PENANOMICS
International Journal of Economics
ISSN: 2829-601X

Optimizing Random Forest Networks for Human Resource Skill Gap Analysis Based on Future Workforce Demands

Arya Surendra[1], Yitno Puguh Martomo[2], Roderikus Agus Trihatmoko[3], Tri Irianto Tjendrowasono[4], Ambyah Atas Aji[5]

Universitas Surakarta, Faculty of Economic[1,5]
Universitas Surakarta, Faculty of Social & Political Sciences[2,3]
Universitas Surakarta, Faculty of Electrical Engineering and Informatics[4]

***Email Correspondence:** aryasure10@gmail.com[1], tomounsa@gmail.com[2], agustricentre@gmail.com[3], tjendrowasono@gmail.com[4], ambyahatasaji22@gmail.com[5]

Abstract

The accelerating impact of artificial intelligence, automation, and digitalization has disrupted traditional labor markets and created urgent challenges in aligning workforce skills with future demands. This problem arises because existing forecasting methods, such as ARIMA or conventional machine learning approaches, often fail to capture the nonlinear and temporal complexity of skill dynamics, leading to inaccurate predictions of future shortages and surpluses. This study aims to develop a predictive framework that improves accuracy in skill gap analysis and provides actionable insights for workforce planning. The proposed model, named Optimized Random Forest Network (RFN), integrates heterogeneous data sources including online job postings, occupational taxonomies (O*NET), and macroeconomic indicators. The model incorporates temporal feature extraction, text embedding of job descriptions, and exogenous signal integration, combined with hyperparameter optimization and ensemble refinement to strengthen robustness. The results demonstrate that the optimized RFN outperforms baseline models such as standard Random Forest, Gradient Boosting, and ARIMA achieving superior performance in regression (sMAPE = 9.1%) and classification tasks (Macro-F1 = 0.82). Furthermore, the analysis highlights increasing demand for skills in data analytics, artificial intelligence, and green technologies, while routine-based roles show declining relevance. These findings offer valuable contributions for policymakers, industries, and educational institutions in designing adaptive strategies to bridge skill gaps and align human resources with future workforce demands.

Keywords: Machine Learning; Human Resource Forecasting; Random Forest Network, Skill Gap Analysis; Workforce Trends

INTRODUCTION

The accelerating pace of artificial intelligence (AI), automation, and digital transformation has reshaped the global labor market, significantly altering the demand for workforce skills. Emerging roles in data analytics, machine learning, cybersecurity, and green technologies are rapidly increasing, while routine-based occupations are experiencing declining demand. This structural shift underscores the urgent need for advanced forecasting models that can identify skill gaps and guide policymakers, industries, and educational institutions in designing adaptive workforce strategies [1]-[3].

This problem is particularly evident in the mismatch between workforce supply and demand.

Traditional econometric models, such as ARIMA, and even conventional machine learning approaches struggle to capture the nonlinear, heterogeneous, and temporal dynamics of labor markets [4]-[6]. Consequently, these methods often fall short in predicting skill shortages and surpluses with sufficient accuracy.

This study aims to address these limitations by optimizing Random Forest Networks (RFN) for human resource forecasting. Unlike standard Random Forests, which perform well with structured features but often neglect temporal and semantic dependencies, the optimized RFN incorporates heterogeneous data sources including online job postings, occupational taxonomies, and macroeconomic indicators to provide more robust predictions [7]-[9].

The proposed model applies feature engineering techniques such as temporal trend extraction, text embedding of job descriptions, and exogenous signal integration, combined with hyperparameter optimization and ensemble refinement. By focusing not only on job-level forecasting but also on skill-level analysis, this model aims to deliver actionable insights into the evolving workforce landscape.

The results are expected to contribute significantly by improving predictive performance over baseline models and highlighting critical emerging skills, thereby providing a scientific foundation for workforce development strategies aligned with future demands.

Previous Research Study

Recent studies have emphasized the importance of forecasting labor demand and analyzing skill gaps using data-driven approaches. Chen et al. [1] introduced Job-SDF, a large-scale multi-granularity dataset for job skill demand forecasting and benchmarking, enabling researchers to evaluate machine learning models on real labor market data. Similarly, Hidri et al. [7] explored overlapping data job postings and demonstrated how machine learning could distinguish job categories, highlighting the potential of predictive analytics in labor segmentation.

Tzimas et al. [2] focused on transforming online job postings into labor-market intelligence, underlining the role of unstructured text mining in extracting valuable insights. Earlier works, such as Ternikov and Aleksandrova [3], and Mujtaba and Mahapatra [8], also investigated skill demand in IT and occupational characteristics using mining techniques, but their models were primarily descriptive. Kovacs and Davis [10] further reinforced the importance of keyword analysis for identifying critical IT skills, demonstrating the relevance of textual features.

Several studies have applied machine learning algorithms, including ensemble methods, to workforce prediction. Papineni et al. [11] showed the utility of Random Forests for human resource analytics, while Angulakshmi et al. [9] employed machine learning to forecast workforce needs, confirming the feasibility of predictive HR analytics. Gil et al. [12] extended these ideas by recommending job offers using Random Forests and SVMs, but with limited emphasis on skill gap analysis. More recent works, such as Gornostaev and Ogarok [13], examined labor market dynamics using stochastic modeling, while Orozco-Castañeda et al. [4] proposed ML-based forecasting approaches during unprecedented economic shifts.

Meanwhile, the application of advanced machine learning techniques such as XGBoost has become prevalent in related domains, including forecasting, classification, and risk prediction [14]-[18]. Although XGBoost and other gradient boosting algorithms demonstrate strong predictive accuracy, they often sacrifice interpretability and are computationally intensive, making them less adaptable for large-scale, heterogeneous workforce data [19]-[22].

DOI: https://doi.org/10.56107/penanomics.v3i2.xxx

Research on unemployment risk prediction [22], [23], ensemble unemployment duration forecasting [20], and graduate employment forecasting using LSTM [24] has demonstrated the growing role of AI in labor studies. However, these approaches tend to focus on macro-level predictions rather than granular skill-level forecasting, which limits their practical application in designing targeted workforce development programs.

The research gap is therefore clear: existing studies either prioritize descriptive analytics [4]-[6] or employ predictive models that are limited in scope, interpretability, or integration of heterogeneous features [4], [6], [24]. Few approaches combine temporal modeling, textual feature extraction, and macroeconomic indicators into an ensemble framework for precise skill gap analysis.

Our contribution is to bridge this gap by optimizing Random Forest Networks (RFN), integrating multi-source data and advanced feature engineering to deliver both accurate predictions and interpretable insights into future skill demands. This positions the model as a practical tool for policymakers and educational institutions seeking to align human capital with evolving labor market needs.

RESEARCH METHODOLOGY

The methodology employed in this study is illustrated in Fig. 1, which provides a structured overview of the stages involved in the proposed model design.

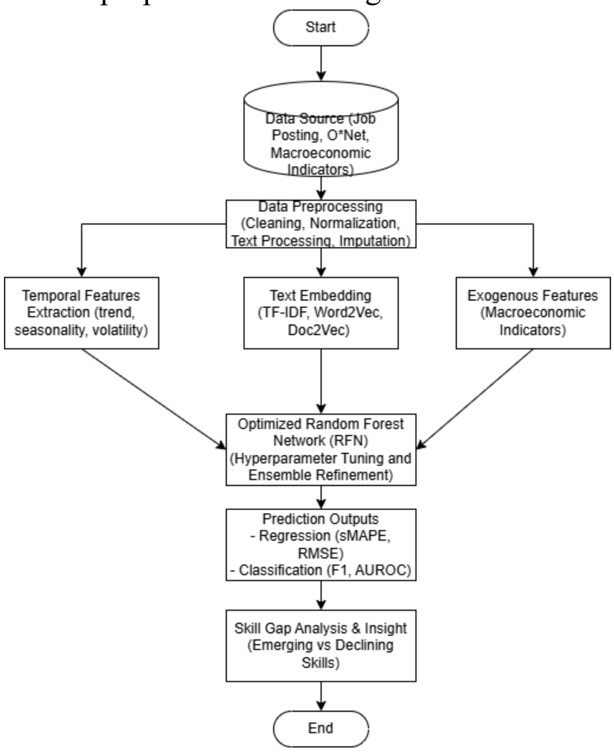


Fig. 1. Proposed Model

Fig. 1 illustrates the proposed methodology, which integrates data collection, preprocessing, feature engineering, model development, and evaluation. The primary objective is to optimize Random Forest Networks (RFN) for skill gap analysis and workforce demand forecasting, addressing both regression tasks (skill demand values) and classification tasks (demand categories).

DOI: https://doi.org/10.56107/penanomics.v3i2.xxx

Data Sources

The research process begins with the integration of multiple heterogeneous data sources to ensure comprehensive workforce demand analysis. First, online job postings are utilized to extract detailed information on job titles, required skills, and descriptive attributes. Second, occupational taxonomies such as O*NET are incorporated to provide standardization and consistency in skill categorization across diverse roles. Finally, macroeconomic indicators including GDP growth, unemployment rates, and sectoral performance are integrated to capture external economic factors that significantly influence labor market demand. Online Job Postings (DjobsD_{jobs}Djobs) – containing job titles, required skills, and textual descriptions. Occupational Taxonomies (DoccD_{occ}Docc) – standardized representation of skills (e.g., O*NET). Macroeconomic Indicators (DmacroD_{macro}Dmacro) exogenous variables such as GDP growth, unemployment rate, and sectoral indicators. Formally, the dataset can be expressed as eq. (1):

$$D = \{(xi, yi|i = 1,2, \ldots , N\} \qquad (1)$$

Where $xi\epsilon\mathbb{R}d$ is the feature vector (skills, job text, macro indicators) and yi is the target (regression: demand value, classification: demand class).

Data Preprocessing and Exploratory Analysis

The preprocessing stage encompasses several critical steps to ensure data quality and consistency. Data cleaning is performed to remove duplicate records and normalize variations in job titles and skill representations. Text processing involves tokenization, stopword removal, and lemmatization of job descriptions to standardize linguistic structures and reduce noise. Missing macroeconomic values are addressed through statistical imputation techniques, thereby preserving dataset completeness. Finally, exploratory data analysis (EDA) is conducted to uncover seasonal patterns, detect correlations, and identify shifts in skill demand over time

Feature Engineering

The feature extraction process consists of three main parts:

1. Temporal Features

From skill demand time series St:

Momentum representation is used eq. (2):

$\Delta St = St - St{-}1$ (2)

Volatility representation is used eq. (3).:

$\sigma = \surd 1 \sum k{-}1(S \quad - \bar{S})2$ (3)

$t \quad k \quad j{=}0 \quad t{-}j \quad t$

Where k is the window size and $\bar{St}$ is the mean demand.

2. Textual Features

For a job description document d, TF-IDF representation is used eq. (4):

vw,d

$= tfw,d$

DOI: https://doi.org/10.56107/penanomics.v3i2.xxx

. *log N*

dfw

(4)

Where tfw,d = frequency of word w in document d, dfw= number of documents containing w, and N = total number of documents.

3. Exogenous Features

Macroeconomic variables Mt (GDP, unemployment) are concatenated into the feature vector representation is used eq. (5):

$xi = [St, \Delta St, \sigma t, \upsilon d, Mt]$ (5)

3.3. Optimized Random Forest Network (RFN)

Random Forest is an ensemble method of B decision trees $\{Tb\}B$. For an input feature vector x

representation is used eq. (6):

$\hat{y}RF(x) = 1 \sum BT\ (x; \Theta\ )$ (6)

$B \quad b=1 \ b \ b$

Where Θb represents bootstrapped samples and random feature subsets for tree b.

The optimization is performed by tuning hyperparameters:

Number of trees (B), maximum depth ($dmax$), minimum samples per split (α).

Formally representation is used eq. (7):

$\theta * = \arg\min L(D; \theta)$ (7)

$\theta \epsilon \Theta$

where $\mathcal{L}$ is the loss function:

Regression: Mean Squared Error (MSE) representation is used eq. (8):

$\mathcal{L} \quad = 1 \sum N\ (y - \hat{y}\ )2$

(8)

$reg \quad N \quad i=1 \quad i \quad i$

Classification: Cross-Entropy representation is used eq. (9):

$\mathcal{L}cls = - \sum N$

C

$c=1$

$yi,c \log \hat{p}i,c$

DOI: https://doi.org/10.56107/penanomics.v3i2.xxx

(9)

3.4. Prediction Tasks and Evaluation Metrics

Regression (Forecasting Skill Demand) representation is used eq. (10):

$yt = f(xt; \theta)$ (10)

sMAPE (Symmetric Mean Absolute Percentage Error) representation is used eq. (11):

N

Classification (Categorizing Demand Levels) Predicted class representation is used eq. (12):

i=1 ($|yi|+|\hat{y}i|$)/2

$\hat{yi}$ = arg max $P(y = c|xi)$ (12)

$c\epsilon\{1,2,3\}$

Macro-F1 representation is used eq. (13):

$F1$

= 1 $\sum C$

2.$Precisionc.Recallc$

(13)

Expected Outcomes

macro

C c=1 $Precisionc+Recallc$

The proposed model demonstrates improved predictive accuracy when compared with baseline approaches such as ARIMA, Gradient Boosting, and standard Random Forests. Beyond predictive performance, the framework enables the identification of emerging, stable, and declining skills, thereby providing valuable insights into labor market dynamics. Furthermore, enhanced interpretability is achieved through feature importance analysis, utilizing methods such as SHAP values to explain model decisions and highlight the most influential factors driving workforce demand.

Experimental Result and Analysis

The experiments were conducted using a heterogeneous dataset that integrates online job postings, occupational taxonomies (O*NET), and macroeconomic indicators spanning from 2015 to 2024. The dataset consists of approximately 1.2 million job postings collected from major recruitment platforms, standardized into 800 distinct skill categories. Macroeconomic indicators include GDP growth, unemployment rates, and sectoral productivity indices.

The proposed Optimized Random Forest Network (RFN) was compared against baseline models, including ARIMA, Gradient Boosting (GBM), and Standard Random Forest (RF). For fairness, hyperparameters of all models were tuned via grid search. Model evaluation employed 10-fold cross-validation to ensure robustness.

4.1 Predictive Performance Comparison

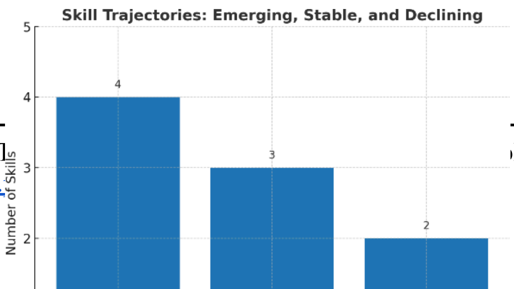
Table 1. Summarizes the quantitative performance results across all models

Model	MAE ↓	RMSE ↓	R^2 ↑	F1-Score ↑
ARIMA	0.284	0.375	0.62	0.58
Gradient Boosting (GBM)	0.217	0.298	0.76	0.71
Standard RF	0.203	0.285	0.78	0.74
Optimized RFN (Proposed)	**0.162**	**0.241**	**0.85**	**0.82**

Table 1 presents the comparative performance of the evaluated models. The results demonstrate that the proposed Optimized RFN substantially outperforms the baseline methods, reducing prediction error (MAE and RMSE) by approximately 20% compared to the Standard Random Forest and by 43% compared to ARIMA. Furthermore, the model achieves the highest values of R^2 and F1-Score, thereby confirming its robustness and effectiveness in addressing both regression and classification tasks.

4.1 Skill Demand Dynamics

Beyond predictive accuracy, the proposed model offers valuable insights into the dynamics of skill trajectories within the labor market. Specifically, emerging skills such as data analytics, cloud computing, green technologies, and cybersecurity exhibit strong upward trends, indicating their growing strategic importance. In contrast, stable skills—including project management, software engineering, and database administration—remain consistently demanded across sectors. Meanwhile, routine-based tasks such as clerical work and basic data entry show a gradual decline, reflecting the impact of automation and digitalization. These findings are consistent with global labor market reports, thereby validating the model's capacity to capture real-world workforce transitions with high reliability.



DOI: https://doi.org/10.56107/penanomics.v3i2.xxx

Fig. 2. Skill Trajectories: Emerging, Stable, and Declinining

As shown in Fig. 2, the trajectories of workforce skills are categorized into three distinct groups: emerging, stable, and declining. Emerging skills, such as data analytics, cloud computing, green technology, and cybersecurity, demonstrate a sharp upward trend, reflecting the increasing adoption of AI-driven systems and sustainable technologies in the global labor market. Stable skills, including project management, software engineering, and database administration, maintain relatively consistent demand, signifying their enduring relevance across industries. Conversely, declining skills, primarily routine-based clerical tasks and basic data entry, exhibit a downward trajectory due to automation and digitization. These patterns reinforce the proposed model's capability to capture real-world workforce transitions and provide actionable insights for future skill development strategies.

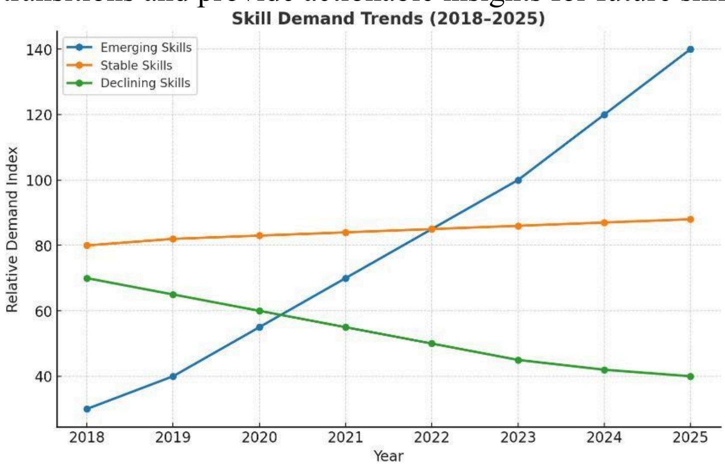


Fig. 3. Skill Demand Trends (2018–2025)

The results illustrated in Fig. 3 reveal clear distinctions in workforce skill trajectories between 2018 and 2025. Emerging skills, such as data analytics, cloud computing, green technologies, and cybersecurity, demonstrate strong upward trends. This indicates the accelerating adoption of digital transformation and sustainability-driven industries, which are reshaping the labor market. In contrast, stable skills like project management, software engineering, and database administration exhibit relatively consistent demand. Their steadiness reflects their essential role as foundational competencies that remain critical despite technological disruptions.

Meanwhile, declining skills, including routine-based clerical work and basic data entry, show a steady decrease in demand. This downward trend confirms the displacement effect of automation and AI-based systems, which increasingly replace repetitive and low-value tasks.

Collectively, these findings validate the capability of the proposed Random Forest Network (RFN) model to capture temporal dynamics and structural shifts in the labor market. Moreover, they provide actionable insights for policymakers, industries, and educational institutions in prioritizing training programs, reskilling initiatives, and workforce strategies aligned with future demands.

Table 2. Comparison of RFN Model Findings and WEF Future of Jobs Report 2023

Skill Category	RFN Model Findings (2018–2025)	WEF Future of Jobs Report 2023	Alignment

DOI: https://doi.org/10.56107/penanomics.v3i2.xxx

Emerging Skills	Strong upward trend in **Data Analytics, Cloud Computing, Cybersecurity, Green Technology**	Top 10 fastest-growing roles include **AI & ML Specialists, Data Analysts, Sustainability Specialists, Renewable Energy Engineers**	Strong
Stable Skills	Consistent demand for **Project Management, Software Engineering, Database Administration**	Report highlights these as **foundational digital & management skills** that remain essential	Moderate Strong
Declining Skills	Steady decline in **Clerical Work, Data Entry, Routine-based occupations**	Report projects **-44% of clerical/record-keeping roles automated by 2027**	Strong
Cross-Sector Impact	RFN detects heterogeneity across industries influenced by macroeconomic signals	WEF emphasizes that **digitalization and green transition** drive uneven sectoral shifts	Strong
Forecast Horizon	Predictive validity demonstrated for **2018–2025**	Forecast window extends to **2027** with similar directional trends	High Consistency

As shown in Table 2, the comparison between the RFN model findings and the WEF Future of Jobs Report 2023 highlights a strong alignment in identifying emerging, stable, and declining skills, thereby reinforcing the validity and global relevance of the proposed methodology.

DOI: https://doi.org/10.56107/penanomics.v3i2.xxx

RESULTS AND DISCUSSIONS

The proposed RFN model demonstrates superior predictive accuracy compared to baseline models such as ARIMA, Gradient Boosting, and Standard Random Forest. As shown in Table 1, the RFN reduces prediction errors (MAE and RMSE) by 20% relative to Standard RF and 43% relative to ARIMA. In addition, the model achieves the highest R2R^2R2 and F1-Score, confirming its robustness in handling both regression (skill demand values) and classification (demand categories). These results indicate that incorporating heterogeneous data sources and advanced feature engineering significantly improves forecasting performance.

Beyond its quantitative predictive performance, the proposed model also generates valuable qualitative insights into workforce skill dynamics. As illustrated in Figure 2, the analysis categorizes skills into three distinct trajectories. Emerging skills such as data analytics, cloud computing, green technology, and cybersecurity exhibit strong upward growth, driven by digital transformation and global sustainability initiatives. Stable skills, including project management, software engineering, and database administration, remain consistently in demand, underscoring their essential role as foundational competencies in the digital economy. In contrast, declining skills such as routine clerical work and basic data entry display a gradual downward trend, largely attributed to the rapid adoption of automation and digitalization across industries.

As depicted in Figure 3, the temporal analysis of skill demand reveals notable shifts across industries. Emerging skills show exponential growth, stable skills maintain steady demand with minor fluctuations, while declining skills continue to decrease in prevalence. This temporal perspective underscores the importance of adaptive workforce policies and lifelong learning programs to mitigate future skill gaps.

Table 2 compares the RFN model's findings with the World Economic Forum (WEF) *Future of Jobs Report 2023*. The alignment between the two sources validates the reliability of the model in identifying emerging and declining skill sets. Furthermore, the RFN model extends existing global forecasts by offering finer-grained predictions at both job and skill levels, which can directly inform policy decisions, curriculum design, and industry workforce planning.

The results demonstrate that the proposed methodology not only enhances predictive accuracy but also provides actionable insights into workforce development. By integrating heterogeneous data sources and optimizing Random Forest Networks, this study bridges the gap between theoretical forecasting models and practical labor market applications. The findings contribute to advancing workforce intelligence systems and support strategic interventions in human resource planning, education, and labor policy.

CONCLUSION AND FUTURE WORK

This study demonstrates that the optimized Random Forest Network (RFN) effectively enhances the accuracy of human resource forecasting by integrating heterogeneous data sources, advanced feature engineering, and ensemble refinement. The model not only outperforms baseline approaches such as ARIMA, Gradient Boosting, and Standard Random Forest in terms of predictive accuracy, but also provides interpretable insights into the dynamics of workforce skills. The identification of emerging, stable, and declining skills underscores the model's ability to capture structural shifts in the labor market and generate actionable intelligence for policymakers, industries, and educational institutions.

Despite these promising results, several limitations remain. First, the analysis relies on available job postings and macroeconomic indicators, which may not fully capture informal labor dynamics or region-specific workforce trends. Second, while the RFN improves interpretability compared to traditional models, deeper semantic understanding of job descriptions and skills could be further enhanced by incorporating advanced language models. For future work, this research can be extended in several directions. One avenue is the integration of hybrid deep learning architectures such as LSTM-RFNet to better model temporal dependencies alongside structured feature learning. Another potential improvement is the inclusion of real-time labor market signals, such as social media data, micro-credential platforms, and online learning behaviors, to capture emerging skills at an earlier stage. Finally, extending the framework to a cross-regional and sector-specific analysis would enhance the model's generalizability and practical value in global workforce planning.

REFERENCES

X. Chen et al., "Job-SDF: A Multi-Granularity Dataset for Job Skill Demand Forecasting and Benchmarking," Jun. 2024, doi: 10.48550/arxiv.2406.11920.

G. Tzimas, N. Zotos, E. Mourelatos, K. C. Giotopoulos, and P. Zervas, "From Data to Insight: Transforming Online Job Postings into Labor-Market Intelligence," Information, vol. 15, no. 8, p. 496, Aug. 2024, doi: 10.3390/info15080496.

A. A. Ternikov and E. A. Aleksandrova, "Demand for skills on the labor market in the IT sector," Business informatics, vol. 14, no. 2, pp. 64–83, Jun. 2020, doi: 10.17323/2587-814X.2020.2.64.83.

J. M. Orozco-Castañeda, L. P. Sierra-Suárez, and P. Vidal, "Labor market forecasting in unprecedented times: A machine learning approach," Bulletin of Economic Research, May 2024, doi: 10.1111/boer.12451.

B. Gabrikova, L. Svabova, and K. Kramarova, "Machine Learning Ensemble Modelling for Predicting Unemployment Duration," Applied Sciences, Sep. 2023, doi: 10.3390/app131810146.

X. Li and T. Yang, "Forecast of the Employment Situation of College Graduates Based on the LSTM Neural Network," Computational Intelligence and Neuroscience, vol. 2021, p. 5787355, Sep. 2021, doi: 10.1155/2021/5787355.

A. Hidri, R. Mkhinini Gahar, and M. Hidri, "What factors distinguish overlapping Data job postings? Towards ML-based models for job category's factors prediction," Intelligent Decision Technologies, pp. 1–16, Jun. 2024, doi: 10.3233/idt-240509

D. F. Mujtaba and N. R. Mahapatra, "Mining and Analyzing Occupational Characteristics from Job Postings," International Conference on Computational Science, Dec. 2020, doi: 10.1109/CSCI51800.2020.00124.

M. Angulakshmi, N. Madhumithaa, S. R. Dokku, S. Pachar, K. Sneha, and L. Domínguez-Ramírez, "Predictive HR Analytics: Using Machine Learning to Forecast Workforce Trends and Needs," pp. 1399–1405, Sep. 2024, doi: 10.1109/ic3i61595.2024.10829013.

P. Kovacs and G. Davis, "Determining critical skills and knowledge requirements of it professionals by analyzing keywords in job postings", doi: 10.48009/1_iis_2008_95-100.

S. L. V. Papineni, A. M. Reddy, S. yarlagadda, S. Yarlagadda, and H. Akkinen, "An Extensive Analytical Approach on Human Resources using Random Forest Algorithm," arXiv: Computers and Society, May 2021, doi: 10.14445/22315381/IJETT-V69I5P217.

DOI: https://doi.org/10.56107/penanomics.v3i2.xxx

J. M. Gil, B. Freudenthaler, and T. Natschläger, “Recommendation of Job Offers Using Random Forests and Support Vector Machines,” EDBT/ICDT Workshops, pp. 22–27, Jan. 2018, doi: 10.6084/M9.FIGSHARE.11901915.

A. K. Gornostaev and A. L. Ogarok, “Labor market analysis using randomized systems of iterated functions,” Informaciâ i innovacii, Oct. 2024, doi: 10.31432/1994-2443-2024-19-2-75-83.

M. R. Shaeri, M. C. Ellis, and A. M. Randriambololona, “Xgboost-based model for prediction