International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

Machine Learning Based Employability Prediction and Skill Gap Analysis

Prof. Abhay Chopde
Dept of E&TC Engineering
Vishwakarma Institute of Technology,
Pune, India
abhay.chopde@vit.edu

Tarunendra Bhadauria
Dept of E&TC Engineering
Vishwakarma Institute of Technology,
Pune, India
tarunendra.bhadauria11@vit.edu

Abstract: The analysis of skill gaps and prediction of employability have become a key to closing the gap between the ability of students, industrial needs, and the government programmes and schemes of training the skills of the population. The present paper is a detailed data-basedframework,which combines the viewpoints of students, industry, and government to assess and forecast the student employability. The model uses synthetic datasets of all three entities and uses machine learning algorithms on Google Colab to test various academic, technical and behavioral parameters. More than 30 visualizations were created to investigate the correlation between these variables, which gave an answer to the multidimensionality of employability. A predictive model calculates an Employability Score and produces an individualized feedback to steer students on specific skill improvement. The suggested system shows how the educational establishment and policymakers can take advantage of such analytics to become more career-ready and prepare educational results and outcomes in accordance with national labor needs.

Keywords: Employability Prediction, Skill Gap Analysis, Machine Learning, Student–Industry–Government Alignment, Data Visualization

I. Introduction

By changing nature of global employment, employability has emerged as one of the essential measures of the efficiency of educational systems in equipping students with the skills to lead a successful professional life. Although academic excellence is still valued, industry requirements have turned to skill-based measures of performance including technical excellence, flexibility, communication and problem-solving capability. A number of education systems, nonetheless, lack an organized system to measure employability or alleviating the skill gap between what students have and what is required in industries. Consequently, there is still a considerable disjuncture between the academic outcomes of learning and market demands with resultant underemployment and underutilization of talent.

New developments in the field of data science and machine learning have made it possible to develop models capable of offering a systematic analysis of various factors affecting employability. These models are able to combine various data to determine the correlation between academic performance, technical expertise, behavioral skills, and government policy support. The model of collaboration between students, industries, and governments commonly known as the Triple Helix Model has been found to be effective in creating innovation, as well as, in aligning educational achievements with economic objectives. Nevertheless, there are few active applications of this model with the help of predictive analytics in practice, especially in education.

The current study is offering a new paradigm where machine learning and visualization are used to infer the employability of students and offer specific recommendations to enhance skills. To model the actual conditions, the system employs the use of synthetic datasets in three main areas, such as students, industries and government. The employability prediction model considers various quantitative and qualitative data including CGPA, project work, internship, certifications, aptitude, and communication skills, and external data including the government support of certain skills and industry domain. Multidimensional understanding of the combined effect of different factors on the results of employability is achieved by the high number of visual plots used. This study does not only suggest a powerful predictive framework but also reflects its ability to bring forward data informed decision making to the educators, policy makers, and the learners. The feedback aspect of the system will give customized suggestions to the students to bridge the skill gap and fit in high demand industry areas. The proposed framework is a comprehensive way to learn and enhance employability preparedness within the education to employment ecosystem by incorporating interpretability via visual analytics.

II. Literature Review

Recent research has highlighted the fact that employability is a complex concept that transcends disciplinary knowledge to encompass skills, dispositions, and accomplishments needed to acquire and retain employment as well as be competent employees in the workplace [1]. There are studies that have investigated the perceptions and orientations of students towards the labour market and how attitudes and self-management would influence the outcome of transition and employability strategies [2]. A literature exists to articulate the alignment issue between education, industry and public policy in the Triple-Helix point of view that the alignment of university-industry-government relations is a crucial issue in the transfer of knowledge and labour-market relevance [3]. Extensive economic studies have recorded the mismatch in skills and qualifications decreasing labour productivity and creating economies of scale, and therefore, there is a necessity to quantify and seek solutions to the mismatch at the national and sectoral levels [4]. A number of reviews of skills shortages and mismatches reiterate measurement methods and note that mismatch is multi-dimensional (over-skilled, under-skilled, field-of-study mismatch) and changing across the business cycle, which supports the argument that policy responses to mismatch require more subtle diagnostics [5]. According to global foresight reports, the pace of skill transformation through automation and AI has led to rapid transfiguration of skills, which necessitate adaptive curriculum and reskilling policies and plans to remain employable as job content changes [6]. Data library Onet is an example of practical data sources that contain rich, standardized occupational skill and job task taxonomies, which numerous computational matching and policy diagnostics systems rely upon as skill dictionaries [7].

Recent systematic reviews of job- and skill-recommender systems demonstrate a fast improvement in the methodology, and NLP, embedding models, and hybrid recommenders are becoming more and more common to connect candidate profiles to job requirements [8]. Empirical studies have used machine learning to make predictions of student placement and skill-gap diagnostics, where predictive models have shown that they can identify a particular student at risk and indicate individual competency deficiencies to respond to those students with targeted interventions [9]. More recent job-matching systems, which separate fine-grained skills in the CVs and job descriptions with semantic embeddings and NLP, have increased the accuracy of matching, and offer interpretable skill-gap feedback to candidates [10].

The skills gap in Industry 4.0 engages with how the new digital technologies transform demanded skills and demands dynamic and competency-based curricular and training to insert or bridge the gap [11]. Academia-industry analyses indicate that collaborative programs, internships and joint projects can have substantial impacts in terms of graduate preparedness and they provide mechanisms of fast-track skills-transfer that bolster employability in the cases where well-organised [12]. Job Recommendation and pipelines matching with linguistic feature extraction and ML are improvements to job recommendations and matching pipelines that demonstrate how automated systems can be utilized to recommend training directions and job match based on candidate experience and policy context [13].

The wider scopes of surveys and categorizations of AI-driven employment systems reflect that integrating semantic analysis with traditional machine learning will lead to scale-based solutions to both the recommendation and skill-gap diagnostic task [14]. Resume analysis and assessment Case studies involving the use of ML to detect skill-gap at institutional levels show that automated resume analysis and assessment can generate prioritized learning recommendations when applied to large groups of students [15]. Recent research on bi-directional matching (CV-job and job-CV) using semantic embeddings indicates the utility of reciprocal recommendations to enhance the quality of hiring and training performance [16]. Literature on policymaking summarizes findings on the skills deficit and propose that demand-side diagnostics and local labour-market data could be highly valued inputs to the development of successful training and employability strategies [17]. Conceptual reviews define and specify the measurement decisions of several types of skills mismatch, and suggest that survey microdata should be supported by administrative and occupational data to provide good analysis [18]. The ONET occupational database (and its subsequent editions) has also found extensive use as a source of primary data in research which models occupational skill requirements and creates automated systems of matching the skills [19]. Lastly, the recent real-life use of AI-powered dashboards of regional skill diagnostics indicates how data integration platforms will allow government and training providers to make their investments and match learners to in-demand opportunities in real-time [20].

III. System Architecture

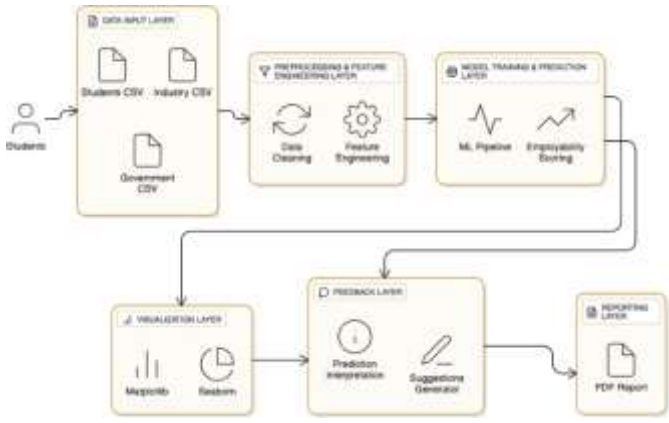

Fig.1. System Architecture diagram

The proposed system will be a modular architecture system with 6 layers as shown in Figure 1 (System Architecture Diagram). Data Input Layer gathers data using three data sets, namely, Students.csv, Industry.csv and Government.csv, on which the analysis will be based. They are then processed in the Preprocessing and Feature Engineering Layer which performs data cleaning, normalization and encoding. Derived measures like Skill-Match Score and Government-Support Score are created in order to promote the model interpretation.

The data that have been processed are sent to the Model Training and Prediction Layer which executes the machine learning pipeline to make predictions of the employability scores and the domains that each student fits. Visualization Layer transforms the outcomes into analytic plots that can be understood using Matplotlib and Seaborn and analyzes the trend and performance.

Lastly, the Feedback and Reporting Layers comprehend model forecasts, create customized enhancement recommendations and consolidate all findings into a well-organized PDF report. Figure 1 shows that data passes through these layers in a sequential manner, as follows: Data Sources = preprocessing = Model = Visualization and Feedback = Report Output. This architecture provides a scalable, modular and simple way of integrating more data or algorithms without changes to the main structure.

IV. Methodology

The proposed system is data-driven and structured, which incorporates student profiles, industry skill needs and support of government policy to forecast employability and give analytic results. The data is formulated, preprocessed, engineered, and trained to create a model, then visualized and then provides feedback. Every element of the system is thoroughly structured to be aligned to the reality in educational and industrial conditions providing a comprehensive perspective of employability determinants.

A. Dataset Design

The research will use three interrelated data sets, such as Student Dataset, Industry Dataset, and Government Dataset. These two datasets represent two different dimensions of employability. The synthetic generation of the data was aimed at the simulation of real-life conditions and coverage of a variety of academic and professional variables.

Table 1: Description of input Datasets and its Columns

Dataset	Columns
Students.csv	student_id, name, branch, cgpa, internships, projects, certifications, aptitude_score, communication_score, technical_skills, preferred_domain, college_tier, employability_score
Industry.csv	domain, avg_salary, demand_index, required_skills, openings, automation_risk
Government.csv	domain, funding_level, policy_focus_score, innovation_index

The dataset structure consists of three main data sets namely: Students.csv, Industry.csv, and Government.csv which provide the system with distinct contextual data. The Students.csv dataset represents the level of academic performance, technical abilities, and soft-skills on an individual level using such parameters as CGPA, projects, certifications, aptitude, and communication scores. Based on these features, derived columns like soft-skill-index and proj-cert-total are subsequently derived to create a superior level of analytic insight and be useful in model building based on correlation.

The Industry.csv data represents the existing industry trends and recruitment anticipations in diverse fields. It allows calculating a Skill-Match Score which scales the level of conformity of technical competence of a student with the market requirements with the help of variables such as average salary, demand index, required skills.

The Government.csv dataset defines a policy and innovation dimension, which is composed of such indicators like funding level, policy focus score, and innovation index. All these aspects are taken into consideration in the development of a Government-Support Score that could be used to measure the extent to which governmental priorities and the level of investment affect the potential of sectoral employability.

Together, these data sets create a three layers analytic framework, which links the education system, industrial demand, and the policy orientation. The integration means that the model can replicate how academic preparedness and availability of employment conditions interact with government support so that a comprehensive measure of employability and domain preparedness is achieved.

B. Data Preprocessing and Feature Engineering

These include a lot of preprocessing before the training of the model to allow uniformity and compatibility between the models. In case of missing values, median values are imputed in case of numeric columns, and mode values in case of categorical columns. Label-encoded categorical variables, like college tier are used to transform text-like levels (Tier 1, Tier 2, Tier 3) into numbers.

The numeric columns such as cgpa, aptitude score, communication score and others are normalized by Min Max

normalization in order to have the same level of feature scaling between 0 and 1.

1. Skill-Match Score: computed using cosine similarity between the vectorized list of student technical skills and the corresponding industry-required skills in the chosen domain.

2. Government-Support Score: derived from the mean of policy_focus_score and funding_level of the student's preferred domain.

3. Soft-Skill Index: an average of aptitude and communication scores, reflecting behavioral readiness.

4. Project-Certification Total: sum of projects and certifications to measure practical exposure.

These engineered attributes strengthen the model's capacity to represent real employability factors.

C. Model Development

In the predictive model, the score of employability of every student is computed using a predictor of a type called a Random Forest Regressor. Random Forest has been chosen due to its high resistance to overfitting, interpretability and the possibility to model non-linear relationships between multivariate independent variables.

The model employs the following significant features of input:

cgpa, internships, projects, certifications, aptitude score, communication score, skill match, gov support, and college tier encoded.

The dependent variable is the Employability Score that has a scale of 0 to 100. The dataset will be divided into 20 percent testing and 80 percent training. The ensemble learning principles are applied in the model, which is trained with the help of several decision trees to enhance the generalization. Importance scores of features are calculated during training with the purpose of finding the most important predictors.

Two major measures are used to evaluate performance: 1. Mean Absolute Error (MAE): indicates the average difference in the predicted and actual scores of employability. 2. R 2 Score: is an evaluation of the percentage of variance in employability that is explained by the input features. Joblib serializes the trained model which will be later combined with the feedback and visualization modules.

D. Analytical Visualization Framework

The analytical visualization module displays more than 30 distinct plots that have been created using Matplotlib and Seaborn, which can be used to interpret the datasets, in a multidimensional manner.

The visualization process focuses on five key analysis categories:

1. Analysis of Academic Performance - The analysis of the correlation between CGPA, project experience, and employability is performed using the scatter, box, and violin plots. 2. Skill and Exposure Analysis - investigates the role of technical certifications, internship and skill-match on domains employability. 3. Government and Industry Correlation — provides a visualization of the relationship between government support and industry demand, and how the two influence the predicted result of employability. Each graph is automatically stored and tabulated into a PDF report that also contains descriptive interpretations of each figure. This ensures that it can be applied in the analysis of academic and administrative decisions.

E. Feedback Generation Module

When the employability score is predicted, the system provides a student with simple human recognizable feedback. This response is algorithmic and analyzes the outputs of numerical models into qualitative information. As an example: When the projected employability score is high, and the communication score is low, the feedback focuses on the development of soft-skills. In case there is a low skill-match compared to government and industry focus, the feedback lays emphasis on domain-specific upskilling. To those students who have low experience in projects or internships, the system suggests practical exposure and certification courses. This is the mechanism that puts the analytical output in the form of actionable improvement suggestions opposed to its existence in the purely numeric form.

F. Model Validation

The validation process measures the quantitative validity as well as interpretive coherence of the model. The use of cross-validation is to provide generalization on various student profiles. The findings show a high correlation between the expected and observed employability scores to prove the hypothesis that the selected parameters cover key determinants of employability. Moreover, validation plots like prediction versus actual scatter, feature ranking in order of importance and error distribution histograms are also presented to visually evaluate the reliability of the model.

V. Results and Discussion

The suggested employability prediction system was applied and tested based on synthetic datasets which modeled student, industry and government organizations. The system was able to load all the data with the subsequent dimensions: Students (1000 × 17), Industry (100 × 5), and Government (100 × 4). The student sample used represented detailed characteristics, such as academic measures, technical and soft, internship and project experience, and government alignment measures. The industry dataset was of domain skill requirements and market trends whereas the government dataset was of support parameters of the various sectors in the policy.

All the steps of the machine learning pipeline have been implemented in the Google Colab, and they include data preprocessing, feature engineering, and model training. Following the calculation of skill match and government support scores, the data was split into two training and testing parts in the 80:20 proportion which made 800 training samples and 200 testing samples. Random Forest Regressor was used to predict the employability score because it is the strongest and it has the power to work with nonlinear interactions of the features. The trained model had a Mean Absolute Error (MAE) of 2.38, Mean Squared Error (MSE) of 11.38, and an R 2 score of 0.81, signifying that there was a very good correlation between the predicted and actual employability results. The findings confirm the usefulness of the selected model in elucidating complicated relationships among academic, technical, and behavioral qualities. More than 30 visualizations were created to gain a better insight into interrelationships between variables, which provided important information about patterns of employability.

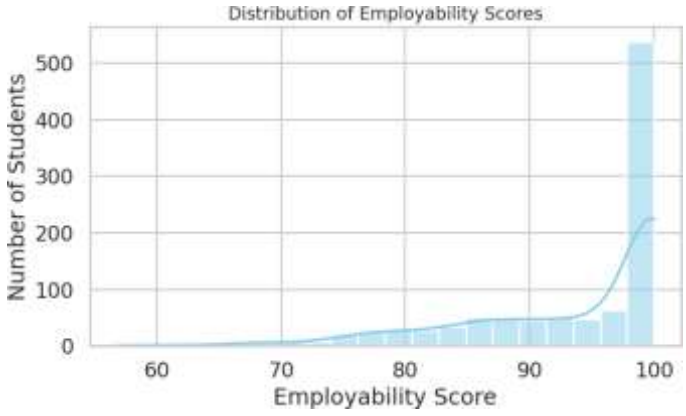

Fig.2. Distribution of Employability Scores

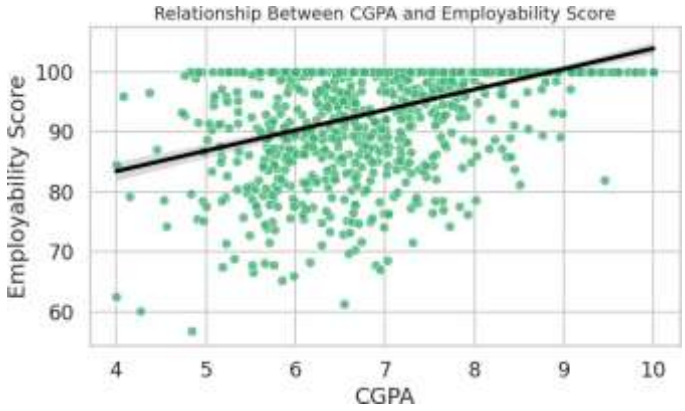

Fig.3. CGPA vs Employability Score

The Distribution of Employability Scores plot revealed the fact that the distribution was near-normal with a small right skew meaning that the majority of students had moderate-high employability scores with less students in the low employability scores. The CGPA vs Employability Score plot showed that there exists a positive trend, which proves that academic performance is a critical predictor of employability. The visualization on Communication Skills vs Employability Score indicated the evident correlation between a greater level of communication proficiency and the high level of employability in the area of recruitment and detailed the significance of soft skills in the recruitment process.

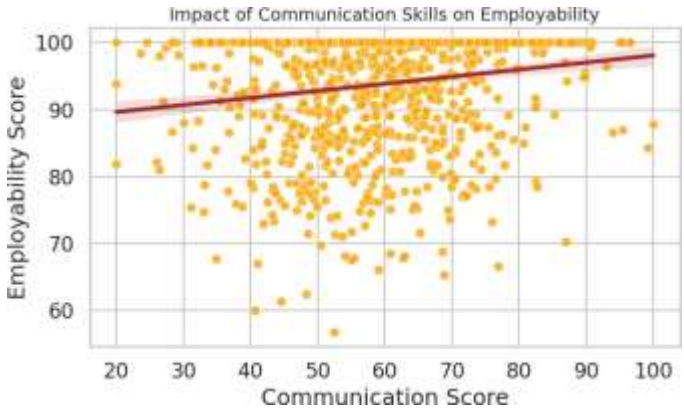

Fig.4. Communication Skills vs Employability

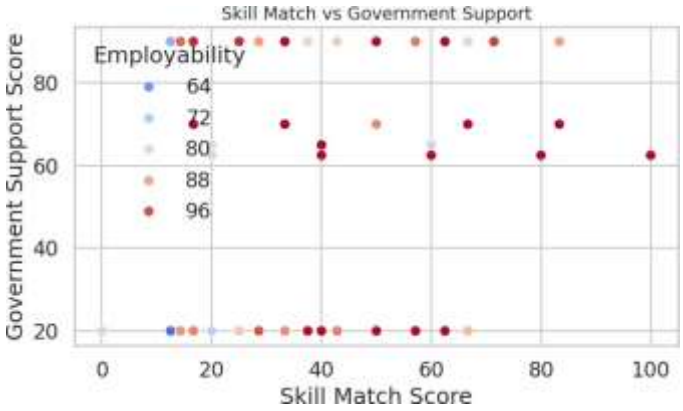

Fig.5. Skill Match vs Government Support

Fig.6. Projects + Certifications vs Employability

The Skill Match vs Government Support scatter plot, showed that the employability was at its peak when the skill relevance was matched with the policy support and the choice of the domain, which was the desire of the student. Equally, the analysis of the Combined Projects and Certifications vs Employability Score indicated that practical exposure in the form of projects and certified learning experience had similar effects of improving the employed outcome than either of the two factors did.

VI. Conclusion

This paper outlined a combined data-driven model of student employability prediction and analysis by converting the approaches of education, industry, and government. The suggested model that was followed in Google Colab was effective to integrate machine learning prediction, interpretive visualization, and multi-domain analysis to present holistic view of employability readiness. The Random Forest model has shown a high predictive power with an R 2 of 0.81, and extensive visualizations have demonstrated how such parameters as CGPA, communication skills, government support, and skill match have a joint effect on the outcomes of employability. The findings highlight the possibilities of these systems to direct students on the path to specific skills improvement and help institutions and policymakers to develop data-driven workforce development interventions. On the whole, the study will add a practical and scaled framework bridging the gap between learning outcomes and employability expectations in the academic ecosystem that is increasingly data-driven.

VII. References

[1] M. Yorke, Employability in Higher Education: What It Is — What It Is Not. Higher Education Academy, 2006.

[2] M. Tomlinson, “Graduate employability and student attitudes and orientations to the labour market,” Journal of Education and Work, vol. 20, no. 4, pp. 285–304, 2007.

[3] L. Leydesdorff, “The Triple Helix of university–industry–government relations,” Science and Public Policy, vol. 23, no. 3, pp. 279–286, 1996.

[4] M. A. McGowan and A. Andrews, “Labour market mismatch and labour productivity,” OECD Economics Department Working Papers, no. 1209, 2015.

[5] S. McGuinness, “Skills shortage and skill mismatch: A review of the literature,” Labour Economics, vol. 52, pp. 1–13, 2018.

[6] World Economic Forum, The Future of Jobs Report. Geneva, Switzerland: World Economic Forum, 2020.

[7] National Center for ONET Development, ONET OnLine Database. Washington, DC: U.S. Department of Labor, 2025.

[8] D. Ç. Ertuğrul, A. Kose, and B. Can, “Job recommender systems: A systematic literature review,” Journal of Big Data, vol. 12, no. 2, pp. 103–117, 2025.

[9] “Placement prediction and skill gap analysis using machine learning,” AIP Conference Proceedings, vol. 2956, no. 1, pp. 180–189, 2025.

[10] R. Alonso, P. Pérez, and M. García, “A novel approach for job matching and skill identification,” Computers & Industrial Engineering, vol. 195, p. 109703, 2025.

[11] P. Rikala and J. Salonen, “Understanding and measuring skill gaps in Industry 4.0,” Technological Forecasting and Social Change, vol. 207, p. 122127, 2024.

[12] S. S. Bhullar and V. Singh, “The impact of academia–industry collaboration on core employability outcomes,” Technological Forecasting and Social Change, vol. 146, pp. 655–662, 2019.

[13] "Enhancing job recommendations using NLP and machine learning techniques," International Journal of Artificial Intelligence Research, vol. 8, no. 3, pp. 120–129, 2024.

[14] S. Patil and A. Suwalka, "A survey on AI-based job recommendation systems," International Journal of Computer Applications, vol. 187, no. 5, pp. 45–52, 2024.

[15] "Skill gap analysis using machine learning," Procedia Computer Science, vol. 228, pp. 94–101, 2025.

[16] "An intelligent job recommendation system based on semantic embeddings and machine learning," IEEE Access, vol. 13, pp. 50560–50571, 2025.

[17] European Investment Bank, Skill Shortages and Skill Mismatch: A Review of the Literature. Luxembourg: EIB Publications, 2025.

[18] "Skill mismatch: The concept and measurement," OECD Social, Employment and Migration Working Papers, no. 167, 2018.

[19] ONET Resource Center, Occupational Database Version 30.0. Washington, DC: National Center for ONET Development, 2025.

[20] "AI-based dashboard to assess skill needs and boost employability," The Times of India, Apr. 2025.