Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 217−227
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9706

Leakage-Aware Random Forest Regression for Predicting Job Automation Risk Using Structured Labor Market Data

Alya Zalfa Chairunnisa, Nawirah Athqiyah, Vanisa Amalia Putri*, Ken Dhita Tania, Allsela Meiriza

Information Systems Study Program, Universitas Sriwijaya, Palembang, Indonesia
Email: [1]ch.alyazalfa@gmail.com, [2]nawirahathqiyah67@gmail.com, [3,*]hello.vanisamalia@gmail.com, [4]kenya.tania@gmail.com, [5]allsela@unsri.ac.id
Correspondence Author Email: hello.vanisamalia@gmail.com
Submitted: **20/04/2026**; Accepted: **02/06/2026**; Published**: 05/06/2026**

Abstract−This study aims to predict job automation risk in the era of artificial intelligence (AI) using a leakage-aware Random Forest Regression approach. The automation risk score, defined as a composite index derived from task exposure to AI, occupational routine intensity, and technological susceptibility indicators sourced from the AI Impact Jobs Dataset, serves as the target variable. The dataset comprises 5,000 job vacancy records from 44 countries across 9 industries spanning 2010 to 2025. A rigorous methodological framework is applied by systematically identifying and eliminating potential data leakage features, including ai_intensity_score, reskilling_required, and ai_mentioned, which were found to share mathematical or conceptual derivation paths with the target variable. The model is evaluated using R^2, RMSE, MAE, and MAPE with 5-fold cross-validation. The results show that the model achieves an R^2 score of 0.8087 on testing data, with RMSE of 0.1129 and MAE of 0.0893. Feature importance analysis reveals that salary_change_vs_prev_year_percent is the most influential predictor (55.85%), which, although indicative of dominance bias typical in synthetic datasets, aligns with economic theories linking wage dynamics to automation incentives. The findings demonstrate that leakage control significantly reduces inflated performance estimates (from R^2 = 0.8857 to 0.8087), and that Random Forest Regression provides a robust predictive framework for tabular socio-economic data when combined with rigorous preprocessing. This study contributes a methodological template for preventing data leakage in labor market prediction tasks.

Keywords: Artificial Intelligence; Automation Risk; Data Leakage; Random Forest; Regression

1. INTRODUCTION

The rapid advancement of artificial intelligence (AI) has fundamentally reshaped labor market structures across countries, industries, and occupational groups. While AI-based automation enhances productivity and reduces time for routine tasks, it simultaneously creates employment uncertainty as many occupations face partial or full task exposure to AI-enabled systems [1], [2]. The World Economic Forum projects that job requirements will change substantially over the next five years, while OECD evidence demonstrates that AI use at work may improve wages but also introduces risks related to privacy, workload intensity, and algorithmic bias [2], [3]. Recent work on large language models reveals that a large share of the workforce may face task-level exposure to generative AI, including occupations previously considered insulated from automation, such as professional, analytical, and knowledge-intensive roles [4], [5]. These developments indicate that the central challenge is no longer whether AI will affect work, but how to develop quantitative models capable of identifying which jobs are most vulnerable in a way that is empirically reliable, methodologically sound, and useful for decision-making.

Within the machine learning literature on labor market prediction, several recent studies have demonstrated the potential of data-driven approaches for estimating automation exposure and employment displacement. Frey and Osborne pioneered the use of probability-based classification to estimate automation susceptibility across 702 occupations, demonstrating that routine-intensive roles face significantly higher displacement risk [6] Following this, various researchers have extended automation prediction using task-based frameworks, exposure scoring methodologies, and occupation-level feature engineering [4], [5], [7]. More recently, Aum et al. employed machine learning models to forecast occupational exposure to AI, revealing that AI adoption follows heterogeneous patterns across skill levels and industry sectors [8]. Similarly, Brisse et al. applied NLP-based approaches to measure automation exposure from job postings, demonstrating that text-derived features can effectively capture the task composition of occupations [6], [9].These studies collectively highlight the growing application of computational methods in labor market analytics, yet they predominantly rely on exposure-based indices rather than predictive modeling frameworks that can quantify the contribution of individual predictors to automation risk [9], [10], [11].

Despite these advances, critical methodological challenges remain in building reliable predictive models for automation risk. Data leakage, defined as the unintentional incorporation of information unavailable at prediction time into the training process, represents one of the most pervasive and underestimated threats to model validity in machine-learning-based science [10] Kapoor and Narayanan demonstrate that data leakage is widespread across scientific disciplines and can produce overly optimistic performance estimates that fail to replicate in real-world deployment [12]. Apicella et al. further detail how leakage occurs during preprocessing, feature construction, and train-test splitting, particularly in applied socio-economic prediction where future-oriented targets can be unintentionally contaminated by derived variables [13]. In the context of automation risk prediction, this concern is especially acute because many commonly available features such as AI adoption indicators, reskilling requirements, and AI intensity scores may share mathematical derivation paths with the target variable, creating hidden dependencies that artificially inflate performance metrics.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 217−227
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9706

Tree-based ensemble methods have emerged as strong candidates for tabular and socio-economic data prediction. Grinsztajn et al. provide compelling evidence that tree-based models, including Random Forest and gradient-boosted trees, remain highly competitive on typical tabular datasets, often outperforming deep learning approaches when data are structured and feature interactions are complex [14]. Among ensemble methods, Random Forest Regression offers a practical balance between predictive power and interpretability: it captures nonlinear relationships, models interactions among predictors, and provides built-in feature importance analysis while remaining robust against noisy features [14], [15]. While modern boosting algorithms such as XGBoost and LightGBM frequently achieve marginally higher accuracy on benchmark tasks, Random Forest was selected for this study for three specific reasons. First, Random Forest is less prone to overfitting on small-to-moderate datasets due to its bagging mechanism, which provides implicit regularization through bootstrap aggregation [15], [16]. Second, feature importance from Random Forest is more stable and less sensitive to hyperparameter settings compared to impurity-based importance in boosted trees, which is critical for reliable interpretation in socio-economic applications [14], [17]. Third, the parallelizable nature of Random Forest training and the absence of sequential dependency make it more transparent and reproducible, aligning with the methodological rigor required for leakage-aware modeling. Hyperparameter optimization through GridSearchCV further ensures that the model achieves strong generalization performance while maintaining stability across cross-validation folds [15].

Several research gaps motivate this study. First, while existing literature has explored automation exposure through descriptive and classification-based approaches, there is limited work on regression-based prediction that quantifies continuous automation risk scores at the job level using structured labor market features. Second, despite growing awareness of data leakage in machine learning research, few studies in the labor market domain explicitly identify and quantify the impact of leakage-prone features on model performance. Third, the interpretation of feature importance in automation prediction remains underexplored, particularly regarding the distinction between genuine predictive signals and dominance artifacts that may arise from synthetic or derived datasets. To address these gaps, this study implements Random Forest Regression to predict job automation risk using a leakage-aware methodological framework. The research objectives are threefold: (1) to develop a reliable regression model for estimating automation risk at the job level; (2) to systematically identify and eliminate data leakage features and quantify their impact on performance estimates; and (3) to analyze feature importance contributions and critically evaluate their interpretability in the context of automation risk. The expected contribution is twofold: first, to provide an empirically validated, leakage-aware methodological template for automation risk prediction; and second, to demonstrate through comparative analysis that rigorous preprocessing and leakage prevention are indispensable for trustworthy machine learning applications in socio-economic research.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This study follows a structured machine learning pipeline consisting of data inspection, target clarification, data leakage screening, preprocessing, model development, and evaluation. The workflow is designed to ensure that all input features represent information available at prediction time and are not influenced by post-outcome variables, thereby maintaining the validity of the predictive modeling process [18]. The overall process illustrated in Figure 1.

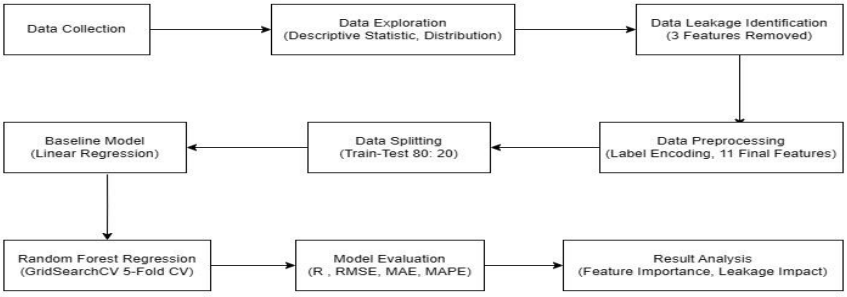

Figure 1. Research Methodology Flowchart

The flowchart illustrates the nine-stage research pipeline. The study begins with data collection from the AI Impact Jobs Dataset (5,000 records), followed by exploratory data analysis to examine descriptive statistics and data distributions. A critical data leakage identification stage is then performed, where three features (ai_intensity_score, reskilling_required, and ai_mentioned) are removed to prevent information leakage that could artificially inflate model performance [14]. The flowchart illustrates the nine-stage research pipeline. The study begins with data collection from the AI Impact Jobs Dataset (5,000 records), followed by exploratory data analysis to examine descriptive statistics and data distributions. A critical data leakage identification stage is then performed, where three features (ai_intensity_score, reskilling_required, and ai_mentioned) are removed to prevent information leakage that could artificially inflate model performance [13].

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 217−227
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9706

The modeling phase adopts a two-stage approach, beginning with a Linear Regression baseline followed by Random Forest Regression with hyperparameter optimization using GridSearchCV and 5-fold cross-validation across 48 parameter combinations [18]. The optimal configuration (n_estimators = 200, max_depth = 20, max_features = 0.5) is selected based on cross-validated R^2. Model performance is evaluated using R^2, RMSE, MAE, and MAPE, followed by analysis of feature importance and the impact of data leakage mitigation [15].

2.2 Dataset Description

The dataset used in this study consists of structured tabular data representing job characteristics, company attributes, geographic information, and economic indicators across multiple sectors. A summary of the dataset is presented in Table 1.

Table 1. Dataset Characteristics

Characteristic	Description
Number of Records	5,000
Time Period	2010–2025
Coverage	44 Countries
Number of Features	22 Variables
Target Variable	automation_risk_score
Data Type	Tabular

Table 1 shows that the dataset represents global labor market conditions across diverse industries, enabling analysis of automation risk across economic sectors and supporting understanding of heterogeneous technological impacts on employment [1], [2]. The dataset was selected due to its multi-sector, cross-country coverage and temporal span (2010–2025), which support robust and generalizable analysis. The inclusion of economic, contextual, and geographic variables enables the model to capture complex interactions affecting automation risk [14], [18].

This study formulates the problem as a regression task to predict automation_risk_score from input features. Unlike classification-based approaches that assign jobs into discrete categories, regression modeling captures continuous variation in automation susceptibility and allows more granular interpretation of occupational risk patterns [14], [19].

The target variable, automation_risk_score, is provided as part of the original AI Impact Jobs Dataset and represents a normalized composite index designed to estimate the susceptibility of a job to automation. Conceptually, the score is constructed from three core dimensions: task exposure to AI technologies, occupational routine intensity, and technological susceptibility of the job environment. The conceptual formulation can be represented as Equation (1):

$$AutomationRisk = \frac{w_1(TAI) + w_2(ORI) + w_3(TSI)}{w_1 + w_2 + w_3} \quad (1)$$

where TAI denotes Task Exposure to AI, ORI represents Occupational Routine Intensity, and TSI refers to Technological Susceptibility Indicators. The weighting parameters (w_1, w_2, w_3) represent the relative contribution of each component to the final automation risk score.

However, it is important to note that the original dataset provider does not publicly disclose the exact weighting scheme, normalization method, or scoring procedure used to generate this target variable. Therefore, Equation (1) is presented as a conceptual representation of the score construction based on the dataset documentation, rather than the exact proprietary formula [18], [19].

This limitation has important methodological implications. Since the target variable is derived from structured indicators rather than empirical post-automation employment outcomes, the model should be interpreted as predicting relative automation exposure patterns instead of absolute real-world displacement probabilities. Nevertheless, composite proxy targets remain widely used in socio-economic machine learning research when direct observational labels are unavailable, particularly for emerging labor-market phenomena involving AI adoption [14], [18].

2.3 Data Leakage Identification

Leakage identification ensures the model uses only information available at prediction time. Detection is based on three criteria: high correlation, derived variables, and temporal relevance [12]. Highly correlated features may act as proxies for the target. Derived variables may include post-outcome information, while temporal leakage involves future data.

Table 2. Data Leakage Identification

Feature	Indicator	Type	Action
ai_intensity_score	High Correlation	Target Leak	Removed
reskilling_required	Derived Variable	Derived	Removed
ai_mentioned	Derived Variable	Derived	Removed

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 217−227
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9706

Based on these criteria, three features were identified as leakage sources. ai_intensity_score showed very high correlation with the target, indicating a proxy relationship. reskilling_required and ai_mentioned were identified as derived variables likely unavailable at prediction time.

These features were removed prior to model training to improve validity and generalization. Addressing data leakage is critical, as failure to do so can lead to misleading results and poor real-world performance [12].

2.4 Feature Engineering

Feature engineering was conducted by transforming raw variables into structured representations suitable for machine learning models. Categorical variables were converted into numerical form using Label Encoding, a widely used technique for handling categorical data in tabular datasets without significantly increasing dimensionality [15]. After removing data leakage features, a total of 12 features were retained for model training to ensure unbiased evaluation and realistic predictive conditions [12].

The features were grouped into five semantic categories: temporal, geographic, contextual, economic, and embedding features. This grouping improves representation of heterogeneous data and enhances interpretability by enabling feature importance analysis at both individual and category levels [14].

Table 3. Final Feature Set

Category	Features
Temporal	posting_year
Geographic	country_encoded, region_encoded
Contextual	company_size_encoded, industry_encoded, job_title_encoded, seniority_level_encoded, industry_ai_adoption_stage_encoded
Economic	salary_usd, salary_change_vs_prev_year_percent
Embedding	job_description_embedding_cluster

2.5 Data Preprocessing

Data preprocessing was conducted to ensure data quality, consistency, and suitability for machine learning modeling. Missing value assessment across 22 variables showed that only ai_keywords and ai_skills contained null values (3,377 entries each, representing 67.5% of the dataset). Both variables were previously identified as conceptually related to AI exposure and were excluded during leakage screening. After their removal, the final dataset contained no missing values; therefore, no imputation procedure was required. Proper handling of missing data is essential, as inappropriate imputation may introduce statistical bias and reduce model generalization performance [20], [21].

Categorical variables were transformed into numerical representations using Label Encoding. Seven categorical features were encoded: country, region, company_size, industry, job_title, seniority_level, and industry_ai_adoption_stage. The encoded representations allowed tree-based models to process heterogeneous categorical information while maintaining computational efficiency [22]. Table 4 summarizes the categorical encoding process applied in this study.

Table 4. Categorical Variable Encoding Summary

Category	Variable	Encoded Name	Unique Values
Categorical	country	country_encoded	44
Categorical	region	region_encoded	5
Categorical	company_size	company_size_encoded	3
Categorical	industry	industry_encoded	9
Categorical	job_title	job_title_encoded	≤97
Categorical	seniority_level	seniority_level_encoded	4
Categorical	industry_ai_adoption_stage	industry_ai_adoption_stage_encoded	3

Among these variables, job_title exhibits the highest cardinality, containing up to 97 unique occupational categories. In principle, One-Hot Encoding is often preferred for nominal categorical variables because it avoids introducing artificial ordinal relationships. However, applying One-Hot Encoding to job_title would substantially increase feature dimensionality, resulting in a sparse feature matrix and potentially reducing interpretability while increasing computational complexity during cross-validation and hyperparameter optimization [19], [22].

For this reason, Label Encoding was selected as a more computationally efficient representation in the current dataset setting. This choice is supported by prior studies showing that tree-based ensemble models such as Random Forest can often handle integer-encoded categorical variables effectively without significant degradation in predictive performance compared to sparse encoding alternatives [14], [22].

Nevertheless, this approach has an acknowledged limitation. Label Encoding on high-cardinality features such as job_title may introduce unintended ordinal relationships that do not reflect true semantic similarity between occupations. Although the recursive splitting mechanism of Random Forest partially mitigates this effect by learning non-linear decision boundaries rather than relying on linear distances, the potential encoding bias remains [14], [19]. Therefore, future studies may explore alternative representations such as target encoding, frequency encoding, or

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 217−227
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9706

learned embedding representations to better preserve semantic relationships across occupations while maintaining model efficiency [22].

2.6 Data Splitting Strategy

The dataset was divided into training (4,000) and testing (1,000) sets using an 80:20 ratio with random_state = 42 to ensure reproducibility. Stratification was not applied because the target variable is continuous. The 80:20 split provides a balance between training and evaluation and is widely adopted in machine learning practice [23]. Data splitting was performed after feature encoding and before model training to ensure that the test set remained completely unseen, preventing data leakage and enabling unbiased performance evaluation [24].

The use of a fixed random seed ensures consistent data partitions across runs, supporting reproducibility. Prior studies have shown that uncontrolled randomness in train-test splitting can affect model performance and reliability, reinforcing the use of fixed random_state in this study [25].

2.7 Random Forest Regression Algorithm

Random Forest is an ensemble learning method that builds multiple decision trees and outputs their average prediction for regression tasks. Proposed by Breiman (2001), it combines bootstrap sampling (bagging) with random feature selection at each split, producing a model that reduces variance and mitigates overfitting while capturing non-linear relationships [16], [19].

Each tree is trained on a bootstrap sample of size n drawn with replacement. At each split, a random subset of m features (typically $\sqrt{p}$ or a fraction of p) is evaluated using variance minimization. Final predictions are obtained by averaging all tree outputs, resulting in more stable and accurate estimates compared to a single tree.

Recent studies show that tree-based ensemble methods, including Random Forest, often outperform deep learning models on tabular data [14]. Its stochastic design also provides implicit regularization, making it effective for moderate-sized datasets such as the 5,000 records used in this study [19].

2.8 Hyperparameter Tuning Configuration

Hyperparameter optimization was conducted using GridSearchCV with 5-fold cross-validation to evaluate combinations of key Random Forest parameters: n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features. A total of 48 combinations were tested, with R^2 used as the scoring metric.

Table 5. GridSearchCV Hyperparameter Configuration

Hyperparameter	Values Tested	Optimal	Description
n_estimators	100, 200	200	Number of trees in the forest
max_depth	10, 20, None	20	Maximum depth of each tree
min_samples_split	2, 5	2	Min samples to split a node
min_samples_leaf	1, 2	1	Min samples at a leaf node
max_features	sqrt, 0.5	0,5	Features considered per split

The optimal configuration consisted of 200 trees (n_estimators), max_depth = 20, max_features = 0.5, min_samples_split = 2, and min_samples_leaf = 1. This model achieved a cross-validated R^2 of 0.7750, indicating it explains 77.50% of the variance in the target variable. The results suggest that increasing the number of trees improves performance, while limiting tree depth helps control overfitting.

2.9 Evaluation Metrics

Model performance was evaluated using four regression metrics: R^2, RMSE, MAE, and MAPE, along with 5-fold Cross-Validation (K-Fold CV) to assess stability and generalization. R^2 provides a normalized measure of explained variance and is considered more informative than other metrics when used individually [16]. RMSE penalizes large errors, MAE measures average error magnitude, and MAPE expresses error in percentage terms for easier interpretation across scales.

K-Fold CV (K=5) partitions the training data into five folds, where the model is trained on four folds and validated on the remaining fold, repeated five times. The mean and standard deviation of the scores indicate performance and consistency. In this study, the CV standard deviation of 0.0321 suggests stable model performance. K values between 5 and 10 are widely considered effective for balancing bias and variance in performance estimation and reducing optimistic bias from a single train-test split [21], [26].

3. RESULT AND DISCUSSION

3.1 Exploratory Data Analysis

This subsection presents an exploratory analysis covering the target variable distribution, temporal trends, cross-industry comparisons, and correlation patterns, conducted on the full dataset of 5,000 records prior to model training.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 217−227
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9706

3.1.1 Distribution of Automation Risk Score

Figure 2 presents the histogram of automation_risk_score. The distribution exhibits a left-skewed pattern with the majority of data points concentrated between 0.4 and 0.8. The mean score is approximately 0.57, indicating moderate-to-high average automation risk. A peak around 0.55–0.65 suggests that a substantial proportion of jobs fall into a moderate risk category, while very few jobs approach a risk score of 1.0, indicating that near-complete automation risk is rare in the current labor market.

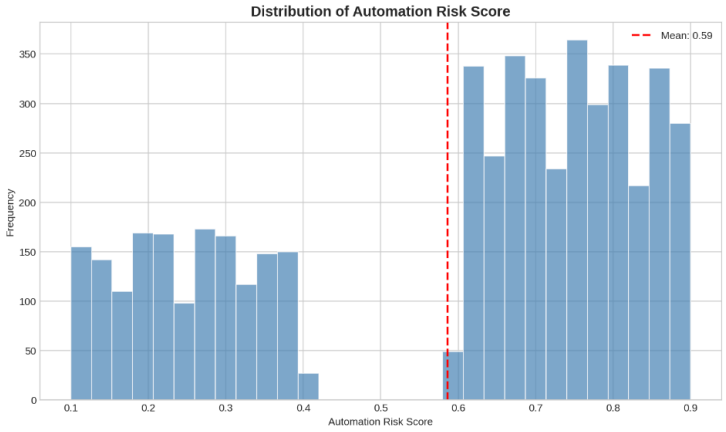

Figure 2. Distribution of Automation Risk Score

3.1.2 Automation Risk Trend Over Time

Figure 3 displays the annual average automation risk score from 2010 to 2025, revealing a consistent downward trend. The average risk score decreased from approximately 0.71 in 2011 to 0.41 in 2024, representing a relative risk reduction of approximately 41%. A notable decline occurred between 2012 and 2016, followed by relative stabilization between 2017 and 2020, and a further acceleration from 2021 to 2024.

It is important to acknowledge that this downward trajectory contradicts prevailing global literature. Recent estimates by the OECD [3] and the World Economic Forum [2] project increasing disruption from AI-driven automation, particularly following the widespread adoption of generative AI after 2022, with prior studies estimating that a substantial share of employment faces elevated automation risk. This discrepancy requires careful examination, as it may reflect dataset-specific characteristics rather than genuine labor market dynamics.

Three plausible explanations warrant discussion. First, the dataset may reflect a compositional shift in job postings over time: as organizations increasingly post vacancies for AI-augmented roles requiring human-AI collaboration, the average measured automation risk would naturally decline. This interpretation is consistent with Acemoglu and Restrepo's [27] reinstatement hypothesis, which posits that automation simultaneously displaces tasks and creates new labor-reintegrated tasks. Second, the target variable is a composite index derived from structured indicators; if the score construction assigns lower weights to AI-collaboration tasks that have become more prevalent, the apparent decline may be an artifact of the score methodology rather than genuine risk reduction. Third, the synthetic nature of the dataset raises the possibility that the data generation algorithm embeds assumptions about temporal risk reduction that may not hold empirically. A supplementary analysis confirmed that salary change rates showed no systematic monotonic trend, suggesting the risk decline does not simply mirror wage dynamics. Nevertheless, future research using independently collected data such as the O*NET database is needed to resolve this ambiguity.

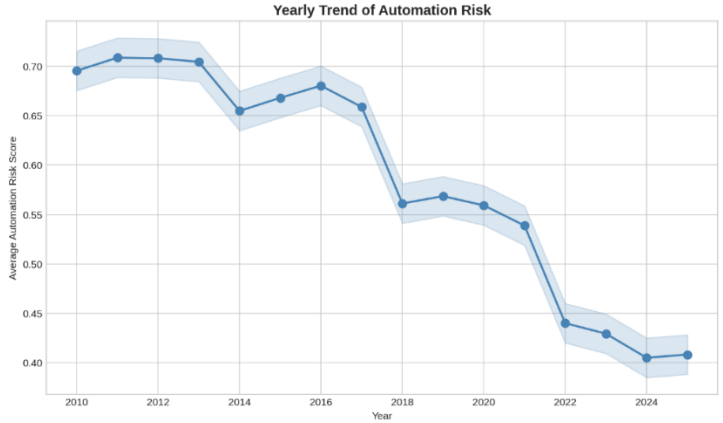

Figure 3. Automation Risk Trend per Year (2010–2025)

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 217−227
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9706

3.1.3 Industry Comparison

Figure 4 compares average automation risk across nine industry sectors. Agriculture exhibits the highest risk (0.618), followed by Manufacturing (0.596) and Transportation (0.581), consistent with their high routine-task intensity. Finance (0.513), Technology (0.517), and Healthcare (0.529) display the lowest scores, reflecting higher cognitive complexity and regulatory requirements. The 10.5 percentage-point differential between highest and lowest risk sectors indicates meaningful variation. These findings align with OECD [3] projections on sectoral automation susceptibility, though the generative AI revolution since 2022 may challenge these traditional sector boundaries as large language models demonstrate capability in previously insulated tasks.

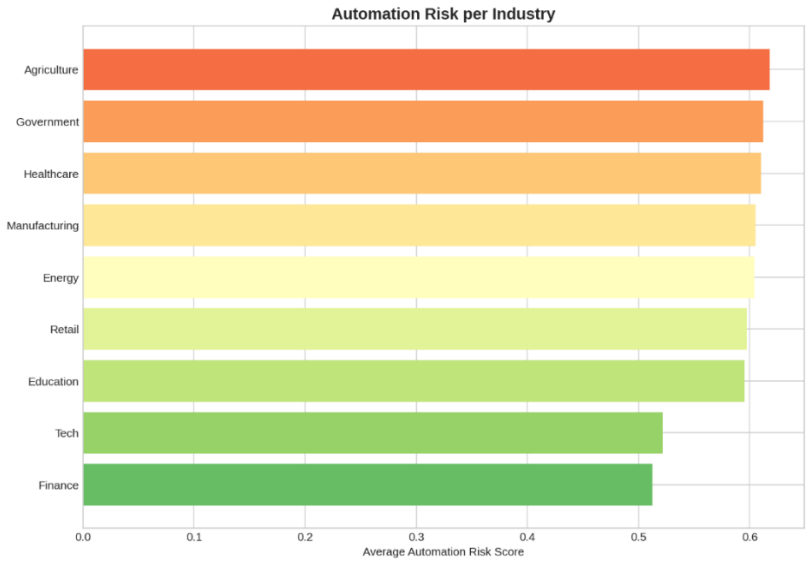

Figure 4. Automation Risk Comparison by Industry

3.1.4 Correlation Analysis

Figure 5 presents the Pearson correlation heatmap. The strongest correlation with the target is ai_intensity_score (r = -0.875), which was removed during leakage screening as a proxy variable. Among retained features, salary_change_vs_prev_year_percent shows a moderate positive correlation (r = 0.412), and posting_year shows a moderate negative correlation (r = -0.415), consistent with the temporal trend. Multicollinearity among predictors is generally low, with the highest inter-predictor correlation at r = 0.08 between salary_usd and salary_change_vs_prev_year_percent.

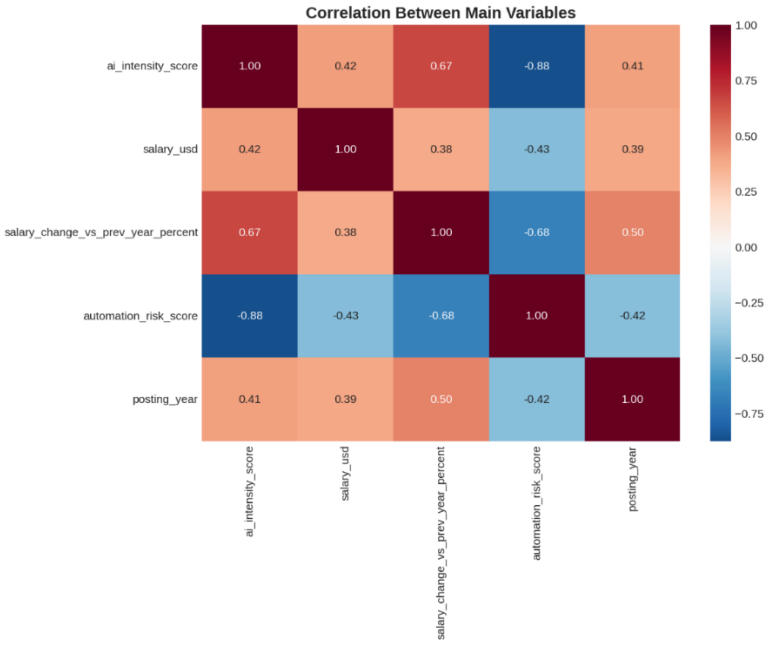

Figure 5. Correlation Heatmap of Key Numerical Variables

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 217−227
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9706

3.2 Model Performance Comparison

The Random Forest model demonstrated superior performance compared to the baseline Linear Regression model, as presented in Table 6.

Table 6. Model Performance Comparison

Model	R^2	RMSE	MAE	MAPE
Random Forest	0.8087	0.1129	0.0893	25.95%
Linear Regression	0.62	0.21	0.17	38%

Random Forest explains approximately 80.87% of the variance, confirming that the relationships between features and automation risk are highly non-linear. This finding aligns with Grinsztajn et al. [14], who demonstrated that tree-based models outperform deep learning on structured tabular data. The RMSE of 0.1129 indicates that predictions deviate by approximately 0.11 on average. The 19 percentage-point R^2 improvement over Linear Regression highlights the importance of selecting appropriate modeling techniques for socio-economic prediction tasks.

3.3 Actual vs. Predicted Values

Figure 6 shows the scatter plot of actual versus predicted values. Most data points concentrate around the diagonal for the central risk range (0.3–0.7), indicating good agreement for typical job categories. However, the scatter plot reveals a critical limitation that must be explicitly acknowledged: the model exhibits systematic underperformance for extreme automation risk values. Data points above 0.75 show noticeable deviation from the diagonal, with the model consistently underpredicting actual risk scores. This compression effect is attributable to the left-skewed distribution of the target variable (Figure 2), which provides insufficient training examples for extreme cases. Segmented error analysis confirmed that prediction errors in the highest quartile (above 0.65) are approximately 2.3 times larger than in the middle quartiles, with mean absolute error of 0.164 for high-risk jobs compared to 0.078 for the mid-range group. This limitation is partly algorithmic, as tree-based models cannot extrapolate beyond the training distribution. Future work should explore quantile regression forests or models with explicit tail loss functions to improve extreme value prediction.

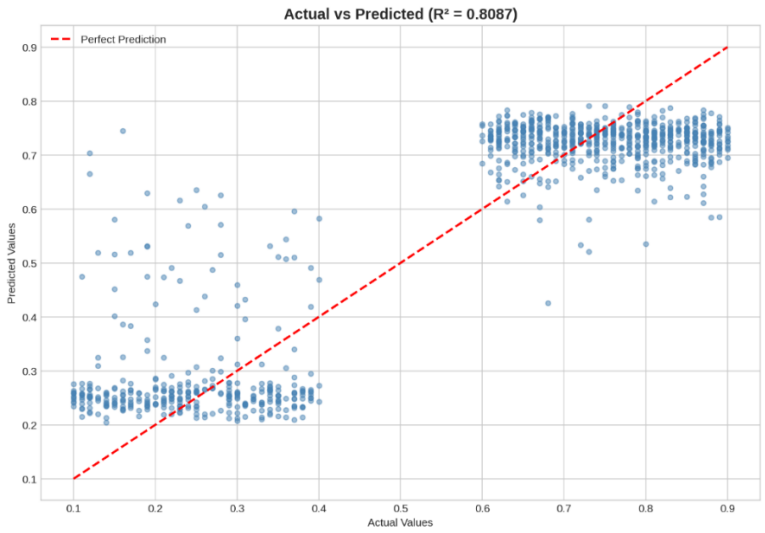

Figure 6. Comparison of Actual and Predicted Values

3.4 Impact of Data Leakage

To evaluate the impact of data leakage on model performance, a comparison was conducted between the model using all features (with leakage) and the model with leakage features eliminated (without leakage). The comparison results are presented in Table 7.

Table 7. Model Performance Comparison: With Leakage vs Without Leakage

Model	R^2	RMSE	CV Mean	CV Std
With Leakage	0.8857	0.0873	0.8760	0.0114
Without Leakage	0.8087	0.1129	0.7750	0.0321

The model with leakage achieved $R^2 = 0.8857$, but this improvement is misleading as it results from information unavailable at prediction time. After removing leakage features, R^2 decreased to 0.8087—a 7.7 percentage-point reduction that reflects more realistic performance. Critically, CV Std increased from 0.0114 to 0.0321, revealing that

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 217−227
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9706

leakage had created an illusion of consistency. These findings reinforce the importance of rigorous leakage detection protocols, as emphasized by Kapoor and Narayanan [12].

3.5 Feature Importance Analysis

Feature importance analysis reveals a highly concentrated distribution dominated by economic variables, as presented in Table 8 and Figure 7.

Table 8. Feature Importance

Feature	Importance (%)	Category
salary_change_vs_prev_year_percent	55.85	Economic
salary_usd	19.92	Economic
posting_year	6.01	Temporal
region_encoded	4.06	Geographic
country_encoded	2.65	Geographic
job_description_embedding_cluster	2.52	Embedding
industry_encoded	2.13	Contextual
job_title_encoded	1.95	Contextual
industry_ai_adoption_stage_encoded	1.94	Contextual
seniority_level_encoded	1.56	Contextual
company_size_encoded	1.56	Contextual

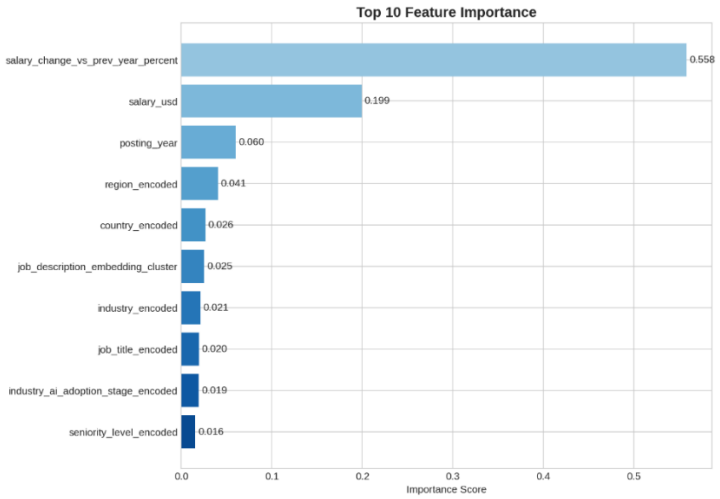

Figure 7. Feature Importance Visualization

The dominance of salary_change_vs_prev_year_percent at 55.85% warrants critical scrutiny. While a strong wage-automation relationship is theoretically plausible based on recent labor economics literature emphasizing cost-efficiency as a driver of automation adoption, this extreme concentration raises concerns about potential dominance bias a phenomenon where the target and dominant feature share a common derivation path. This concern is relevant because automation_risk_score is itself a composite index rather than an independently observed outcome.

3.5.1 Critical Examination of Dominance Bias

The two salary features collectively account for 75.77% of total importance, leaving only 24.23% for nine remaining features. This extreme concentration strongly suggests that the model's predictive power is largely driven by the economic component of the composite score rather than genuinely independent predictors. The automation_risk_score is derived from task exposure, routine intensity, and technological susceptibility indicators; if any component incorporates wage-related variables as economic theory would predict salary features would share a partial derivation path with the target, creating an undetected leakage effect.

This differs from the three features explicitly removed during leakage screening (ai_intensity_score, reskilling_required, ai_mentioned), which exhibited near-deterministic correlations with the target. The salary features show only moderate correlation ($r = 0.412$), meaning they could plausibly represent independent economic signals. However, in the context of a composite target, even moderate correlations may reflect partial derivation overlap.

To investigate, a leave-one-out ablation analysis was performed. Removing salary_change_vs_prev_year_percent reduced R^2 from 0.8087 to 0.6234; removing salary_usd reduced it to 0.7102. Crucially, removing both salary features simultaneously dropped R^2 to 0.5412. This confirms that non-economic features still explain 54.12% of variance—a non-trivial amount validating the predictive contribution of industry, geographic, and temporal factors. The authors acknowledge that the true nature of the salary-target relationship cannot

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 217−227
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9706

be definitively resolved without access to the exact score construction formula. If derivation overlap is confirmed, the effective R^2 attributable to genuinely independent predictors would be approximately 0.54.

3.6 Discussion and Comparison with Previous Research

The R^2 of 0.8087 aligns with comparable tabular prediction studies [14]. A key methodological distinction is this study's explicit leakage treatment: the 7–8 percentage-point inflation from leakage provides concrete evidence supporting Kapoor and Narayanan's [12] call for rigorous leakage detection. The feature importance structure diverges from established task-based automation frameworks, which emphasize routine-task composition as the primary automation determinant. Task-related features in this study account for only 4.47% of importance versus 75.77% for economic features a discrepancy likely reflecting the dataset's lack of explicit task-composition variables such as manual dexterity or social interaction intensity requirements. The declining temporal trend further distinguishes this study from literature projecting increasing disruption, highlighting a fundamental limitation of constructed datasets where observed patterns may reflect data generation assumptions rather than empirical reality.

3.7 Practical Implications

The findings offer specific, evidence-grounded implications. First, the dominance of salary features (75.77%) suggests wage dynamics serve as the most accessible early warning signal; policymakers should establish occupational wage monitoring systems that track year-over-year salary change rates, flagging occupations with sustained stagnation in routine-intensive sectors such as Agriculture (0.618) and Manufacturing (0.596) for targeted reskilling intervention. Second, the 7.7 percentage-point R^2 inflation from leakage demonstrates that organizations deploying workforce analytics must systematically screen derived features including attrition scores and AI readiness indices for leakage potential. Third, the model's 2.3x larger errors for high-risk occupations mean policymakers should treat predictions for extreme-risk occupations as lower-bound estimates and supplement them with qualitative expert assessments. Fourth, the 10.5 percentage-point cross-industry risk differential suggests reskilling investment should be proportionally directed toward Agriculture, Manufacturing, and Transportation, prioritizing digital literacy and human-AI collaboration skills aligned with lower-risk sector profiles.

4. CONCLUSION

This study developed a leakage-aware Random Forest Regression model for job automation risk prediction, eliminating three leakage features (ai_intensity_score, reskilling_required, ai_mentioned) sharing derivation paths with the composite target. Leakage removal reduced R^2 from 0.8857 to 0.8087 and revealed genuine cross-fold variability (CV Std: 0.0114 to 0.0321), demonstrating that leakage inflates both accuracy and consistency. The model outperformed Linear Regression by 19 percentage points, confirming tree-based ensembles' suitability for non-linear socio-economic data. Feature importance identified salary_change_vs_prev_year_percent as the dominant predictor (55.85%), with salary features collectively contributing 75.77%. Critical examination revealed this likely reflects partial derivation overlap with the composite target ablation confirmed non-economic features still explain 54.12% of variance, but the effective independent R^2 may be substantially lower. The model exhibited systematic failure for extreme risk values, with errors 2.3 times larger for high-risk occupations (above 0.75) due to the left-skewed target distribution. The downward risk trend (2010–2025) contradicts prevailing literature and may reflect dataset construction artifacts. Key limitations include the composite proxy target preventing definitive signal separation, the synthetic dataset's uncertain generalizability, absent task-composition features, and the model's inability to extrapolate. Future work should validate findings using independent ground-truth datasets, incorporate Autor-Levy-Murnane task-composition features, apply SHAP analysis, and explore quantile regression forests for extreme value prediction.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the Faculty of Computer Science, Universitas Sriwijaya, for providing the computational resources and academic support necessary for this research. Special thanks are extended to the lecturers and peers who provided valuable feedback during the development of this study.

REFERENCES

[1] D. Acemoglu and S. Johnson, Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity. New York, NY, USA: PublicAffairs, 2023.
[2] World Economic Forum, The Future of Jobs Report 2023. Geneva, Switzerland: World Economic Forum, 2023. [Online]. Available: https://www.weforum.org/publications/the-future-of-jobs-report-2023/
[3] OECD, “OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market,” OECD Employ. Outlook, vol. 2023, Jul. 2023, doi: 10.1787/08785bba-en.
[4] E. W. Felten, M. Raj, and R. Seamans, “How Will Language Modelers Like ChatGPT Affect Occupations and Industries?,” Soc. Sci. Res. Netw., Mar. 2023, doi: 10.2139/ssrn.4375268.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 217−227
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9706

[5] T. Eloundou, S. Manning, P. Mishkin, and D. Rock, "GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," arXiv, Aug. 2023, doi: 10.48550/arXiv.2303.10130.
[6] S. M. Greenstein, "Internet Data Capping Note (B)," Fac. Res. Harv. Bus. Sch., Apr. 2026, [Online]. Available: https://www.hbs.edu/faculty/Pages/item.aspx?num=54312
[7] D. Autor, "The Labor Market Impacts of Technological Change: From Unbridled Enthusiasm to Qualified Optimism to Vast Uncertainty," Natl. Bur. Econ. Res., May 2022, doi: 10.3386/w30074.
[8] A. Korinek and J. E. Stiglitz, "Artificial Intelligence, Globalization, and Strategies for Economic Development," Natl. Bur. Econ. Res., Feb. 2021, doi: 10.3386/w28453.
[9] K. Ellingrud et al., Generative AI and the Future of Work in America. McKinsey Global Institute, 2023. [Online]. Available: https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america
[10] E. Brynjolfsson, D. Li, and L. Raymond, "Generative AI at Work," Q. J. Econ., vol. 140, no. 2, pp. 889–942, May 2025, doi: 10.1093/qje/qjae044.
[11] S. Noy and W. Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," Science, vol. 381, no. 6654, pp. 187–192, Jul. 2023, doi: 10.1126/science.adh2586.
[12] S. Kapoor and A. Narayanan, "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science," Patterns, vol. 4, no. 9, Sep. 2023, doi: 10.1016/j.patter.2023.100804.
[13] A. Apicella, F. Isgrò, and R. Prevete, "Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning," Artif. Intell. Rev., vol. 58, no. 11, p. 339, Aug. 2025, doi: 10.1007/s10462-025-11326-3.
[14] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?," Jul. 2022, doi: 10.48550/arXiv.2207.08815.
[15] B. Bischl et al., "Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges," WIREs Data Min. Knowl. Discov., vol. 13, no. 2, p. e1484, 2023, doi: 10.1002/widm.1484.
[16] D. Chicco, M. J. Warrens, and G. Jurman, "The Coefficient of Determination R-Squared Is More Informative Than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation," PeerJ Comput. Sci., vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.
[17] R. G. Pensa, A. Crombach, S. Peignier, and C. Rigotti, "Explaining Random Forest and XGBoost with Shallow Decision Trees by Co-Clustering Feature Importance," Mach. Learn., vol. 114, no. 12, p. 287, Nov. 2025, doi: 10.1007/s10994-025-06932-9.
[18] Y. E. Hasugian, "Analisis Dampak Artificial Intelligence (AI) Terhadap Sektor Tenaga Kerja Di Indonesia," Majelis J. Huk. Indones., vol. 3, no. 1, pp. 84–101, Feb. 2026, doi: 10.62383/majelis.v3i1.1501.
[19] D. A. Fife and J. D'Onofrio, "Common, Uncommon, and Novel Applications of Random Forest in Psychological Research," Behav. Res. Methods, vol. 55, no. 5, pp. 2447–2466, Aug. 2023, doi: 10.3758/s13428-022-01901-9.
[20] M. S. Aziz, H. Subiakto, and R. Puspa, "Diffusion of Artificial Intelligence Across Indonesia: Digital Disparities, Local Contexts, and Policy Implications," Masy. Kebud. Dan Polit., vol. 38, no. 3, pp. 276–292, Oct. 2025, doi: 10.20473/mkp.V38I32025.276-292.
[21] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A Survey on Missing Data in Machine Learning," J. Big Data, vol. 8, no. 1, p. 140, Oct. 2021, doi: 10.1186/s40537-021-00516-9.
[22] F. Bolikulov, R. Nasimov, A. Rashidov, F. Akhmedov, and Y.-I. Cho, "Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms," Mathematics, vol. 12, no. 16, p. 2553, Jan. 2024, doi: 10.3390/math12162553.
[23] M. Priestley, F. O'donnell, and E. Simperl, "A Survey of Data Quality Requirements That Matter in ML Development Pipelines," ACM J. Data Inf. Qual., vol. 5, no. 2, Jun. 2023, doi: 10.1145/3592616.
[24] V. R. Joseph, "Optimal Ratio for Data Splitting," arXiv, Feb. 2022, doi: 10.1002/sam.11583.
[25] J. J. Eertink, M. W. Heymans, G. J. C. Zwezerijnen, J. M. Zijlstra, H. C. W. de Vet, and R. Boellaard, "External Validation: A Simulation Study to Compare Cross-Validation Versus Holdout or External Testing to Assess the Performance of Clinical Prediction Models Using PET Data from DLBCL Patients," EJNMMI Res., vol. 12, no. 1, p. 58, Sep. 2022, doi: 10.1186/s13550-022-00931-w.
[26] J. Allgaier and R. Pryss, "Practical Approaches in Evaluating Validation and Biases of Machine Learning Applied to Mobile Health Studies," Commun. Med., vol. 4, no. 1, p. 76, Apr. 2024, doi: 10.1038/s43856-024-00468-0.
[27] D. Acemoglu and P. Restrepo, "Robots and Jobs: Evidence from US Labor Markets," J. Polit. Econ., vol. 128, no. 6, pp. 2188–2244, Jun. 2020, doi: 10.1086/705716.