



## Enhancing Transparency in Income Prediction: A SHAP-based Explainable AI Approach with Random Forest

Marsani Asfi<sup>1</sup>, Budi Warsito<sup>2</sup>, Adi Wibowo<sup>3</sup>

<sup>1</sup>Doctoral Program of Information System, School of Postgraduate Studies, Diponegoro University, Semarang, Indonesia

<sup>1</sup>Dept. of Information System, Faculty of Information Technology, Catur Insan Cendekia University, Cirebon, Indonesia

<sup>2</sup>Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia

<sup>3</sup>Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia

\*[marsani.asfi@cic.ac.id](mailto:marsani.asfi@cic.ac.id)

### Abstract

*Explainable Artificial Intelligence (XAI) has gained increasing importance as modern machine learning models operate as opaque black-box systems, limiting transparency and trust in automated decision-making. While most XAI frameworks merely focus on static post-hoc visualizations, this study goes beyond passive diagnostics by introducing an algorithmic recourse framework for income prediction. By integrating SHAP (SHapley Additive exPlanations) with an optimized Random Forest classifier trained on the UCI Adult Income dataset, we partition socio-economic attributes into actionable features (e.g., weekly working hours, education level) and immutable demographic variables (e.g., age, sex). The dataset was preprocessed using missing-value handling, categorical encoding, class balancing via SMOTE, and feature scaling. Hyperparameter tuning optimization determined that the configuration of 150 trees, a maximum depth of 15, and a minimum split size of 15 yielded peak training accuracy rates of 85.49%, 86.13%, and 86.05% respectively. The Random Forest model achieved strong predictive performance, yielding an accuracy of 84.68%, precision of 85.15%, recall of 84.68%, F1-score of 84.88%, and a high ROC-AUC of 90.95%. Global SHAP analysis identified relationship, marital status, and education as the primary drivers of income classification. Crucially, local instance-level counterfactual optimization demonstrates that strategically adjusting actionable features can successfully neutralize negative SHAP momentum and flip unfavorable low-income predictions ( $\leq 50K$ ) into positive outcomes ( $> 50K$ ) while holding immutable traits perfectly constant. This integration demonstrates that combining Random Forest with an algorithmic recourse pipeline transforms AI from a descriptive transparency tool into an active, prescriptive mechanism that empowers data-driven decision-making and human solutions in socio-economic applications.*

**Keywords:** explainable artificial intelligence; explainability model; income classification; random forest; SHAP

*How to Cite:* M. Asfi, B. Warsito, and A. Wibowo, "Enhancing Transparency in Income Prediction: A SHAP-based Explainable AI Approach with Random Forest", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 5, pp. 1215–1223, Sep. 2026, doi: [10.29207/resti.v10i5.7461](https://doi.org/10.29207/resti.v10i5.7461).  
*Permalink/DOI:* <https://doi.org/10.29207/resti.v10i5.7461>

*Received:* January 9, 2026

*Accepted:* September 5, 2026

*Available Online:* September 28, 2026

*This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)*

*Published by [Ikatan Ahli Informatika Indonesia](https://www.lkatanahliinformatika.com/)*

### 1. Introduction

Explainable Artificial Intelligence (XAI) has become an essential research area in response to the lack of transparency associated with complex machine learning (ML) models that often operate as opaque "black-box" systems. By providing interpretable insight into model decision mechanisms, XAI enables stakeholders to understand how predictions are generated and why particular outputs occur. The central objective of XAI is to improve interpretability and support trust in ML-driven decision processes across various application domains [1].

One of the most widely recognized methods in XAI is SHAP (Shapley Additive Explanations), a framework

derived from cooperative game theory that attributes each feature's contribution to a model prediction fairly and accurately [2], [3]. Built upon Shapley value theory originally introduced by Lloyd Shapley in 1953 to allocate collective gains proportionally among contributors [4-6]. SHAP applies the same fairness principle to machine learning, where features act as players and model predictions represent the total payout. SHAP has gained prominence due to its strong mathematical foundation, reliability, and suitability for explaining complex ML models, establishing it as a core approach for model interpretation. SHAP-based explainability also has been shown to reliably identify key predictive factors while improving the transparency

and trustworthiness of machine learning models in high-stakes decision-making contexts [7].

From a theoretical perspective, SHAP adheres to fundamental properties such as local accuracy, efficiency, and consistency, ensuring equitable attribution of feature importance based on their contribution to prediction outcomes [3]. In practical implementations, SHAP provides TreeExplainer for efficient computation of SHAP values in tree-based models such as Random Forest and Gradient Boosting [8]. Moreover, SHAP is model-agnostic, supporting interpretability across diverse ML architectures and enabling both global and local explanations to interpret overall model behaviour and individual prediction reasoning [8], [9]. The technique enhances transparency and trust in high-stakes sectors such as healthcare and finance, where justification of model decisions is crucial [10], [11]. Empirical studies also demonstrate SHAP's effectiveness in data-driven decision-making and alignment with interpretability evaluation metrics [12], [13], while increasing compliance with regulatory frameworks that require transparent and traceable AI systems [13], [14].

As a flexible and analytically robust interpretability technique, SHAP has been widely used due to its reliability in generating stable visual explanations across different ML models [12], [15], [16]. Its consistency enables identification of influential features, facilitating deeper understanding of model predictions in sensitive domains such as Facial Emotion Recognition (FER) and medical diagnostics [12], [15] and predicting maternal health risks by integrating ensemble learning methods with SHAP (Shapley Additive exPlanations) for interpretability [17]. In autism spectrum disorder (ASD) classification research, for instance, SHAP identified TG, AGE, and LDL as dominant predictors influencing classification outcomes [18], highlighting its practical capability to reveal critical feature interactions.

Despite its strengths, SHAP is not without limitations. The computational cost of SHAP value generation can become substantial for large datasets and highly complex model architectures, and its underlying assumption of feature independence can reduce explanation reliability when multicollinearity exists [3] [12]. Nonetheless, its balance of interpretability and predictive performance sustains SHAP's role as a central method in contemporary XAI research. Comparative studies frequently examine SHAP alongside alternative frameworks such as LIME (Local Interpretable Model-Agnostic Explanations), showing that SHAP generally offers more consistent and trustworthy visual explanations, particularly in complex domains like facial emotion recognition and hyperspectral image analysis [15], [19]. While both enhance interpretability, SHAP demonstrates superior robustness across diverse implementations, including file fragment classification and predictive maintenance in data centres [20], [21].

Unlike prior work such as Hu et al. (2023)[22], which applies XAI primarily to strengthen trust and support iterative model improvement in customer segmentation, the present study focuses on SHAP-based interpretability within a socio-economic income prediction context. Specifically, the study integrates Random Forest with SHAP to achieve clear, traceable explanations without requiring multi-stage model retraining. This research applies SHAP to analyse decision behaviour in a Random Forest model trained on the Adult Income Dataset, prioritizing transparency over raw predictive accuracy [23]. Comprehensive global and local SHAP visualizations are presented to evaluate feature influence in income classification, providing a framework that promotes stakeholder understanding and facilitates bias detection. The findings contribute empirical evidence demonstrating SHAP's value in enhancing accountability and interpretability in socio-economic decision-support systems, establishing a practical bridge between performance and explainability in machine learning.

Although the Adult Income Dataset is a legacy benchmark widely used in explainability research, prior studies predominantly utilize XAI methods like SHAP merely as post-hoc diagnostic tools to display static feature importance. This leaves a critical gap in prescriptive analytics: knowing why a model predicts a low income does not inherently provide individuals with a clear pathway to improve their socio-economic classification. To bridge this gap, this study leverages the Adult dataset as a controlled environment to establish an algorithmic recourse framework. The specific novelty of this work lies in utilizing local SHAP value calculations (Waterfall and Force plots) to map out data-driven, personalized recommendations. By isolating actionable attributes (e.g., weekly working hours) from immutable demographic features (e.g., age, sex), we demonstrate how individuals can systematically identify the minimum lifestyle or educational adjustments required to strategically reverse a sub-optimal model prediction.

## 2. Methods

### 2.1 Research Approach

As shown in Figure 1, this research adopts an experimental methodology aimed at interpreting the decision-making process of machine learning models trained to predict income categories using the Adult Income Dataset. The study integrates XAI through SHAP to enhance transparency and interpretability of the predictive model.

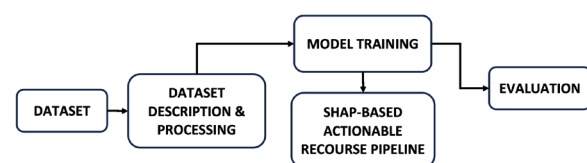


Figure 1. Research Methodology Workflow

## 2.2 Dataset Description and Preprocessing

This study utilizes the Adult Income Dataset (also referred to as the *Census Income Dataset*) obtained from the UCI Machine Learning Repository, consisting of 48,842 anonymized records derived from the 1994 U.S. Census [23]. The dataset contains fourteen demographic and socio-economic attributes, including age, workclass, education, education-num, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours-per-week, and native-country, along with a binary target variable indicating whether an individual's annual income exceeds US\$50,000. The primary modelling objective is a binary classification task for predicting income category (>50K vs ≤50K).

Prior to model development, the dataset underwent a comprehensive data preprocessing stage to address issues such as missing values, inconsistent data types, outliers, and the absence of column headers. Several numerical features were originally stored as object-type fields and required appropriate type conversion, while scaling was applied to ensure consistency across numerical variables. Ensuring high-quality input data is essential for producing machine learning models that are interpretable and reliable, particularly within XAI frameworks.

## 2.3 Model Training and Evaluation

After completing data preprocessing, including handling missing values and preparing numerical and categorical features, the next step was model development using the Random Forest algorithm. Random Forest-based ensemble models have been widely adopted for classification on imbalanced data due to their robustness and generalization capability, and SHAP has been effectively used to provide feature-level explanations by quantifying each variable's contribution to the model's predictions, thereby enhancing interpretability without sacrificing predictive performance [24]. The model was trained on the processed training dataset and optimized through hyperparameter tuning with cross-validation techniques to achieve the best predictive performance while minimizing the risk of overfitting. Once the optimal configuration was obtained, the model was evaluated using an unseen testing set to objectively assess its generalization capability.

Model performance was assessed using standard evaluation metrics, namely accuracy, precision, recall, and F1-Score, calculated based on the classification results as shown in Equations 1 - 4.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (2)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (3)$$

$$\text{F1 - Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4)$$

In these equations, TP denoted true positives, TN denoted true negatives, FP denoted false positives, and FN denoted false negatives. Accuracy reflects overall correctness of predictions. Precision measures the proportion of true positive results among positive predictions. Recall indicates the model's sensitivity in detecting actual positive cases, and F1-Score represents the harmonic balance between precision and recall.

## 2.4 Explainability Model

To enhance explainability and provide insight into the decision-making process of the Random Forest model, XAI was applied using SHAP.

SHAP allows both global and local explanations through visualization, showing the contribution of each feature to the final prediction and enabling a transparent understanding of which input variables most strongly influence the classification outcomes [2].

## 2.5 Algorithmic Recourse Setup

To transform the post-hoc explanations provided by SHAP into prescriptive and operational solutions, this study establishes an algorithmic recourse framework. Algorithmic recourse aims to identify the minimum, realistic adjustments an individual can make to reverse an unfavorable model prediction (i.e., shifting the classification from ≤50K to >50K).

The baseline formulation as presented in Equation 5 begins by partitioning the complete feature space  $X$  into two distinct subsets based on their real-world mutability:

$$\text{Features}(X) = \{X_{\text{actionable}}\} \cup \{X_{\text{immutable}}\} \quad (5)$$

$\{X_{\text{immutable}}\}$  represents attributes that cannot be altered by individual intervention or are socio-demographically fixed within short-term planning horizons. In the context of the Adult Income dataset, this subset includes features such as age, sex, race, native-country, relationship, and marital-status.

$\{X_{\text{actionable}}\}$  denotes attributes that an individual can strategically modify through actionable effort or lifestyle adjustments. This subset encompasses variables such as education-num (attainment through training/schooling), hours-per-week (labor participation intensity), and workclass (employment sector).

Mathematically, given an individual instance  $x$  where the optimized Random Forest model yields an unfavorable prediction  $f(x) < \theta$  (where  $\theta = 0.5$  is the classification threshold), the objective of the recourse framework is to find a counterfactual perturbation vector  $\Delta X_{\text{actionable}}$  that satisfies the following conditions as presented in Equations 6-8.

$$\text{Minimize } \|\Delta X_{\text{actionable}}\|_2 \quad (6)$$

$$\text{Subject to: } f(x + \Delta X_{\text{actionable}}) \geq \theta \quad (7)$$

$$\text{and } X_{\text{immutable}} = \text{constant} \quad (8)$$

The optimization workflow utilizes local SHAP calculations (derived from Waterfall and Force plots) as

a directional guide. Instead of employing an exhaustive random search, the algorithm analyzes the sign and magnitude of the local SHAP values for all variables within  $X_{\text{actionable}}$ . If an actionable feature exhibits a negative SHAP value, it signifies that the current state of this feature is suppressing the high-income prediction.

The framework systematically computes the required incremental shift ( $\Delta$ ) for these specific features. By keeping the immutable socio-demographic constraints perfectly constant, the system isolates and simulates "What-If" scenarios. This pipeline ensures that the generated recommendations are not only mathematically optimal for confusing the model's boundary but are also practically feasible for human execution.

### 3. Results and Discussions

This section presents the experimental results of the Random Forest model and the SHAP-based explainability analysis. The findings are discussed from both predictive performance and interpretability perspectives

#### 3.1 Dataset Description and Preprocessing

This research used the Adult Income dataset from the UCI Machine Learning Repository, which initially includes 48,842 samples describing demographic and occupational characteristics related to income classification ( $\leq 50K$  or  $>50K$ ) [23]. After preprocessing steps such as managing missing values, encoding categorical variables, balancing classes, and scaling numerical features, the data was prepared for model training and interpretability evaluation. The original dataset is distributed across two files (adult.data and adult.test). By removing meta-information, standardizing label formats, and utilizing only the primary training file (adult.data), the final dataset used in this study consists of 32,561 records and 15 feature.

The dataset exhibits a class imbalance, with the majority class ( $\leq 50K$ ) significantly outnumbering the minority class ( $>50K$ ). To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied during the preprocessing stage. SMOTE generates synthetic instances of the minority class through nearest-neighbor interpolation, resulting in a more balanced class distribution [25]. By mitigating the bias toward the majority class, this approach enhances the model's ability to learn representative patterns from both income groups and contributes to more reliable classification performance.

#### 3.2 Model Training and Performance Evaluation

A Random Forest classifier was developed using the pre-processed dataset, with model optimization conducted through cross-validation-based hyperparameter tuning to enhance predictive performance. The tuning process evaluated multiple configurations of key parameters, including the number

of trees ( $n_{\text{estimators}}$ ), tree depth ( $\text{max\_depth}$ ), and minimum samples required for an internal split ( $\text{min\_samples\_split}$ ). The results indicated that the optimal configuration was achieved using 150 trees, which yielded an accuracy of 85.49% as shown in Figure 2, a maximum depth of 15 resulting in an accuracy of 86.13% as shown in Figure 3, and a minimum split size of 15 as shown in Figure 4, producing an accuracy of 86.05%. These optimal values were subsequently applied to train the final model, ensuring a balance between predictive performance and overfitting control.

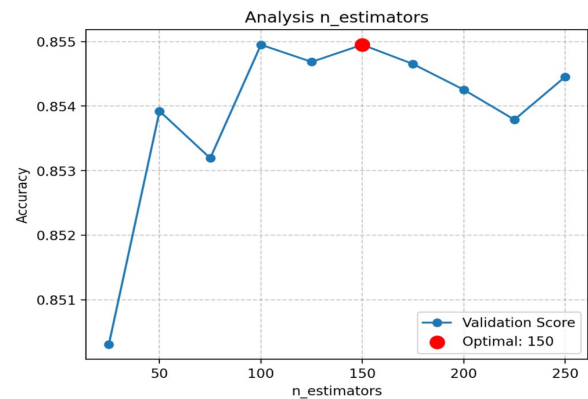


Figure 2. Optimal  $n_{\text{estimators}}$ : 150 (Acc: 85.49%)

Evaluation metrics including accuracy, precision, recall as shown in Equations 1-4, and AUC-ROC were calculated on the test dataset to assess classification effectiveness.

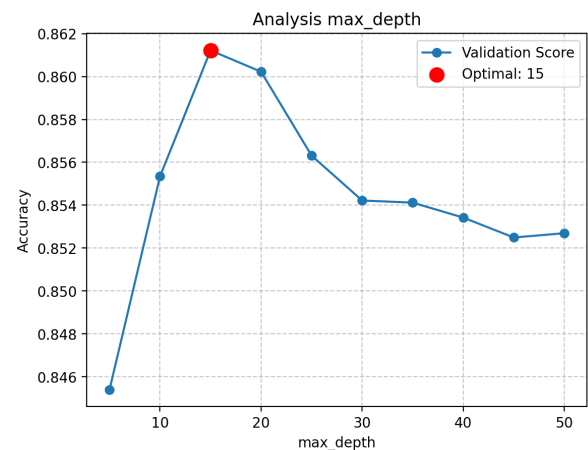


Figure 3. Optimal  $\text{max\_depth}$ : 15 (Acc: 86.13%)

The performance evaluation results indicate that the Random Forest model achieved strong classification performance for income prediction. As shown in Table 1, the model obtained an accuracy of 84.68%, meaning that most instances were correctly classified into their respective income categories.

Table 1. Random Forest Model Performance

| Metric        | Metric Evaluation |           |        |          |
|---------------|-------------------|-----------|--------|----------|
|               | Accuracy          | Precision | Recall | F1-Score |
| Random Forest | 0.8468            | 0.8515    | 0.8468 | 0.8488   |

The model achieved a precision of 85.15%, indicating that the majority of individuals predicted as belonging to the high-income group (>50K) were correctly identified. In addition, the recall of 84.68% shows that the model successfully detected a large proportion of actual high-income individuals. The similar values of precision and recall suggest a balanced classification performance.

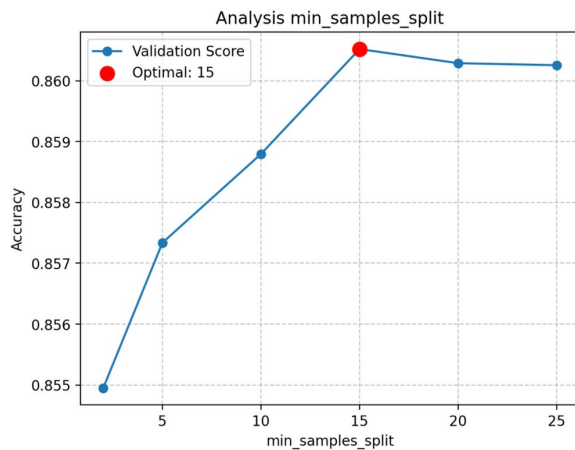


Figure 4. Optimal min\_samples\_split: 15 (Acc: 86.05%)

The F1-Score of 84.88% further confirms the model's consistency by balancing precision and recall.

Moreover as shown in Figure 5, the model achieved a ROC-AUC score of 0.9095 (90.95%), which indicates excellent ability to distinguish between the two income classes. A ROC-AUC value above 0.90 demonstrates that the model can effectively separate individuals with income above and below \$50,000.

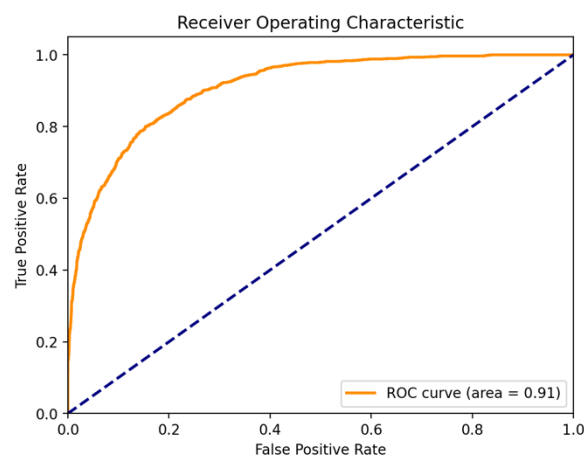


Figure 5. ROC Curve for Random Forest Classifier

The high ROC-AUC score indicates that the model has strong discriminative power, while the balanced precision and recall values suggest that it does not favor one class over the other. Therefore, the SHAP explanations generated from this model can be considered reliable for interpreting the factors that influence income classification.

Overall, these results show that the Random Forest model provides reliable predictive performance and

serves as a strong foundation for the subsequent SHAP-based explainability analysis.

### 3.3 Explainability using SHAP Feature Importance

To enhance model transparency, SHAP were applied to determine global feature importance. Figure 6 illustrates the mean absolute SHAP values for all features, ranking them according to their contribution to model output.

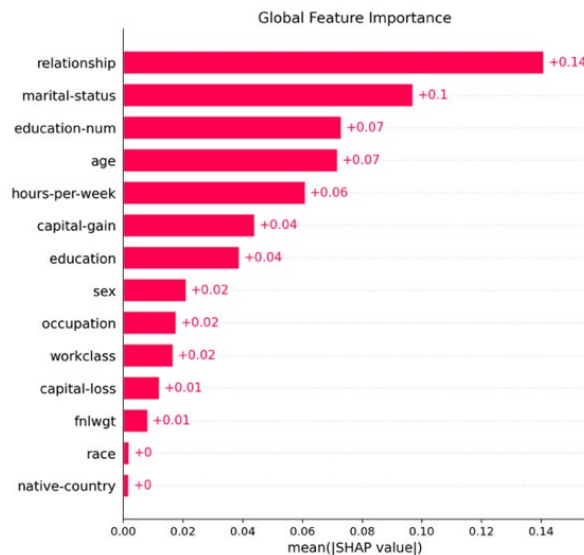


Figure 6. SHAP Global Feature Importance

The SHAP global feature importance analysis as shown in Figure 6 reveals that relationship, marital-status, education-num, age, hours-per-week, and capital-gain are the most influential predictors of income classification. Among these features, relationship exhibits the highest mean absolute SHAP value, indicating that an individual's family or household role has the strongest overall influence on the model's predictions.

The dominance of *relationship* and *marital-status* suggests that household structure and social status play an important role in determining income levels. Individuals who are married or belong to economically stable household structures may benefit from greater financial resources and labor-market opportunities. This finding is consistent with previous socioeconomic studies that identify family structure as an important determinant of income and economic well-being.

The strong contribution of *education-num* is consistent with Human Capital Theory, which argues that higher educational attainment increases productivity, skills, and earning potential. Likewise, *age* can be interpreted as a proxy for work experience and career progression, while *hours-per-week* reflects labor participation and employment intensity. The influence of *capital-gain* further indicates that investment returns and asset ownership contribute substantially to distinguishing higher-income individuals from lower-income groups.

Overall, the agreement between the SHAP results and established socioeconomic theories suggests that the

Random Forest model captures meaningful patterns in the Adult Income dataset rather than relying on spurious correlations. This strengthens confidence in both the predictive performance of the model and the reliability of the generated explanations.

A SHAP summary plot as shown in Figure 7 was generated to reveal how individual feature values shift predictions towards either class direction ( $\leq 50K$  or  $>50K$ ).

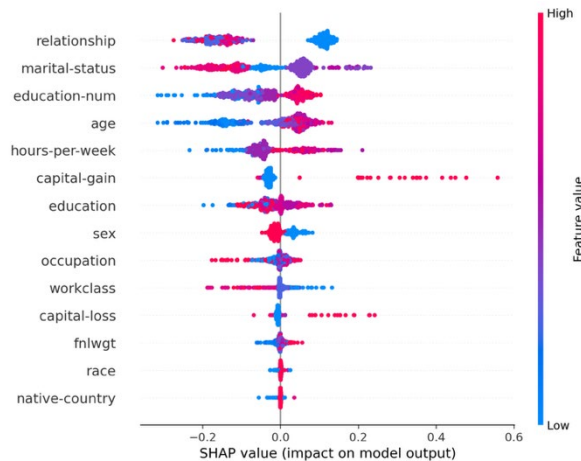


Figure 7. SHAP Summary

The distribution indicates heterogeneous contributions across instances, demonstrating that the model captures complex interactions between demographic and employment variables.

### 3.4 Local Instance-Level Explanations and Algorithmic Recourse Implementation

While global feature importance establishes an aggregate overview of model behavior, socio-economic decision-support systems require granular validation at the individual level. To demonstrate the transition from post-hoc explainability to actionable algorithmic recourse, this section evaluates individual predictions using SHAP Waterfall and Force plots, paired with counterfactual simulations on low-income instances.

Figure 8 presents a local SHAP Waterfall plot for a specific sample profile. The model initializes its decision path at the expected baseline value  $E[f(X)] = 0.5$ . For this particular individual, the final prediction reaches  $f(x) = 0.75$ , successfully placing them in the high-income category ( $>50K$ ).

The visual decomposition reveals that the individual's household structure, represented by 0 = relationship, provides the most dominant positive momentum, driving the score upward by +0.25. This is further reinforced by demographic attributes such as 0 = sex (+0.08) and institutional qualifications like 10 = education-num (+0.06). Conversely, the model captures restrictive pressures from 30 = age (-0.06) and 40 = hours-per-week (-0.04).

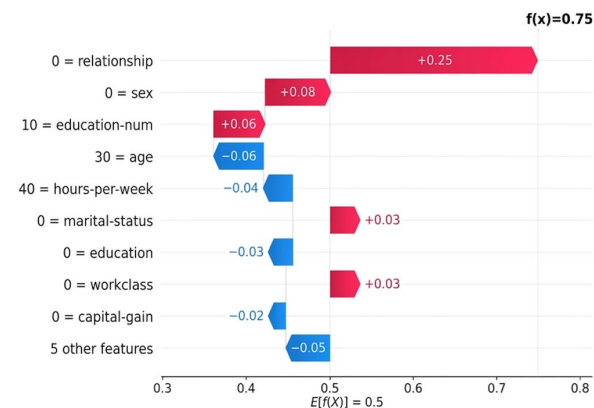


Figure 8. SHAP waterfall plot explanation for sample

Similarly, the SHAP Force Plot in Figure 9 visualizes another successful case with a prediction output  $f(x) = 0.693$ . The horizontal force interaction clearly depicts how the positive pushes from education-num = 10, workclass = 0, and relationship = 0 collectively outweigh the negative balancing forces exerted by age = 30 and hours-per-week = 40.

These high-income instances empirically validate that within the learned logic of the Random Forest model, educational thresholds and stable household roles act as powerful catalysts to surpass the baseline prediction boundary

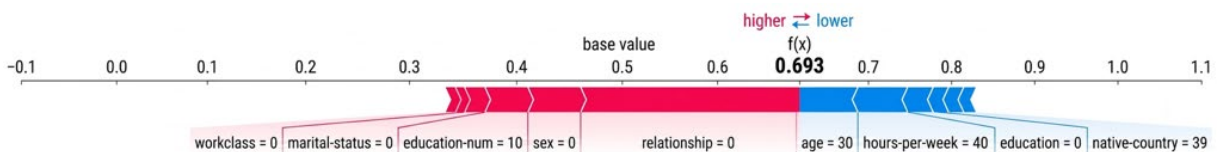


Figure 9. SHAP Force Plot for Individual Prediction

To operationalize the recourse framework, we analyze the sub-optimal profiles documented in Table 2, where all individuals initially failed to cross the high-income threshold ( $f(x) < 0.5$ ). Rather than treating these metrics as stagnant outcomes, the algorithmic recourse pipeline isolates the top negative features that are actionable, while keeping immutable variables structurally constant.

**Case 2 Analysis (Labor Intensity Intervention):** In Case 2, the individual is assigned a low high-income probability of 24.63%. Although their family role (relationship) contributes a positive momentum of +0.086, the prediction is severely suppressed by labor participation intensity (hours-per-week), which inflicts a massive negative SHAP value of -0.244. Since working hours are highly actionable  $X_{\text{actionable}}$ , the recourse algorithm simulates a "What-If" intervention. By keeping immutable demographic traits constant and

incrementally increasing hours-per-week from the current sub-standard level toward a full-time or overtime threshold, the negative SHAP momentum is systematically neutralized, shifting the final prediction past the  $\theta = 0.5$  boundary.

**Case 4 Analysis (Educational Attainment Intervention):** In Case 4, the model yields an income probability of only 16.40%. Despite receiving a positive push from their household status (relationship = +0.108), the individual's earning potential is constrained by a restricted educational background (education = -0.185). Utilizing the recourse framework, this negative asset is targeted for optimization. The simulation establishes that upgrading the educational qualification (equivalent to increasing the underlying education-num parameter) can mathematically counteract the structural deficits, providing a concrete, data-driven pathway for the individual to reverse the model's unfavorable verdict.

To geometrically map the structural transition of sub-optimal instances, Figure 10 visualizes the counterfactual optimization trajectories for both Case 2 and Case 4 against the baseline classification boundary. The horizontal axis delineates the model's high-income classification score, where a critical threshold line is mathematically fixed at a probability of 0.5. The visualization clearly separates the original failed state (red markers) located in the low-probability zone ( $\leq 0.5$ ) from the post-recourse success state (green markers) positioned within the high-income classification region ( $> 0.5$ ).

As illustrated in the trajectory of Case 2 (Labor Intensity Reciprocation), the individual initially resides deep within the unfavorable zone with a failed original prediction probability of 24.63%. By freezing all immutable demographic nodes and applying the recourse algorithm to shift the 'hours-per-week' parameter upward, the instance travels along a positive horizontal trajectory. This tactical labor escalation successfully neutralizes the negative SHAP pressures, effectively driving the counterfactual prediction score to cross the decision boundary up to a successful post-recourse probability of 51.5%.

A structurally identical pattern is observed in Case 4 (Educational Attainment Intervention). The initial state of the profile displays an even more severe deficit, stalling at a low-income probability of 16.40%. By utilizing local SHAP gradients as a directional vector, the recourse engine simulates an educational upgrade via the 'education-num' variable. While keeping inherent baseline demographics static, this targeted asset optimization successfully forces the prediction to bypass the threshold line, securing an inverted high-income probability of 55.2%.

A deeper comparative analysis of both trajectories reveals key differences in intervention magnitude and structural efficacy. Case 4 exhibits a substantially larger probability delta ( $\Delta P = +38.80\%$ ) compared to Case 2 ( $\Delta P = +26.87\%$ ), despite starting from a lower baseline score. This variance underscores the higher leverage capacity of educational attainment relative to labor hours within the Random Forest decision tree architecture, indicating that long-term structural human capital investments yield a stronger positive momentum shift per unit of change.

Furthermore, the trajectories demonstrate the algorithmic principle of minimal recourse cost. Rather than attempting to push predictions to extreme certainty levels (e.g.,  $> 0.90$ ), the recourse optimizer terminates immediately once the post-recourse instance lands safely beyond the decision threshold ( $\theta = 0.50$ ). This design constraint prevents unrealistic or unnecessarily costly recommendations for the user, ensuring that the prescribed behavioral modifications remain feasible and efficient.

Mapping counterfactual trajectories transforms passive model explainability into active prescriptive guidance by providing explicit, quantified targets to reverse low-income predictions. As demonstrated in Figure 10, embedding a counterfactual layer onto post-hoc SHAP metrics offers a practical mechanism to guide individuals across automated classification thresholds. Ultimately, this approach bridges the gap between diagnostic transparency and active socio-economic empowerment, turning automated auditing into actionable human solutions.

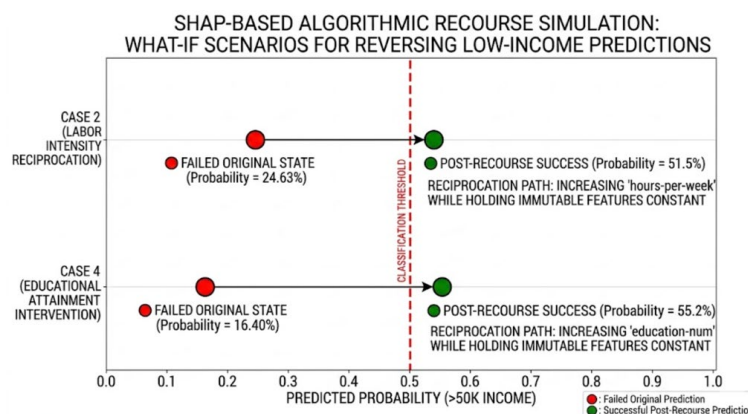


Figure 10. SHAP-Based Algorithmic Recourse Paths for Low-Income Profiles

The local SHAP results in Table 2 highlight how specific features influence individual predictions toward the higher-income class (>50K). The probability column indicates the model's estimated likelihood of an instance being classified as high income, while the positive and negative feature columns show the most influential contributors driving the prediction either upward or downward.

Table 2. Example Cases with Feature Contribution for Prediction <=50K

| No | Prob<br>(>50K) | Top Positive<br>Feature    | Top Negative<br>Feature    |
|----|----------------|----------------------------|----------------------------|
| 1  | 5.08%          | Age<br>(+0.034)            | Marital-status<br>(-0.171) |
|    | 24.63%         | Relationship<br>(+0.086)   | Hours-per-week<br>(-0.244) |
| 3  | 1.24%          | Age<br>(+0.030)            | Relationship<br>(-0.210)   |
| 4  | 16.40%         | Relationship<br>(+0.108)   | Education<br>(-0.185)      |
| 5  | 32.67%         | Marital-status<br>(+0.168) | Age<br>(-0.148)            |

Across the examples, relationship, age, marital-status, education and hours-per-week consistently appear as the strongest drivers. For instance, in case 5, the model predicts a 32.67% probability of earning >50K, primarily supported by marital-status with a positive SHAP contribution of +0.168, suggesting that being married (typically "Married-civ-spouse") strongly increases the likelihood of higher income. Conversely, age contributes negatively (-0.148), indicating that a younger age within this sample decreases the predicted income likelihood.

Similarly, Case 2 demonstrates that although relationship contributes positively (+0.086) toward a 24.63% probability of high income, the feature hours-per-week has a stronger negative influence (-0.244), implying insufficient working time is a major factor reducing income potential. In Case 4, education has a negative effect (-0.185), indicating lower education level reduces income probability despite positive impact from relationship status.

Overall, the results illustrate that SHAP enables detailed insight into why individual predictions vary, revealing that marital-status and relationship roles strongly support higher income predictions, while low working hours, younger age, or limited education reduce predicted earning potential. These instance-level explanations complement global feature importance and confirm the model's ability to capture realistic socio-economic patterns.

To position this study within existing research, a comparison is in [20] which examines XAI in machine learning for customer segmentation within product development. Their research employs feature-based and data-based explanation methods, including LIME and data-valued Shapley, to support iterative model refinement and enhance user trust by enabling designers to interpret model outputs throughout development cycles.

### 3.5 Limitations and Future Work

The SHAP-based explainability analysis effectively uncovered significant patterns related to factors influencing income classification and offered clear, transparent explanations for the decisions made by the Random Forest model. These results demonstrate that the integration of SHAP with Random Forest achieves not only robust predictive performance but also reliable interpretability, reinforcing the value of post-hoc explainability techniques in transforming complex black-box models into more understandable and trustworthy systems. Such transparency is essential for responsible and ethical adoption of AI solutions in socio-economic contexts where decision justification is required.

Despite these promising outcomes, several limitations remain. SHAP estimates can be affected by multicollinearity, which may impact the stability and consistency of feature attribution. Additionally, the results are dependent on the characteristics of the Adult Income dataset and may not be representative of different populations or time periods. It is also important to note that SHAP explanations reflect contribution within the learned model structure rather than causal relationships. Future work should incorporate fairness auditing, comparison with other XAI techniques such as LIME or counterfactual explanations enhance the interpretability of complex machine learning models by clarifying the rationale behind individual predictions and indicating minimal changes required to obtain alternative outcomes [26], and evaluation by domain experts to enhance practical usefulness and generalizability across broader socio-economic settings.

## 4. Conclusions

This study successfully expanded the utility of SHAP from a conventional post-hoc visualization technique into an active algorithmic recourse framework for income prediction using a Random Forest model. While the model achieved a strong discriminative power with a ROC-AUC of 0.9095, the primary contribution lies in the systematic deployment of personalized, actionable recommendations. By separating actionable socio-economic features from unalterable demographic traits, this framework provides traceable guidance on how individuals can realistically invert low-income predictions. Future work will focus on integrating automated counterfactual optimization libraries to standardize the recourse pathways across highly correlated variables.

## Declarations

### AI Usage Statement

The authors used Gemini solely to improve the readability and language of the manuscript; all scientific content, analysis, and conclusions are the authors' own.

### Author Contributions

M.A.: conceptualization, methodology, data curation, formal analysis, writing—original draft; B.W.: supervision—review and correction; A.W.: supervision—review and correction.

### Funding Statement

The authors received no specific research funding from public, commercial, or not-for-profit agencies.

### Conflict of Interest Statement

The authors declare no conflict of interest.

### Data and Code Availability

The dataset and code supporting the findings of this study are publicly available at <https://archive.ics.uci.edu/dataset/2/adult> and [https://github.com/marsani/Adult\\_Income\\_XAI](https://github.com/marsani/Adult_Income_XAI), respectively.

### References

- [1] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, "Explainable AI Methods - A Brief Overview," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2022, pp. 13–38. doi: [10.1007/978-3-031-04083-2\\_2](https://doi.org/10.1007/978-3-031-04083-2_2).
- [2] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017. doi: [10.48550/arXiv.1705.07874](https://doi.org/10.48550/arXiv.1705.07874).
- [3] M. Li, H. Sun, Y. Huang, and H. Chen, "Shapley value: from cooperative game to explainable artificial intelligence," *Autonomous Intelligent Systems*, vol. 4, no. 1, p. 2, Feb. 2024, doi: [10.1007/s43684-023-00060-8](https://doi.org/10.1007/s43684-023-00060-8).
- [4] P. Jamshidi, S. Nowaczyk, and M. Rahat, "Analysis of characteristic functions on Shapley values in Machine Learning," in *2024 International Conference on Intelligent Environments (IE)*, IEEE, Jun. 2024, pp. 70–77. doi: [10.1109/IE61493.2024.10599897](https://doi.org/10.1109/IE61493.2024.10599897).
- [5] L. Wu, "A review of the transition from Shapley values and SHAP values to RGE," *Statistics (Berl.)*, vol. 59, no. 5, pp. 1161–1183, Sep. 2025, doi: [10.1080/02331888.2025.2487853](https://doi.org/10.1080/02331888.2025.2487853).
- [6] B. Warner, E. Ratner, and A. Lendasse, "X-ELM: A Fast Explainability Approach for Extreme Learning Machines," 2023, pp. 411–422. doi: [10.1007/978-3-031-43078-7\\_34](https://doi.org/10.1007/978-3-031-43078-7_34).
- [7] J. Pinem, W. Astuti, and Adiwijaya, "Explainable Ensemble Learning Framework with SMOTE, SHAP and LIME for Predicting 30-Day Readmission in Diabetic Patients," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 5, pp. 983–990, Sep. 2025, doi: [10.29207/resti.v9i5.6977](https://doi.org/10.29207/resti.v9i5.6977).
- [8] P. S. R. Aditya and M. Pal, "Local Interpretable Model Agnostic Shap Explanations for machine learning models," Oct. 2022, doi: <https://doi.org/10.48550/arXiv.2210.04533>.
- [9] A. Scrivano, "Explainable Artificial Intelligence (XAI): Investigating Methods to Make AI Algorithms More Interpretable and Transparent," Aug. 2025, doi: [10.31224/5205](https://doi.org/10.31224/5205).
- [10] H. Byeon and S. B. Shaheema, "Explainable AI: Techniques and Tools for Interpreting Deep Learning Models," in *Cutting-Edge Technologies in Innovations in Computer Science and Engineering*, San International Scientific Publications, 2022, doi: [10.59646/csebookc10/004](https://doi.org/10.59646/csebookc10/004).
- [11] D. Kulaklıoğlu, "Explainable AI: Enhancing Interpretability of Machine Learning Models," *Human Computer Interaction*, vol. 8, no. 1, p. 91, Dec. 2024, doi: [10.62802/z3pde490](https://doi.org/10.62802/z3pde490).
- [12] A. M. Salih et al., "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," *Advanced Intelligent Systems*, vol. 7, no. 1, Jan. 2025, doi: [10.1002/aisy.202400304](https://doi.org/10.1002/aisy.202400304).
- [13] N. L. Rane, S. K. Mallick, Ö. Kaya, and J. Rane, "Explainable and trustworthy artificial intelligence, machine learning, and deep learning," in *Applied Machine Learning and Deep Learning: Architectures and Techniques*, Deep Science Publishing, 2024. doi: [10.70593/978-81-981271-4-3\\_9](https://doi.org/10.70593/978-81-981271-4-3_9).
- [14] M. Asfi, B. Warsito, and A. Wibowo, "Enhancing Explainable AI: Leveraging SHAP for Transparent Decision-Making in Machine Learning," in *2024 Ninth International Conference on Informatics and Computing (ICIC)*, IEEE, Oct. 2024, pp. 1–6. doi: [10.1109/ICIC64337.2024.10956244](https://doi.org/10.1109/ICIC64337.2024.10956244).
- [15] S. Lorch, J. Gebele, and P. Brune, "Towards Trustworthy AI: Evaluating SHAP and LIME for Facial Emotion Recognition," 2025. doi: [10.24251/HICSS.2025.900](https://doi.org/10.24251/HICSS.2025.900).
- [16] H. Malik, D. Dampier, and N. Jabeur, "Decoding Performance Testing Results: Empowering Trust with Explainable Artificial Intelligence (XAI)," in *NAECON 2023 - IEEE National Aerospace and Electronics Conference*, IEEE, Aug. 2023, pp. 116–121. doi: [10.1109/NAECON58068.2023.10365764](https://doi.org/10.1109/NAECON58068.2023.10365764).
- [17] L. Widyawati and N. Sulistianingsih, "Explainable Ensemble Learning for Maternal Health Risk in Low-Resource Settings," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 5, pp. 1121–1130, Oct. 2025, doi: [10.29207/resti.v9i5.6765](https://doi.org/10.29207/resti.v9i5.6765).
- [18] Y. Bokka, R. N. V. J. Mohan, and M. C. Naik, "SHAP-Driven Interpretability of Autism Risk in Pregnancy Using Explainable AI," in *2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS)*, IEEE, Nov. 2024, pp. 1–7. doi: [10.1109/ICIICS63763.2024.10859415](https://doi.org/10.1109/ICIICS63763.2024.10859415).
- [19] İ. Şahin, A. Ertürk, and E. Aptoula, "Band-Based Interpretability with SHAP for Hyperspectral Classification," in *2023 31st Signal Processing and Communications Applications Conference (SIU)*, IEEE, Jul. 2023, pp. 1–4. doi: [10.1109/SIU59756.2023.10223795](https://doi.org/10.1109/SIU59756.2023.10223795).
- [20] R. Jinad, A. Islam, and N. Shashidhar, "Interpretability and Transparency of Machine Learning in File Fragment Analysis with Explainable Artificial Intelligence," *Electronics (Basel)*, vol. 13, no. 13, p. 2438, Jun. 2024, doi: [10.3390/electronics13132438](https://doi.org/10.3390/electronics13132438).
- [21] Y. Gebreyesus, D. Dalton, D. De Chiara, M. Chinnici, and A. Chinnici, "AI for Automating Data Center Operations: Model Explainability in the Data Centre Context Using Shapley Additive Explanations (SHAP)," *Electronics (Basel)*, vol. 13, no. 9, p. 1628, Apr. 2024, doi: [10.3390/electronics13091628](https://doi.org/10.3390/electronics13091628).
- [22] X. Hu, A. Liu, X. Li, Y. Dai, and M. Nakao, "Explainable AI for customer segmentation in product development," *CIRP Annals*, vol. 72, no. 1, pp. 89–92, 2023, doi: [10.1016/j.cirp.2023.03.004](https://doi.org/10.1016/j.cirp.2023.03.004).
- [23] B. K. R. Becker, "Adult dataset (UCI Machine Learning Repository)," 1996. doi: <https://doi.org/10.24432/C5XW20>.
- [24] Rais, Agus Mohamad Soleh, and Budi Susetyo, "Rotation Double Random Forest Algorithm to Predict The Food Insecurity Status of Households," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 1, pp. 33–41, Jan. 2024, doi: [10.29207/resti.v8i1.5540](https://doi.org/10.29207/resti.v8i1.5540).
- [25] H. Hairani, T. Widiyaningtyas, and D. Dwi Prasetya, "Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *JOIV: International Journal on Informatics Visualization*, vol. 8, no. 3, p. 1310, Sep. 2024, doi: [10.62527/joiv.8.3.2283](https://doi.org/10.62527/joiv.8.3.2283).
- [26] S. Tonati, M. Di Vece, R. Pellungini, and F. Giannotti, "Ensemble Counterfactual Explanations for Churn Analysis," 2025, pp. 332–347. doi: [10.1007/978-3-031-78980-9\\_21](https://doi.org/10.1007/978-3-031-78980-9_21).