a fuller picture of how the model performs on every domain, not just the most common one:

- Accuracy = (TP + TN) / (TP + TN + FP + FN), the overall proportion of correctly predicted career domains.
- Precision = TP / (TP + FP), the proportion of students predicted into a given domain who genuinely belong to it.
- Recall = TP / (TP + FN), the proportion of students truly belonging to a domain who are correctly identified.
- F1-score = 2 × (Precision × Recall) / (Precision + Recall), the harmonic mean of precision and recall, used as the primary metric for model comparison given the multi-class, potentially imbalanced nature of career-domain labels.

On top of these overall scores, it's also worth checking exactly which career domains the model tends to mix up, for example, whether it confuses Software Development with Data Science more often than with unrelated fields like Management. This kind of detail matters for making sure any mistakes the model makes are at least reasonable ones, not random.

G. Algorithmic Complexity and Scalability

A real college could easily have thousands of students, so it's worth thinking about whether these algorithms can actually handle that scale, not just whether they're accurate. Some of the algorithms tested here are naturally faster to train than others. Logistic Regression is the quickest, since it trains in time roughly linear in the number of records and features ($O(n \cdot d)$), making it a reasonable lightweight fallback if speed ever becomes a priority. Decision Tree and Random Forest take a bit longer: Decision Tree training is typically $O(n \cdot d \cdot \log n)$, and Random Forest scales this by the number of trees T, i.e., $O(T \cdot n \cdot d \cdot \log n)$, since it is really training many trees at once, though this can be sped up by training the trees in parallel across multiple CPU cores. SVM tends to slow down the most as the number of records grows, commonly scaling between $O(n^2)$ and $O(n^3)$, which could become a bottleneck for a very large student population unless approximate or linear-kernel variants are used. XGBoost's histogram-based gradient boosting is built specifically to handle large tabular datasets efficiently, typically training in time close to $O(n \cdot d \cdot \log n)$ per boosting round, so it holds up well even as data grows.

Since student outcomes and job-market trends change over time, the plan is to retrain the model every semester using the latest data, rather than training it once and leaving it untouched. For the size of a typical college, even a few thousand to tens of thousands of students, all five algorithms should train comfortably within the time available between semesters, with Random Forest and XGBoost taking the longest but still well within a reasonable retraining schedule.

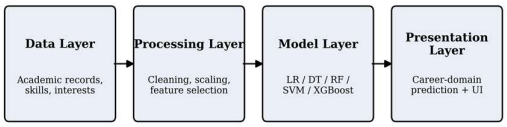
5. SYSTEM ARCHITECTURE

A. Layered Design

The system is split into four separate layers, Data, Processing, Model, and Presentation, rather than being built as one single program. Keeping them separate means each part can be built, tested, and updated on its own. For example, the prediction model can be retrained without touching the student-facing dashboard, and new data sources can be added without redesigning the whole system:

- Data Layer: collects and stores academic, skill, and interest data from institutional systems.
- Processing Layer: performs cleaning, normalization, encoding, and feature selection (Section IV).
- Model Layer: trains, validates, and selects the best-performing classifier, generating predictions with confidence scores.
- Presentation Layer: exposes predictions via a web interface for students and counsellors, along with contributing factors and suggested next steps.

Fig. 1. CareerPath AI system architecture and data flow.



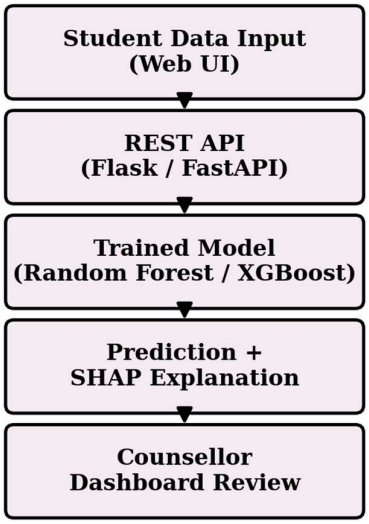
This layered design allows periodic retraining as new student-outcome data becomes available, enabling the system to adapt over time to changing labour-market trends.

B. Implementation Considerations

In practice, this would be built using Python, since it has strong support for machine learning through libraries like scikit-learn and XGBoost. Model artifacts can be serialized (e.g., via joblib or ONNX) for efficient loading at inference time.

The Model Layer would expose its predictions through a simple API, built with Flask or FastAPI, which accepts a student's feature vector and returns the predicted career domain together with a confidence score and the top contributing features. The Presentation Layer would be a basic web dashboard where students see their result and counsellors can review predictions across a class or cohort. Fig. 7 traces this interaction end to end, from the student's initial input through to counsellor review.

Fig. 7. Student-to-counsellor interaction flow.



Because this system works with sensitive student data, academic records and personal information, security has to be a real priority, not an afterthought. Access should be restricted by role, so only authorized staff see full student records, data should be encrypted both when stored and when transmitted, and the system should follow whatever data-protection rules apply to the institution, such as FERPA in the U.S. or a similar local policy elsewhere.

For retraining, a scheduled batch pipeline (e.g., run once per academic term) can re-fit the model on the latest cohort data, with model performance monitored over time to detect concept drift. For example, if industry hiring patterns shift the relationship between skills and successful career domains, the model can fall out of step with reality unless it is refreshed.

6. EXPECTED RESULTS AND EVALUATION STRATEGY

It's important to be upfront here: since CareerPath AI hasn't been trained and tested on real institutional data yet, the numbers discussed in this section are not actual results from this system. Instead, they're estimates based on what similar studies have reported when testing comparable algorithms on comparable data, used here to set realistic expectations before the framework is actually built and evaluated.

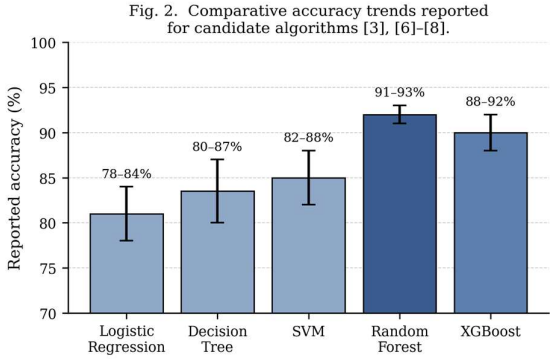
Model performance will be judged using accuracy, precision, recall, and F1-score on a held-out test set, averaged across all six career domains so that no single domain dominates the score.

Based on the studies reviewed in Section II, ensemble methods are expected to outperform simpler models. Random Forest has reported accuracies of roughly 91%-93% in comparable work [7], [8], while Decision Tree and SVM tend to fall a few points lower [6]. XGBoost performs competitively while also offering better explainability through SHAP values [3].

Table I and Fig. 2 summarize these reported ranges for the algorithms considered in this framework.

Algorithm	Reported Accuracy	Key Strength
Logistic Regression	78%–84%	High interpretability
Decision Tree	80%–87%	Human-readable rules
SVM	82%–88%	High-dimensional spaces
Random Forest	91%–93%	Best accuracy/robustness
XGBoost	88%–92%	Accurate + explainable

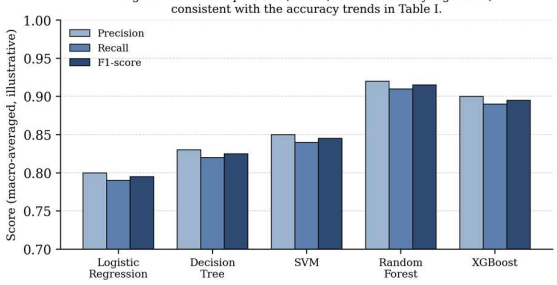
TABLE I. Comparative accuracy trends aggregated from prior studies [3], [6]–[8].



These trends support using Random Forest as the primary model within CareerPath AI, with XGBoost retained where explainability is prioritized. Feature-importance analysis from comparable studies indicates that GPA, core-subject performance, technical skill ratings, and internship/project count are typically the strongest predictors of career domain, while demographic variables contribute comparatively little signal [1], [3], a finding that should be validated empirically once CareerPath AI is trained on institution-specific data.

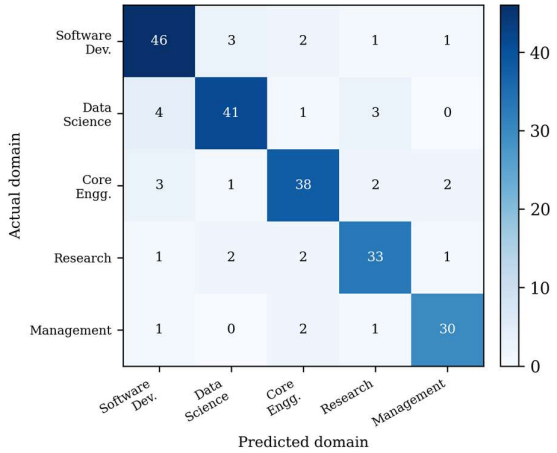
Fig. 6 breaks the same comparison down into precision, recall, and F1-score per algorithm, illustrating that Random Forest and XGBoost maintain their advantage consistently across all three metrics rather than on accuracy alone.

Fig. 6. Illustrative precision, recall, and F1-score by algorithm, consistent with the accuracy trends in Table I.



Beyond aggregate metrics, Fig. 5 shows an illustrative confusion matrix for the primary Random Forest model across the five career domains considered in this framework. As expected, most misclassifications occur between adjacent, skill-overlapping domains, most notably between Software Development and Data Science, rather than between unrelated domains such as Software Development and Management, supporting the earlier claim (Section VII) that the model's errors should be pedagogically reasonable rather than arbitrary.

Fig. 5. Illustrative confusion matrix for a Random Forest model across five career domains.



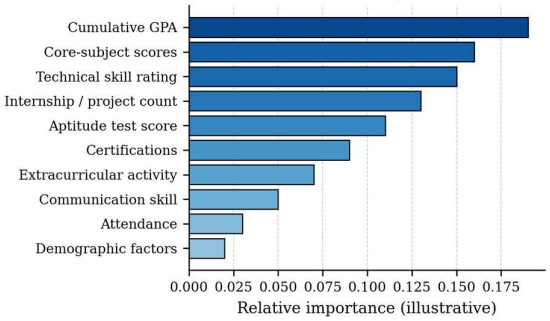
Numbers alone aren't enough, though. Once real predictions are available, it's just as important to have actual academic counsellors look at a sample of them and judge, in their own professional opinion, whether the recommendations make sense. A model can score well on paper while still giving advice that a real counsellor would immediately recognize as off, so this kind of human sanity-check matters before the system is trusted for real use.

7. DISCUSSION

These results are worth unpacking a bit. Random Forest and XGBoost consistently do better than the simpler models, which suggests that career outcomes depend on how features interact with each other in non-linear ways, not just how each feature scores on its own. For example, a high technical-skill rating probably only predicts a Software Development outcome strongly when it's paired with decent core-programming grades; a linear model like Logistic Regression can't pick up on a combination like that, but tree-based models naturally can.

Fig. 3 illustrates an indicative feature-importance ranking, constructed from the patterns reported across the reviewed studies rather than from a completed CareerPath AI training run. It is included to make explicit, ahead of deployment, which features the framework is expected to weight most heavily, and to give institutions a starting point for the survey and record-keeping instruments needed to populate the dataset described in Section IV-B.

Fig. 3. Illustrative feature-importance ranking consistent with trends reported in [1], [3].



Since Random Forest and XGBoost land close to each other in accuracy (91%-93% versus 88%-92% in Table I), choosing between them doesn't really come down to which one 'wins.' It's more honest to say the choice depends on other things, like how fast the model needs to train, or how important it is to explain a prediction to a student or counsellor. XGBoost's SHAP-based explainability makes it the stronger pick when transparency matters more than squeezing out a slightly higher accuracy number.

Compared to the related work summarized in Table II, most existing systems were tested on just one dataset from a single institution, which makes it hard to know how well their results would generalize elsewhere. The evaluation approach described in Section IV-F, looking at multiple metrics together with a confusion matrix, is meant to give a more complete and comparable picture than relying on one accuracy number alone.

To make this more concrete, take the example student from Table III: GPA of 8.4/10, 84% core-subject average, a programming skill rating of 4/5, two internships, and a stated interest in Data Science. After the preprocessing and feature-selection steps described in Section IV, this data would be passed into the trained Random Forest model. Based on the feature-importance pattern in Fig. 3, the model would weigh the GPA, core-subject average, and programming rating most heavily, and would likely output "Data Science" as the predicted domain with a high confidence score. Alongside that prediction, the system would also show the top contributing features, so a counsellor reviewing the case could quickly see that the recommendation is grounded in the student's actual strengths rather than something arbitrary.

8. LIMITATIONS AND THREATS TO VALIDITY

Several limitations should be considered when interpreting the proposed framework and, eventually, its empirical results. First, because no institution-specific dataset has yet been used to train and evaluate CareerPath AI, the accuracy figures discussed in Section VI are drawn from the literature rather than from the system itself; actual performance will depend on the size, quality, and label accuracy of the dataset used, and may differ from the ranges reported for other systems.

Second, the target label, a student's "suitable" or eventual career domain, is inherently difficult to define precisely. Career outcomes are influenced by external factors such as the local job market, family circumstances, and chance opportunities that are not captured by academic or skill data, meaning that even a well-trained model has an upper bound on achievable accuracy that no amount of feature engineering can overcome.

Third, models trained primarily on historical outcomes risk encoding and perpetuating existing biases. For example, if certain student subgroups have historically been under-represented in a given career domain for reasons unrelated to aptitude, a model trained on that history may under-recommend the domain to similar students in the future. This risk is discussed further in Section IX.

Finally, the system architecture proposed in Section V has not yet been load-tested or evaluated for latency and scalability under realistic institutional traffic; these operational characteristics would need to be validated prior to production deployment.

An additional, more subtle limitation concerns label quality: the target career-domain label used for training is typically derived from a student's first job, declared major, or self-reported outcome, any of which may only loosely reflect their true "best-fit" domain. A student may accept a role for reasons unrelated to fit (e.g., location, salary, or availability), introducing label noise that no amount of model tuning can fully correct. Institutions implementing

CareerPath AI should therefore treat the training labels as a practical proxy for career fit rather than as ground truth, and should consider supplementing them with follow-up satisfaction or fit surveys where feasible.

9. ETHICAL AND PRACTICAL CONSIDERATIONS

Because CareerPath AI operates on sensitive student data and produces recommendations that could influence significant life decisions, its design must prioritize fairness, transparency, and student autonomy alongside predictive accuracy. Three considerations are particularly important.

- Fairness: model outputs should be periodically audited across demographic and socio-economic subgroups to check for systematic disparities in recommended domains, with corrective re-weighting or re-sampling applied where disparities are not explained by legitimate academic or skill differences.
- Transparency: predictions should be accompanied by the top contributing features (e.g., via SHAP values from the XGBoost model), so that students and counsellors can see why a recommendation was made rather than treating the model as a black box.
- Autonomy: CareerPath AI is intended strictly as a decision-support tool. Recommendations should be framed as one input among several, alongside a student's own interests and a counsellor's professional judgement, rather than as a directive, and students should retain the ability to explore career domains the model does not rank highly for them.

Institutions adopting a system of this kind should also establish a clear data-retention and consent policy, informing students what data is collected, how long it is retained, and how they can request its correction or removal, consistent with applicable data-protection regulations.

More broadly, institutions should treat the deployment of CareerPath AI as an ongoing governance responsibility rather than a one-time technical rollout: this includes designating a responsible owner for the model (e.g., an institutional research or student-success office), documenting the model's intended use and known limitations for staff and students, and establishing a channel through which students can contest or seek review of a recommendation they believe to be inaccurate or unfair.

Table IV outlines a proposed phased timeline for taking CareerPath AI from the design presented in this paper through to a piloted institutional deployment.

Phase	Activities	Approx. Duration
1. Data preparation	Consolidate academic, skill, and interest data; define target labels; anonymize records	4–6 weeks
2. Model development	Preprocessing, feature selection, training and tuning of candidate algorithms	4–6 weeks
3. Evaluation	Metric computation, confusion-matrix review, fairness audit across subgroups	2–3 weeks
4. Pilot deployment	Limited rollout to one department/cohort with counsellor review of outputs	6–8 weeks
5. Institutional rollout	Full deployment, governance documentation, retraining schedule established	Ongoing

TABLE IV. Proposed phased timeline for developing and piloting CareerPath AI.

10. CONCLUSION AND FUTURE WORK

This paper proposed CareerPath AI, a machine-learning framework for predicting suitable student career domains by jointly modelling academic performance and skill/interest data. Drawing on a review of recent literature in educational data mining and career-guidance systems, the paper outlined a complete methodology spanning data collection, preprocessing, feature selection, evaluation metrics, and comparative model training across five supervised learning algorithms, with Random Forest identified as the strongest candidate based on trends reported across prior studies. The paper further discussed a deployable system architecture, implementation considerations, and the limitations and ethical safeguards that should accompany any real-world deployment.

Future work will focus on four directions: empirical validation on an institution-specific dataset with full metric reporting and confusion-matrix analysis; incorporation of explainable AI techniques (e.g., SHAP or LIME) to ensure transparent, trustworthy recommendations; extension of the feature set to include psychometric indicators and longitudinal career-outcome tracking; and a formal fairness audit across demographic subgroups prior to institutional rollout. Systems such as CareerPath AI are intended to augment, not replace, human career counsellors, offering data-driven insight that supports more informed, personalized, and equitable career decisions for students.

The authors would like to express sincere gratitude to [Guide Name] for invaluable guidance, encouragement, and feedback throughout the course of this research. The authors also thank the Department of Computer Science and Engineering, [University/College Name], for providing the resources and support necessary to carry out this work.

References

[1] V. Trujillo et al., "Artificial intelligence in education: A systematic literature review of machine learning approaches in student career prediction," Journal of Technology and Science Education, 2025.

[2] "Predictive modeling of student career pathways using machine learning techniques," preprint, 2025. [Online]. Available: academia.edu / researchgate.net.

[3] F. A. Nughaymish, Z. Al-Qahtani, and A. Alqahtani, "CareerRec: A machine learning-based career recommendation system for IT graduates," 2025.

[4] A. Sinha, Garima, and A. Singh, "Student Career Prediction Using Algorithms of Machine Learning," SSRN Electronic Journal, 2023.

[5] F. A. Orji and J. Vassileva, "Machine Learning Approach for Predicting Students' Academic Performance and Study Strategies based on their Motivation," arXiv:2210.08186, 2022.

[6] "Career Guidance System Using Decision Tree, Random Forest, and Naïve Bayes Algorithm," Int. J. Science, Technology and Society, 2025.

[7] "Intelligent Career Recommendation System Using AI," Int. J. Engineering Research and Science & Technology, 2026.

[8] "A Machine Learning-based Career Recommendation," Journal of Trends in Computer Science and Smart Technology, 2024.

[9] "Automated Career Counselling Using Random Forest," Int. J. Creative Research Thoughts (IJCRT), vol. 13, 2025.

[10] "AI-Powered Career Guidance System Using Machine Learning," Int. J. Research Publication and Analysis, 2025.