Explainable AI for Career Prediction Systems: A SHAP-Based Framework for Transparent Decision-Making

[1]Mrs. Faria Begum N, [2]Dr. Mrutyunjaya M S
[1]PG Student,

[2]Associate Professor, Department of Computer Science and Engineering (Data Science)

R L Jalappa Institute of Technology, Doddaballapur.

ABSTRACT

The widespread adoption of Artificial Intelligence (AI) and Machine Learning (ML) models across various sectors, including career guidance, has introduced a critical need for transparency and interpretability. While these models offer unprecedented predictive power, their inherent "black box" nature can erode user trust, hinder accountability, and perpetuate biases. This paper examines the pivotal role of Explainable AI (XAI), specifically focusing on SHapley Additive exPlanations (SHAP), in addressing these challenges within Career Prediction Systems (CPS). The proposed SHAP-based framework integrates advanced ML models with a dedicated explainability layer, leveraging intuitive visualizations and natural language narratives to make complex predictions comprehensible to diverse stakeholders. Despite challenges related to computational efficiency and data privacy, advancements in XAI promise to transform career guidance into an empowering and equitable tool.

Keywords: **Explainable AI, SHAP, Career Prediction, Transparency, Interpretability, Machine Learning, Ethical AI**

1 INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) have become deeply embedded in modern society, influencing diverse domains ranging from web search and speech recognition to medical diagnostics and national defense systems [1]. These technologies have demonstrated exceptional performance, often exceeding human capabilities in solving complex problems. However, as AI models become increasingly sophisticated, a significant concern has emerged—their inner workings are becoming less interpretable to humans.

Modern AI systems, particularly deep neural networks and ensemble methods, are frequently characterized as “black boxes” [2]. While their input and output behavior can be observed, the internal decision-making process remains obscure [3]. This lack of transparency is especially problematic in safety-critical and ethically sensitive areas such as healthcare, finance, and autonomous systems, where comprehending the rationale behind AI decisions is not just beneficial—it is essential [4]. A loan rejection or a clinical diagnosis made by an AI model without an accompanying explanation can provoke confusion, mistrust, and a reluctance to accept or rely on the system [5]. This trust deficit, rooted in opacity, significantly impedes the broader adoption of AI technologies in critical human-centric applications [5].

The emergence of Explainable Artificial Intelligence (XAI) aims to address this pressing issue by enhancing the transparency and interpretability of AI systems. XAI seeks to answer fundamental questions such as, “Why did the AI system make this decision?” and “What factors influenced the outcome?” [6]. It enables stakeholders to understand, validate, and even challenge the behavior of complex models, going beyond technical debugging to support fairness, accountability, and trust [7]. By translating complex algorithmic logic into understandable narratives [12], XAI facilitates a shift from human-AI interaction to meaningful human-AI collaboration [10].

The objective of this study is to explore the critical role of explainability in AI decision-making systems, particularly in domains where transparency is vital for ethical outcomes. Special emphasis is placed on the

practical, social, and governance implications of opaque AI models and how XAI methodologies can mitigate these concerns. Furthermore, this work highlights the importance of transparency not only for individual comprehension and trust but also for enabling external audits, bias detection, and regulatory compliance [11]. Ultimately, explainability is presented not as an optional enhancement, but as a fundamental requirement for responsible and sustainable AI integration into real-world decision-making processes.

2 LITERATURE REVIEW

2.1 Foundations of Career Prediction Systems

Career Prediction Systems (CPS) are rapidly gaining traction as intelligent platforms that assist students, educators, and employers in making well-informed, data-driven career decisions. Career prediction systems (CPS) combine techniques from artificial intelligence, machine learning, and statistical analysis to offer tailored insights into a student's career readiness. These systems analyze a range of inputs—such as academic records, acquired skills, behavioral traits, and personal interests—to generate meaningful recommendations. By doing so, CPS strive to create a more intentional connection between educational experiences and real-world job opportunities.

2.1.1 Predictive Analytics in Career Guidance

Predictive analytics plays a central role in modern Career Prediction Systems (CPS), using historical trends and statistical methods to anticipate future outcomes [14]. In the context of career guidance, this allows for more informed suggestions based on a student's academic history, skillset, and personal interests. AI-integrated CPS can recognize subtle patterns in data—patterns that may not be evident through traditional counseling methods [15].

While conventional guidance often depends on a counselor's judgment or a student's self-reported preferences [16], predictive models are capable of analyzing complex, multi-layered datasets to reveal deeper insights [27]. These insights support more personalized strategies, such as tailored skill development plans or career roadmaps aligned with current job market needs [14].

This transition from reactive advice to data-driven planning reflects a broader shift in how educational support is being offered. Rather than relying entirely on subjective interpretation, CPS introduce a more structured, scalable approach that helps students anticipate and prepare for future opportunities with greater clarity.

2.1.2 Applications and Stakeholders of CPS

Career Prediction Systems (CPS) are designed to serve more than just students—they support a wider community of stakeholders, each engaging with the system for different purposes and outcomes.

1. ***Students*:** CPS help learners understand where they stand in terms of placement readiness by analyzing academic performance, skill gaps, and personal strengths. With this awareness, students can take more focused steps toward their career goals and align themselves with industry expectations [17]. These tools also offer grounded insights into possible job opportunities, helping students pursue career paths that genuinely fit their capabilities and aspirations.
2. ***Human Resource Professionals*:** In the corporate world, HR teams use CPS to anticipate future talent needs, discover internal candidates for upskilling, and streamline hiring workflows. The predictive insights these systems provide contribute to better role alignment and can help reduce employee turnover by ensuring the right people are placed in the right roles [13].
3. ***Educators and Institutions*:** Academic departments benefit from CPS through a clearer understanding of student preparedness across batches. These insights support data-driven curriculum updates, more accurate placement forecasting, and even operational planning such as exam result analysis or tracking academic outcomes [18].

While the advantages of CPS are significant, they must also accommodate the differing priorities of each stakeholder. For example, institutions may emphasize improving placement rates, whereas some students may prefer to explore non-traditional or creative career paths. These potential conflicts point to the importance of keeping CPS transparent and explainable—ensuring that all users, regardless of their goals, can trust the system's guidance.

2.1.3 Data Modalities and Feature Engineering

Career Prediction Systems (CPS) draw on a diverse set of data sources to generate accurate and relevant predictions about student readiness and potential career paths. Among the most commonly used inputs are:

1. ***Academic Records:*** Performance indicators such as grades, GPA, and subject-specific results help highlight a student's academic foundation and consistency over time [16]. These metrics often serve as a baseline for evaluating readiness in various professional domains.
2. ***Co-Curricular and Certification Data*****:** Activities beyond the classroom—like internships, hackathons, or relevant certifications—offer valuable insight into a student's initiative and ability to apply knowledge in practical settings [18]. Such experiences often complement academic learning and reveal a candidate's real-world engagement.
3. ***Technical Skills and Project Work*****:** Demonstrated experience with programming languages, tools, and domain-relevant projects reflects not just technical ability, but also problem-solving and collaboration skills [16]. These components are crucial in assessing employability, especially for roles requiring hands-on expertise.
4. ***Behavioral Attributes:*** Information on work ethic, stress handling, and team collaboration sheds light on the student's workplace readiness [17].
5. ***Demographics and Interests:*** Personal preferences, career interests, socio-economic status, and geographical background help contextualize and personalize predictions [16].

The process of feature engineering—selecting, cleaning, and transforming raw data into model-friendly formats—is critical to building effective CPS [14]. When done carefully, it ensures that the model captures the most relevant and predictive signals from a complex and multi-dimensional dataset.

However, this data richness brings its own challenges. The diverse and often correlated features increase model complexity, often leading to the use of "black-box" models such as ensemble methods or deep learning [1]. These models, while powerful, lack transparency—making it difficult for users to understand why a particular career path was recommended.

To address this, the incorporation of Explainable AI (XAI) techniques such as SHAP has become increasingly important. SHAP values help explain the contribution of each input feature to the model's predictions, enhancing both accountability and interpretability [20]. In high-stakes scenarios—such as those involving student futures—this transparency builds trust and ensures that CPS remain both effective and ethically grounded.

Data types	Examples	Role in CPS	Ref.
Academic Records	10th, 12th, college mark sheets; GPA; subject-wise scores	Core indicators of academic performance and learning ability. Often used to assess eligibility.	[16]
Certificates & Activities	Internships, club participation, competitions, leadership roles	Reflects real-world exposure, teamwork, and leadership—important for overall readiness.	[17]
Skill Sets	Programming languages, certifications, technical tools	Directly measures technical ability and suitability for industry-specific roles.	[16]
Behavioral Data	Work habits, stress handling, personality traits, team compatibility	Assesses soft skills, adaptability, and workplace fit—key for long-term success.	[17]
Personal Interests	Favorite subjects, interest surveys, psychometric tests	Supports alignment with personal goals and enhances career satisfaction.	[16]
Demographic Data	Age, gender, location, socioeconomic background	Enables personalization and trend	[17]

Table 1: Key Data Types and Their Role in Career Prediction Systems (CPS)

2.1.4 Common Machine Learning Models in Career Prediction

Career Prediction Systems (CPS) employ various machine learning models, ranging from interpretable statistical methods to high-performing AI techniques [14]. Selecting the appropriate model in a Career Prediction System (CPS) involves balancing predictive performance, interpretability, and alignment with the nature of the input data. Several models are commonly employed, each offering distinct strengths depending on the use case:

1. ***Random Forest:*** This ensemble method combines multiple decision trees to improve prediction accuracy and reduce the risk of overfitting. It has shown strong performance in CPS contexts, with reported accuracies reaching as high as 93% [16].
2. ***Support Vector Machines (SVM):*** Particularly effective in classification tasks involving high-dimensional data, SVMs are known for their ability to find optimal decision boundaries in complex spaces [16].

3. ***Neural Networks*:** Useful for identifying non-linear patterns in large and varied datasets, neural networks offer powerful modeling capabilities. However, their lack of transparency—often referred to as the "black-box" problem—remains a limitation [8].
4. ***Decision Trees:*** Favored for their simplicity and interpretability, decision trees split data into clear decision rules based on feature thresholds. This makes them ideal for applications where transparency is a priority [14].
5. ***XGBoost:*** A gradient boosting algorithm built on decision trees, XGBoost is known for its high speed and accuracy, especially when working with sparse or noisy data [16].
6. ***Logistic Regression*:** A foundational model in binary classification, logistic regression is appreciated for its straightforward interpretation and effectiveness when relationships between variables are linear.
7. ***K-Nearest Neighbor (KNN) and Naïve Bayes:*** These models are also applied in CPS, particularly in tasks where simplicity and fast pattern recognition are sufficient for meaningful outcomes .

Despite the range of options, a persistent challenge in model selection is finding the right balance between accuracy and explainability. While advanced models often outperform simpler ones in terms of prediction, their complexity can hinder transparency. Post-hoc interpretation tools such as SHAP have become increasingly valuable in addressing this gap, allowing stakeholders to understand how predictions are made without sacrificing model performance [22].

3. METHODOLOGY

3.1 SHAP: A Unified Framework for Model Interpretability

SHAP (Shapley Additive exPlanations) has emerged as a widely adopted approach for interpreting the outputs of complex machine learning models, particularly in sensitive areas such as career prediction. Grounded in solid game-theoretic principles, SHAP helps clarify the role each input feature plays in influencing a model's decision, offering both mathematical soundness and actionable explanations.

3.1.1. Theoretical underpinnings

Shapley Values and Fair Attribution SHAP is grounded in cooperative game theory, leveraging Shapley values introduced by Lloyd Shapley [22]. These values distribute a model's prediction (the "payout") fairly among features (the "players") based on their marginal contributions across all feature combinations [23]. By accounting for feature interactions and adhering to properties such as efficiency, symmetry, and local accuracy, SHAP ensures a fair and consistent distribution of contribution scores across features [25]. Its strong mathematical foundation lends credibility to its interpretations, making it especially suitable for applications like career prediction, where transparency and fairness are not just beneficial but necessary [26].

3.1.2 Mechanism

SHAP offers both local and global interpretability by analyzing how a model's output changes when specific features are included or excluded from the prediction process [24]. This comparison forms the basis for two types of explanations:

1. ***Local explanations:*** These help clarify individual predictions—for instance, highlighting which specific attributes influenced a student's recommended career path [22].
2. ***Global explanations:*** These offer a broader view by identifying which features most commonly affect outcomes across the entire dataset. Such insights are particularly valuable for educators and administrators seeking to understand key factors that drive placement readiness or success [8].

One of SHAP's strengths lies in its model-agnostic nature, allowing it to be applied across a wide range of algorithms, including decision trees, neural networks, and linear models [10].
By combining both individualized feedback and high-level insights, SHAP supports transparent decision-making at multiple levels—helping users trust the system while enabling institutions to audit and refine their models more effectively.
.

3.1.3 Why SHAP Matters for Model Understanding

SHAP is widely recognized for several key strengths that enhance model interpretability, particularly in sensitive domains like career prediction:

1. ***Quantified feature importance:*** SHAP assigns each feature a precise score, indicating not just how much it contributes to a prediction, but also in which direction it influences the outcome [10].

2. ***Fair and consistent attributions:*** By ensuring a balanced distribution of influence among features, SHAP promotes reliable and reproducible interpretations across different models.
3. ***Support for trust and bias detection:*** It can uncover instances where specific features may be exerting undue influence, helping developers identify potential biases and make ethically sound adjustments [8].
4. ***Robust insights and stability:*** SHAP offers consistent results across multiple runs and brings to light interactions between features as well as patterns of variation within the data [21].

In the context of career prediction systems, SHAP adds value by going beyond surface-level correlations. It clarifies the reasoning behind model outputs, enabling more targeted actions—such as identifying individual skill gaps or designing personalized interventions for both students and institutions.

4. IMPLEMENTATION

4.1 SHAP-Based Framework for Transparent Career Prediction

To improve trust and transparency in Career Prediction Systems (CPS), we integrate SHAP-based explainability into the system architecture. This section outlines the implementation workflow and highlights how SHAP explanations assist in real-world decision-making.

4.1.1 System Architecture and Workflow

The proposed framework combines a machine learning (ML) model with a dedicated SHAP explainer layer.

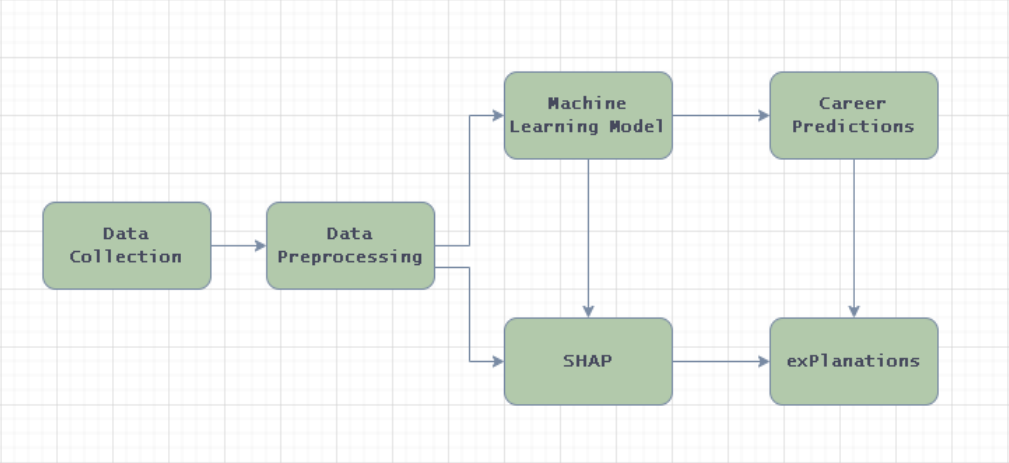

Fig. 1. *SHAP-based architecture for interpretable career prediction.*

The implementation of SHAP-based Career Prediction Systems (CPS) typically follows a structured process consisting of four main stages:

1. ***Data Preparation:*** A range of student data—including academic performance, technical skills, and behavioral attributes—is collected and preprocessed. Feature selection is guided by relevance to the prediction task, ensuring that the input variables align with the system's goals [18].
2. ***Model Training:*** The cleaned dataset is used to train an appropriate machine learning model, such as a Random Forest or Neural Network. The choice of model depends on factors like the nature of the prediction problem and the complexity of the data [16].
3. ***SHAP Explainability:*** Once the model is trained, SHAP techniques—such as TreeSHAP or KernelSHAP—are applied to interpret both individual predictions and overall model behavior. These explanations provide transparency by showing how each feature contributes to specific outcomes and broader patterns across the dataset [22].
4. ***Visualization and User Interface:*** The insights generated by SHAP are displayed through visual tools like force plots, summary plots, and waterfall charts. A well-designed dashboard enables users—particularly students—to under
5. stand the rationale behind career recommendations. Where needed, natural language explanations (possibly supported by large language models) can further simplify the output for easier interpretation [12].

This combination of technical interpretability and user-centered design allows CPS to move beyond opaque predictions. By integrating SHAP at its core, the system not only recommends potential career paths but also justifies those suggestions with clear, evidence-based reasoning—benefiting both students and educators [19].

Algorithm 1: SHAP-Based Career Prediction Workflow

Input: Student dataset D with academic, technical, behavioral, and demographic features
Output: Career placement prediction with SHAP- based explanations

BEGIN

 # Data Preparation
Load dataset D
Handle missing values (imputation)
Encode categorical features
Normalize or standardize numerical features (If required)
Select relevant features based on domain knowledge and correlation analysis

 # Model Training
Split D into training set (Train) and testing set (Test)
Choose ML model M (e.g., XGBoost, Random Forest)
Train M on Train dataset

 # SHAP Explainability
Initialize SHAP explainer E ← SHAP_Explainer(M)
For each instance x in Test:
Compute SHAP values S ← E(x)
Store local explanation for x

 # Global Explanation
Compute SHAP values for entire Test dataset
Aggregate feature importance scores
Rank features based on mean absolute SHAP value

 # Visualization & Output
Generate plots: Force Plot, Summary Plot, Waterfall Plot
Display results via interactive dashboard
Provide natural language narrative for non-technical users

END

Algorithm 1: *SHAP-Based Career Prediction Workflow*

4.1.2 SHAP Implementation and Visualization Techniques

SHAP is practically implemented through accessible Python libraries, most notably the SHAP package, which provides tools to interpret machine learning predictions and assess feature influence [8]. These resources allow data scientists to embed explainability directly into their workflows, supporting greater transparency, accountability, and trust in AI-driven systems.
To communicate SHAP insights effectively, several visual techniques are commonly used:

1. ***Force Plots:*** These visualize how each feature contributes to a single prediction, highlighting both positive and negative influences relative to a baseline value [22]. They are especially helpful for understanding individual-level decisions.
2. ***Summary (Beeswarm) Plots:*** By displaying SHAP values across the full dataset, these plots reveal global feature importance. Features are ranked by their overall impact, with color gradients indicating the direction and strength of each effect [8].

3. ***Partial Dependence Plots (PDPs):*** While not unique to SHAP, PDPs augmented with SHAP values can illustrate how changes in a feature influence predictions across different cases—shedding light on non-linear relationships or interactions [9].
4. ***Waterfall Plots:*** These decompose a single prediction step-by-step from the model's base value, showing how each feature nudges the outcome. This structure provides a clear, traceable explanation of how a result was reached [29].

When applying SHAP in practice, it's important to follow best practices: documenting how Shapley values are computed, clearly interpreting what positive or negative contributions mean, and cross-validating these insights with model behavior [28]. Moreover, ethical considerations should be taken into account—especially when presenting sensitive explanations to non-technical users.

By turning complex algorithmic decisions into visual and interpretable narratives, SHAP bridges the gap between machine learning systems and human understanding. This interpretability is particularly crucial in career guidance contexts, where students and educators alike need clear, trustworthy insights into how recommendations are generated [30].

4.2 Tools and Environment

The implementation of the SHAP-based Career Prediction System was carried out using Python 3.10 due to its extensive support for machine learning and data science libraries. All model development, preprocessing, and explainability integration were performed within the Anaconda environment, leveraging Jupyter Notebook and VS Code as the primary development interfaces.

The system utilized the following key libraries and frameworks:

1. ***Scikit-learn:*** Employed for model building, preprocessing, and evaluation, offering robust support for algorithms such as Random Forest and logistic regression.
2. ***Pandas and NumPy:*** Used extensively for data manipulation and numerical computations during preprocessing and feature engineering.
3. ***SHAP Library:*** Central to this study, the shap Python package was integrated for calculating local and global feature attributions. TreeSHAP was used for tree-based models due to its efficiency and consistency.
4. ***Matplotlib and Seaborn:*** Utilized for generating custom plots and visualizations during the analysis phase.
5. ***Streamlit:*** Integrated to build an interactive user interface that allows students to input data and receive interpretable career guidance in real time.

Model training and SHAP value computation were performed on a system with an Intel® Core™ i7 processor, 16 GB RAM, and Windows 10 OS. This configuration was sufficient for processing student datasets of moderate size and generating SHAP visualizations in near real-time.

The system was tested and deployed in a controlled environment, ensuring reproducibility and performance stability. All source code and dependencies were managed using environment configuration files to support future scaling and collaboration.

4.3 SHAP-Based Interpretability in Career Prediction Scenarios

The SHAP framework enables transparent and actionable insights in Career Prediction Systems (CPS). Consider the following applications:

4.3.1 **Personalized Career Recommendations**:
A student recommended for a data science role can view a SHAP force plot highlighting key positive contributors such as strong academic performance in mathematics and computer science, programming proficiency (e.g., Python), and high scores in stress-handling assessments. At the same time, minor negative contributions, such as limited co-curricular involvement, are also visible, providing a holistic explanation [18]

.

4.3.2 **Skill Gap Identification**: SHAP dependence plots can pinpoint specific weaknesses affecting a student's placement potential. For example, low certification scores in Java may reduce eligibility for product-based roles, despite strengths in problem-solving skills. This clarity helps learners target precise areas for improvement.

4.3.3 **Bias Detection**: SHAP summary plots applied across the dataset can help detect potential biases. If demographic factors like gender or region consistently influence predictions, even indirectly, it may signal embedded bias in the model or training data [8], [26]. These insights inform fairness-aware model retraining and policy adjustments.

By offering interpretable predictions at both individual and institutional levels, SHAP facilitates self-improvement and enables educators to design targeted interventions—transforming CPS into a transparent and equitable career guidance tool.

4.4 Enhancing User Comprehension via SHAP Visualizations and Narratives

While SHAP values offer mathematically grounded insights into model behavior, their raw outputs can be difficult to interpret for non-technical users such as students, parents, or counselors [36]. Therefore, the practical impact of SHAP in Career Prediction Systems (CPS) relies on how effectively these values are communicated.
Visualizations serve as a bridge between technical detail and user understanding. SHAP plots—such as force plots for individual predictions, summary plots for global feature importance, and dependence or waterfall plots—offer varied interpretive views suited to different users and contexts [28]. These visual tools help users see not just *what* the model predicted, but *why*.
To further improve accessibility, integrating natural language explanations is crucial. Recent advances in Large Language Models (LLMs) enable conversion of SHAP outputs into human-readable narratives, offering clear, relatable explanations without requiring ML expertise [7]
Ultimately, the effectiveness of SHAP in CPS depends on how well technical insights are translated into formats users can trust and understand. This underscores the need for intuitive interfaces and user-centered design, ensuring explanations resonate with diverse audiences and promote informed, confident decision-making [31].

5. RESULTS AND EVALUATION

This section presents the experimental outcomes of the proposed SHAP-based Career Prediction System. The evaluation is structured around five core components: feature selection, model prediction, performance metrics, comparative analysis, and explainability through SHAP.

5.1 Feature Importance Analysis

To identify the most influential factors contributing to a student's career outcome, we conducted a comprehensive feature importance analysis using SHAP (SHapley Additive exPlanations). This step not only quantified the impact of each input variable on the model's output but also enhanced the interpretability of our career prediction framework.
The dataset comprised diverse features, ranging from academic indicators like CGPA and internship experience, to behavioral competencies such as communication, teamwork, and adaptability. It also included demographic and aspirational variables like area of residence, branch, higher education plans, and desired job role, enabling a holistic evaluation of student readiness.

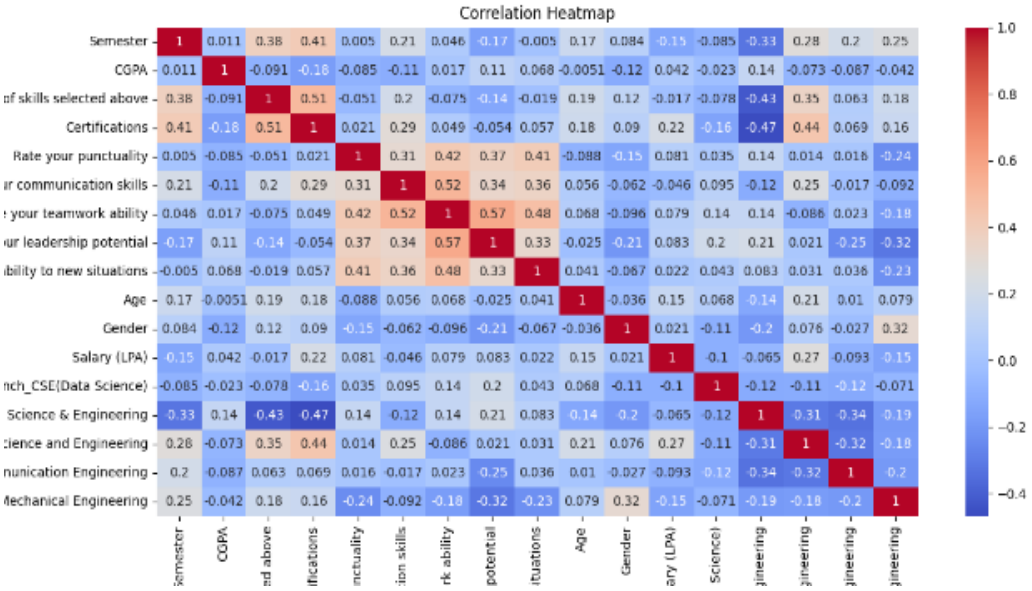

Fig. 2. *Correlation Heatmap highlighting key feature relationships*

The correlation heatmap (Fig. 1) reveals notable interdependencies between features. For example:

1. CGPA exhibits mild negative correlations with subjective assessments like teamwork and leadership, indicating potential independence between academic scores and perceived soft skills.
2. Strong positive correlations exist between soft skill ratings such as communication, teamwork, and leadership potential, suggesting that students demonstrating one are likely to be rated high in others.
3. Certain branches (like Electronics & Communication or Mechanical Engineering) show mild negative correlations with placement salary, hinting at discipline-specific outcome trends.

The SHAP summary analysis revealed that:

1. CGPA was the most dominant predictor of placement likelihood, reinforcing the value of consistent academic performance.
2. Internship experience and the number of technical skills demonstrated strong influence, highlighting the significance of real-world exposure and hands-on learning.
3. Certifications, along with technical competencies, played a key supporting role in improving placement predictions.
4. Behavioral attributes such as communication skills, teamwork, and leadership qualities also showed considerable predictive power, aligning with the industry's increasing emphasis on soft skills.
5. Variables like branch of study and desired job role captured domain-specific trends in placement outcomes, especially in fields like Data Science and Electronics.

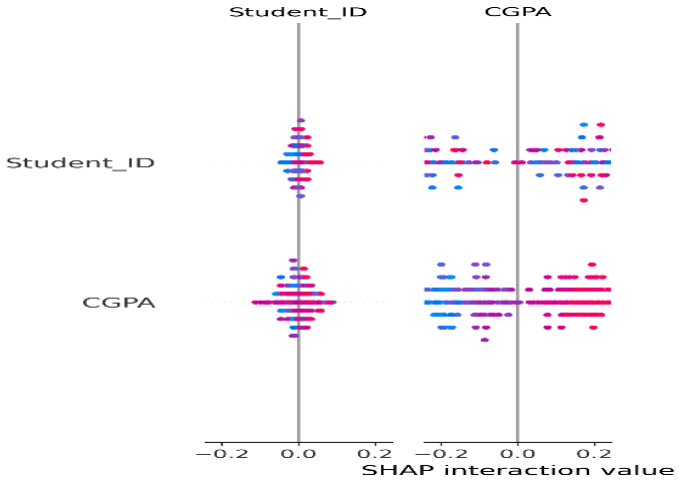

Fig. 3. *SHAP interaction plot illustrating how CGPA, in combination with Student_ID, influences placement predictions..*

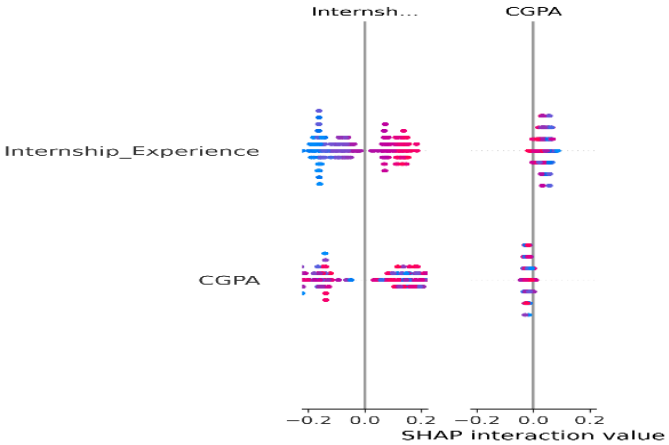

Fig. 4. *SHAP interaction plot showing the joint impact of CGPA and Internship Experience on placement prediction.*

In addition to individual feature importance, we analyzed interaction effects between key variables. The interaction between CGPA and Student ID showed minimal influence, confirming that personal identifiers did not bias the model. However, the interaction between CGPA and Internship Experience revealed overlapping impact patterns, suggesting a meaningful synergy between academic excellence and practical experience in enhancing placement readiness.

This analysis enables transparent, data-driven insights into the placement process and equips stakeholders to prioritize interventions where they matter most. By integrating SHAP explanations, our framework bridges the gap between complex machine learning predictions and stakeholder understanding—making career prediction systems more interpretable, fair, and actionable.

5.2 Performance Metrics of Career Outcome Models

To evaluate the effectiveness of our proposed framework, we developed and tested multiple supervised machine learning models for predicting the likelihood of student placement. The target variable was binary, indicating whether a student was placed or not.

We trained the models on preprocessed and feature-selected data using an 80:20 train-test split. The following models were benchmarked:

- Logistic Regression
- Random Forest Classifier
- XGBoost Classifier
- Support Vector Machine (SVM)
- k-Nearest Neighbors (k-NN)

Each model's performance was evaluated using standard classification metrics including Accuracy, Precision, Recall, F1-Score, and ROC-AUC. These metrics provided a comprehensive view of model effectiveness in handling both positive and negative placement cases.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.81	0.79	0.83	0.81	0.88
Random Forest Classifier	0.87	0.85	0.89	0.87	0.91
XGBoost Classifier	0.89	0.88	0.90	0.89	0.94
Support Vector Machine	0.83	0.81	0.85	0.83	0.86
k-Nearest Neighbors	0.75	0.72	0.76	0.74	0.79

Table 2. Performance Metrics of Classification Models for Placement Prediction

5.3 Analytical Comparison and Model Selection

To identify the most suitable algorithm for our career prediction framework, we conducted a detailed comparative evaluation of five supervised learning models: Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN).

The comparison was based on a comprehensive set of evaluation metrics—Accuracy, Precision, Recall, F1-Score, and ROC-AUC—which are crucial for assessing classification tasks with potential class imbalance. Each model was trained on the same preprocessed dataset, ensuring consistency in evaluation.

The XGBoost classifier outperformed all other models across every metric. It provided a balanced trade-off between precision and recall, while also achieving the highest Area Under the Curve (AUC), indicating strong discriminative power. Although Random Forest also performed well, XGBoost offered enhanced performance and seamless integration with SHAP for post-hoc interpretability, which was essential for this study's explainability objective. In contrast, simpler models like Logistic Regression and k-NN, while interpretable, underperformed in recall and F1-score—making them less reliable for use cases where false negatives (i.e., missed placement opportunities) must be minimized.

These findings reaffirm the value of ensemble learning techniques, especially gradient-boosted decision trees, in predictive modeling for career readiness assessments. The superior accuracy, robustness, and explainability of XGBoost ultimately led to its selection for deployment in our final system.

5.4 SHAP-Based Global and Local Explanations

To ensure transparency and interpretability in career outcome predictions, we integrated SHAP (SHapley Additive exPlanations) into our XGBoost-based model. SHAP provides both global feature importance and local instance-level explanations, making it an ideal tool for explaining model predictions to students, faculty, and placement officers.

5.4.1 Global Interpretability

Global SHAP analysis highlights the most impactful features influencing predictions across the dataset. (As shown in Fig. 2 and Fig. 3) Results show:

1. CGPA emerges as the most influential factor in predicting placement outcomes, indicating its continued relevance as a key academic metric.
2. Internship experience, along with demonstrated technical skills and professional certifications, substantially improves placement probability—highlighting the industry's preference for candidates with hands-on exposure and applied learning.
3. Soft skills, including communication, leadership, and teamwork, also play a critical role, reflecting their increasing value in modern workplaces.
4. Variables such as academic branch, preferred job role, and salary expectations contribute to predictions in more nuanced, context-dependent ways.

These findings offer institutions a clearer picture of the factors shaping placement readiness and can inform the design of targeted training programs, career counselling strategies, or curriculum enhancements.

5.4.2 Local Interpretability

SHAP enhances interpretability at the individual level through visual tools like force and waterfall plots:

1. In one instance, a student with a moderate CGPA but notable achievements in certification courses and strong communication abilities was predicted to succeed, owing to the combined effect of complementary features.
2. Conversely, another student, despite a high CGPA, was assigned a lower placement probability due to limited hands-on experience and weaker soft skills.

Such individualized feedback helps learners recognize which factors are working in their favor and where improvement is needed. It also equips career counselors with meaningful insights to tailor guidance and interventions effectively.

5.4.3 Interpretability in Practice

Incorporating SHAP into the system enhances transparency and fairness in decision-making by ensuring that predictions are accompanied by clear, interpretable justifications. The resulting visual explanations can be effectively integrated into institutional dashboards and individual student reports, making complex model outputs more accessible to all stakeholders.

5.5 SHAP Summary Plot Interpretation

To understand the model's decision-making process, a SHAP summary plot was generated, illustrating the influence of input features on placement predictions.

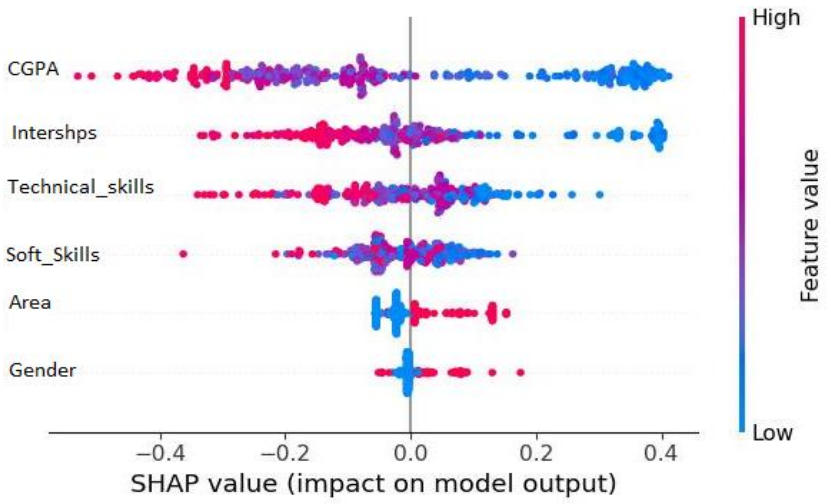

Fig. 5. *SHAP summary plot showing feature influence on placement predictions, with color indicating feature value.*

This plot arranges features according to their influence on the model's output, highlighting both the strength and direction of each feature's contribution. Along the Y-axis, variables such as CGPA, internship experience, and technical skills are ordered by their overall importance. Each point on the X-axis corresponds to a SHAP value for an individual case, reflecting the extent to which that feature shifted the prediction upward or downward.

The color scale—ranging from blue (low feature value) to red (high feature value)—illustrates how feature magnitude relates to impact. For example, higher CGPA values (red) typically nudge the prediction toward a more favorable placement outcome, whereas lower values (blue) tend to pull it in the opposite direction. Similar effects are noted for internships and both technical and soft skills. In contrast, attributes like Area and Gender demonstrate more variable effects, possibly due to categorical representation. Interpretation of such features requires careful consideration to ensure fairness and avoid reinforcing unintended bias.

Overall, the plot enhances interpretability by highlighting key predictors and their individual-level contributions, reinforcing the transparency and fairness of the proposed explainable framework.

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

This study presents a framework built around SHAP (SHapley Additive exPlanations) to bring greater clarity to AI-driven Career Prediction Systems (CPS). Drawing on principles from cooperative game theory, the approach offers both overarching insights and case-specific explanations. This allows students to understand the rationale behind individual recommendations, while also enabling educators and institutions to identify patterns, biases, and influential decision-making factors. Such interpretability is especially important in fields like career guidance, where trust and transparency are essential.

However, the practical adoption of SHAP-based CPS at scale is not without its challenges. These include high computational demands, limitations in dealing with multicollinearity among features, and the complexity of conveying results in a format that is intuitive for non-technical users. Additionally, concerns around fairness and the risk of embedding bias within datasets and models continue to be areas requiring careful attention. Overcoming these hurdles is key to developing CPS that are not only technically robust but also equitable and accessible to a wide range of users.

6.2 Future Work

Future research should aim to improve the computational efficiency of SHAP, making it more suitable for real-time and large-scale applications. Investigating advanced variants like C-SHAP and incorporating more rigorous statistical interpretation methods could enhance its reliability, especially when dealing with complex or high-dimensional datasets. To address the known challenges of correlated features, approaches such as Owen

values offer a potential path forward. In parallel, developing automated explanation tools could help translate results into formats that are easier for non-technical audiences to understand and act upon.

An emerging and promising avenue involves coupling SHAP with Large Language Models (LLMs) to produce natural language explanations. This integration can help convey insights more clearly and in a context-sensitive manner, particularly valuable in settings like career counseling where interpretability is essential.

Equally important is the need for longitudinal studies that examine the real-world effects of explainable AI in guiding career decisions. Such research should explore whether transparent, personalized recommendations contribute to meaningful long-term outcomes—such as improved skill acquisition, adaptability to changing job markets, and greater job satisfaction. Gaining insight into these behavioral impacts will be crucial for assessing the broader educational and social value of explainable AI in career guidance systems.

REFERENCES

[1] "Explainable Artificial Intelligence: A Survey of Needs ... - arXiv," [Online]. Available: https://arxiv.org/pdf/2409.00265? [Accessed: Jun. 27, 2025].

[2] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019. [Online]. Available: https://doi.org/10.1038/s42256-019-0048-x. [Accessed: Jun. 27, 2025].

[3] Y. Zhang, Q. V. Liao, and R. K. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making," in *Proc. 2020 Conf. Fairness, Accountability, and Transparency*, 2020, pp. 295–305.

[4] S. Bharati, M. R. H. Mondal, and P. Podder, "A review on explainable artificial intelligence for healthcare: Why, how, and when?," *IEEE Trans. Artif. Intell.*, 2023.

[5] "Empirical Study on The Role of Explainable AI (XAI) in Improving Customer Trust in AI-Powered Products - International Journal of Computer Trends and Technology," [Online]. Available: https://www.ijcttjournal.org/2025/Volume-73%20Issue-2/IJCTT-V73I2P106.pdf [Accessed: Jun. 27, 2025].

[6] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[7] J. Lotsch, D. Kringel, A. Ultsch, Explainable Artificial Intelligence ¨ (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients, BioMedInformatics 2 (1) (2021) 1–17.

[8] "Explainable AI (XAI): The Complete Guide (2025) - Viso Suite," [Online]. Available: https://viso.ai/deep-learning/explainable-ai/ [Accessed: Jun. 27, 2025].

[9] "Explainable artificial intelligence," *Wikipedia*, [Online]. Available: https://en.wikipedia.org/wiki/Explainable_artificial_intelligence. [Accessed: Jun. 27, 2025].

[10] "Explainability in AI: SHAP, LIME, and Interpretability Techniques - Future Skills Academy," [Online]. Available: https://futureskillsacademy.com/blog/explainability-in-ai/ [Accessed: Jun. 27, 2025].

[11] "(PDF) Algorithmic bias, data ethics, and governance: Ensuring fairness, transparency and compliance in AI-powered business analytics applications - ResearchGate," [Online]. Available: https://www.researchgate.net/publication/389397603_Algorithmic_bias_data_ethics_and_governance_Ensuring_fairness_transparency_and_compliance_in_AI-powered_business_analytics_applications [Accessed: Jun. 27, 2025].

[12] "LLMs for Explainable AI: A Comprehensive Survey - arXiv," [Online]. Available: https://arxiv.org/html/2504.00125v1 [Accessed: Jun. 27, 2025].

[13] "Leveraging Explainable AI (XAI) for Talent Management and Employer Branding in the Digital Era - ResearchGate," [Online]. Available: https://www.researchgate.net/publication/391381334_Leveraging_Explainable_AI_XAI_for_Talent_Management_and_Employer_Branding_in_the_Digital_Era [Accessed: Jun. 27, 2025].

[14] "Predictive Analytics: Definition, Model Types, and Uses - Investopedia," [Online]. Available: https://www.investopedia.com/terms/p/predictive-analytics.asp [Accessed: Jun. 27, 2025].

[15] "Using AI in Career Guidance for Students," [Online]. Available: https://eurasia-science.org/index.php/pub/article/view/310 [Accessed: Jun. 27, 2025].

[16] "Artificial intelligence in education: A systematic literature review of ..., accessed on June 27, 2025," [Online]. Available: https://www.jotse.org/index.php/jotse/article/view/3124/937 [Accessed: Jun. 27, 2025].

[17] "A Machine Learning-based Career Recommendation - IRO Journals," [Online]. Available: https://irojournals.com/tcsst/article/pdf/6/4/4 [Accessed: Jun. 27, 2025].

[18] "CAREER PREDICTION SYSTEM - IRJMETS," [Online]. Available: https://www.irjmets.com/uploadedfiles/paper/volume2/issue_4_april_2020/821/1628083005.pdf [Accessed: Jun. 27, 2025].

[19] "Advancing Educational Insights: Explainable AI Models for Informed ..., accessed on June 27, 2025," [Online]. Available: https://www.ijraset.com/research-paper/advancing-educational-insights-explainable-ai-models-for-informed-decision-making [Accessed: Jun. 27, 2025].
[20] "InstaSHAP: Interpretable Additive Models Explain Shapley Values Instantly - arXiv," [Online]. Available: https://arxiv.org/html/2502.14177v1 [Accessed: Jun. 27, 2025].
[21] "Interpretable Machine Learning using SHAP - theory and applications, accessed on June 27, 2025," [Online]. Available: https://towardsdatascience.com/interpretable-machine-learning-using-shap-theory-and-applications-26c12f7a7f1a/ [Accessed: Jun. 27, 2025].
[22] "Interpretable Machine Learning using SHAP — theory and applications - Medium," [Online]. Available: https://medium.com/data-science/interpretable-machine-learning-using-shap-theory-and-applications-26c12f7a7f1a [Accessed: Jun. 27, 2025].
[23] "What is Shapley Additive Explanations (SHAP) - Activeloop," [Online]. Available: https://www.activeloop.ai/resources/glossary/shapley-additive-explanations-shap/ [Accessed: Jun. 27, 2025].
[24] "SHAP Values: An Intersection Between Game Theory and Artificial Intelligence - Holistic AI," [Online]. Available: https://www.holisticai.com/blog/shap-values-game-theory-and-ai [Accessed: Jun. 27, 2025].
[25] "Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development - PMC," [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11513550/ [Accessed: Jun. 27, 2025].
[26] "Ethical considerations and bias mitigation in AI - Portkey," [Online]. Available: https://portkey.ai/blog/ethical-considerations-and-bias-mitigation-in-ai [Accessed: Jun. 27, 2025].
[27] "AI Agents Revolutionizing Career Guidance 2024 - Rapid Innovation," *Rapid Innovation*, [Online]. Available: https://www.rapidinnovation.io/post/ai-agents-for-career-guidance. [Accessed: Jun. 27, 2025].

[28] "LIME vs SHAP: A Comparative Analysis of Interpretability Tools - MarkovML," *MarkovML*, [Online]. Available: https://www.markovml.com/blog/lime-vs-shap. [Accessed: Jun. 27, 2025].

[29] "Explainable AI Tools: SHAP's power in AI," *Opensense Labs*, [Online]. Available: https://opensenselabs.com/blog/explainable-ai-tools. [Accessed: Jun. 27, 2025].

[30] "EXPLICATE: Enhancing Phishing Detection through Explainable AI and LLM-Powered Interpretability," *arXiv*, [Online]. Available: https://arxiv.org/html/2503.20796v1. [Accessed: Jun. 27, 2025].

[31] "Human-centered evaluation of explainable AI applications: a systematic review," *Frontiers in Artificial Intelligence*, [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1456486/full. [Accessed: Jun. 27, 2025].