Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 13 No. 3 (2026) 555 - 566 e-ISSN: 2580-0760

# A Hybrid PCA, MPK Method for Multi-Objective Human Resource Analytics in Plantation Management

Gilang Ramadhan[1], Rahmad B.Y Syah[2]*, Ahmad Rafiki[3]
[1]Department of Magister Management, Universitas Medan Area, Medan. Indonesia
[2]Department of Informatics, Faculty of Engineering, Universitas Medan Area, Medan. Indonesia
[2]Department of Management, Faculty of Economic and Business, Universitas Medan Area, Medan. Indonesia

*rahmadsyah@uma.ac.id*

**Abstract**

*This study presents a hybrid framework combining Principal Component Analysis (PCA) and Multi-Objective Complexity Prediction (MPK) to extract actionable insights from multidimensional human resource (HR) data in plantation companies. Initially, PCA reduces nine original HR variables including age, tenure, absenteeism, leave, harvest yield, and performance scores to four principal components that capture over 90 % of total variance. These reduced features form the solution space for MPK agents, each of which simultaneously optimizes three objectives: absenteeism risk (predicted via SVR), performance score (via linear regression), and workload imbalance. The model is tuned with 5-fold cross-validation on 80 % of the data, yielding MAE 6.5 days and $R^2$ 0.82 for absenteeism, MAE 4.2 points and $R^2$ 0.78 for performance, MAPE 15 % for workload imbalance, and a Pareto-front hypervolume of 0.90. Validation on the remaining 20 % hold-out set confirms similar generalization (hypervolume 0.88). Sensitivity analysis with ± 10 % input noise provides initial evidence the approach's robustness under moderate perturbations. These results illustrate that PCA-MPK can both streamline high-dimensional HR datasets and deliver reliable, multi-objective forecasts to inform strategic workforce planning. These results demonstrate that PCA-MPK can simplify high-dimensional HR datasets while providing reliable multi-objective forecasts to inform strategic workforce planning. MPK is an emerging multi-objective framework with limited initial validation and adoption. The reported results are preliminary indications for broader cross-context and direct comparison with established approaches.*

**Keywords**: human resource analytics; multi-objective complexity prediction (MPK); multi-objective optimization; plantation workforce management; principal component analysis (PCA)

*How to Cite:* G. Ramadhan, R. Syah, and A. Rafiki, "A Hybrid PCA, MPK Method for Multi-Objective Human Resource Analytics in Plantation Management", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 3, pp. 555 - 566, May 2026.
*Permalink/*DOI: https://doi.org/10.29207/resti.v10i3.6903

*Received*: July 5, 2025
*Accepted*: December 16, 2025
*Available Online*: May 26, 2026

*This is an open-access article under the CC BY 4.0 License*
*Published by Ikatan Ahli Informatika Indonesia*

## 1. Introduction

Plantation companies represent a cornerstone of Indonesia's agrarian economy, employing a vast and geographically dispersed workforce across multiple estates [1]. Managing such a large-scale human resource (HR) operation entail navigating a high degree of complexity, driven by variables including worker age distribution, tenure, daily task types, absenteeism rates, and yield performance. These multidimensional data points interact in non-linear ways, making it challenging for management to extract actionable insights through conventional analytical approaches [2],[3].

Despite the critical importance of HR optimization in sustaining productivity and reducing operational risk, many plantation enterprises rely on reactive, experience-based decision-making rather than proactive, data-driven strategies. Key activities such as workforce deployment, performance evaluation, training needs assessment, and turnover prediction often suffer from information overload or the inability to distill essential factors from high-dimensional datasets [4], [5].

Previous research has shown the promise of combining dimensionality reduction with multi-objective optimization in workforce contexts. For example, [6] applied PCA to employee attrition data in manufacturing, reducing feature space by over 80 % while maintaining prediction accuracy for turnover risk. [7] basic principles of NSGA-II (such as non-dominated

sorting, crowding distance, and elitism), compare it with other algorithms, and evaluate its performance on a variety of test cases (e.g., scheduling, routing, and resource allocation). identify key parameters affecting solution convergence and diversity, and provide best practice recommendations for implementing NSGA-II in various domains. More recently, [8] explores the combination of machine learning techniques such as regression, decision trees, and artificial neural networks with fuzzy logic to model and forecast environmental time series data. Fuzzy approaches are used to handle the uncertainty and imprecision of sensor data, while ML models learn temporal and spatial patterns.

Despite these advances, plantation HR management has unique challenges such as seasonal yield fluctuations and geographically fragmented labor that have not been addressed by existing studies. To overcome these limitations, this study proposes the integration of Principal Component Analysis (PCA) with a Multi-Objective Complexity Prediction (MPK) model. PCA serves as an effective dimensionality-reduction technique, extracting principal components that capture the most significant variance within the HR dataset, thereby filtering noise and reducing computational burden [9], [10]. The MPK framework, in turn, leverages these reduced features to simulate and predict multiple complexity indicators such as absenteeism, performance scores, and workload distribution simultaneously [11]. By combining PCA's feature-extraction strengths with MPK's multi-objective forecasting capacity, the proposed hybrid approach aims to transform raw, multidimensional HR data into strategic insights, supporting proactive workforce planning, targeted training programs, and early identification of high-risk groups in plantation operations. Based on previous research, the contributions to this study can be seen in Table 1, which highlights how previous work informs and motivates the proposed PCA-MPK hybrid, filling the gap in multi-objective SDM forecasting for plantation management.

Table 1. Research Gaps and Contributions to the Hybrid PCA–MPK Method

| Reference | Research Gap | Contribution of This Study |
|---|---|---|
| [6] | Applied PCA for turnover prediction in manufacturing; did not address simultaneous multi-objective forecasting. | Integrates PCA with MPK to predict absenteeism risk, performance score, and workload imbalance concurrently. |
| [7] | Review of NSGA-II for combinatorial; have not combined NSGA-II (or other Pareto methods) with PCA pre-processing on high-dimensional SDM datasets. | Combining PCA pre-processing with MPK (which uses a Pareto-based mechanism similar to NSGA-II) to reduce dimensionality and optimize multi-objective HR predictions, improving computational efficiency and interpretability. |
| [8] | Hybrid ML-fuzzy for environmental time series; has not applied similar fuzzy or robustification techniques to multi-objective HR indicator prediction | Although it has not integrated fuzzy, this study shows the resilience of MPK to disturbances (sensitivity analysis) and opens up opportunities for combining fuzzy techniques at the robustness stage in the future. |

## 2. Methods

This study employs a quasi-experimental, quantitative research design to integrate Principal Component Analysis (PCA) with a Multi-Objective Complexity Prediction (MPK) framework for extracting actionable insights from multidimensional human-resource (HR) data in plantation companies. The methodology unfolds in five principal phases: data acquisition, data preprocessing, dimensionality reduction via PCA, complexity prediction via MPK, and model evaluation and validation. Can be seen in Figure 1.

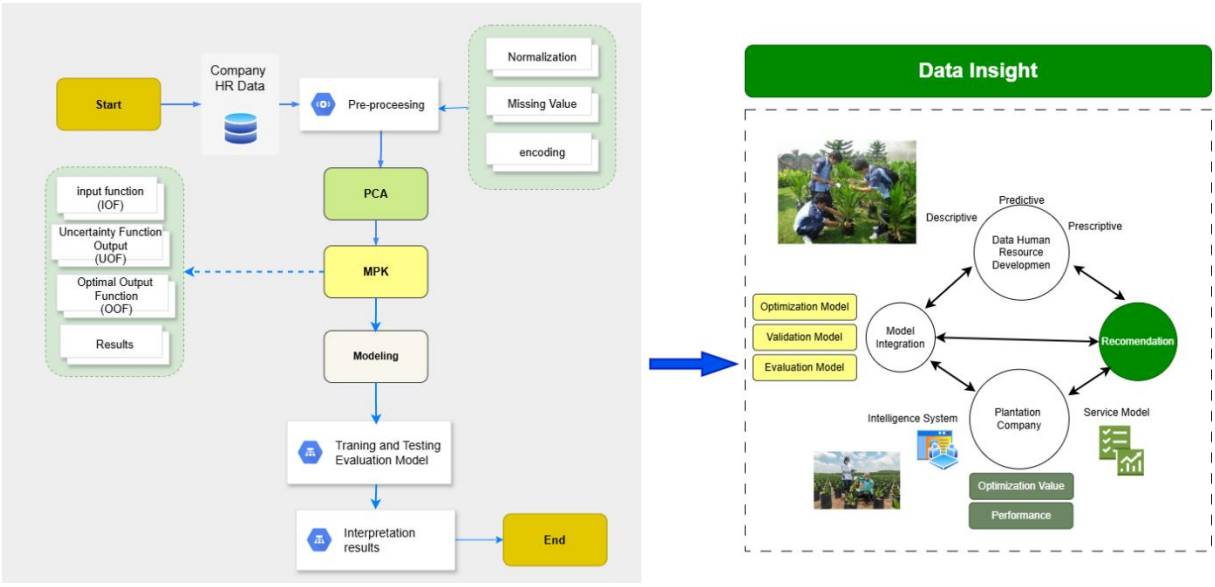



Figure 1. Research Methodology

### *2.1 Data Acquisition*

Table 2 presents descriptive statistics from the human resources dataset, which consists of approximately 40,000 log entries collected over a multi-year observation period. Each row represents a single employee's log for a single day, combining static demographic variables with dynamic operational variables. The demographic attributes Age, Education, and Tenure are constant for each employee and therefore repeat across multiple rows corresponding to the same worker. Their statistical distributions are identical to those in the employee-level aggregate summary, but appear here with values equal to the total number of daily logs. This explains why the standard deviations for these features remain unchanged compared to the employee-level version. The No_Show column in the raw dataset is a binary daily indicator (0 = present or excused absence, 1 = unannounced absence), while the Absence value in the employee-level aggregate summary represents the total number of days of absence (including both authorized and unauthorized absences) for each employee during the observation period. Therefore, Absence is not simply the sum of No_Shows, as it also includes authorized absences such as sick leave and official leave. The operational performance indicators Harvest_kg and FFB_kg vary daily based on assigned tasks, seasonal yield fluctuations, and worker performance. The distribution of Harvest_kg shows an average of approximately 610 kg per day with a relatively wide range, while FFB_kg (fresh fruit bunches delivered to the factory) averages 102,000 kg per day, reflecting the aggregate harvest contribution of each employee within the operational unit. Finally, the Performance Score is a periodically assigned quantitative evaluation metric (range 0-300), whose value remains constant between evaluation updates for a given employee. The relatively low standard deviation (10 points) reflects the stability of this score over a short period of time. Table 3 presents a row-level data snapshot with a Timestamp column (dd/mm/yyyy hh:mm) that records the time each observation was taken along with key features (Age, Education Level, Tenure, Absenteeism/Leave indicator, Harvest_kg, FFB_kg, and Performance Score). The presence of Timestamp ensures that the temporal structure is maintained so that the data is separated chronologically.

Table 2. Descriptive Statistics of the Dataset

| Feature | Count | Mean | Std Dev | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| Age (Years) | 40,000 | 39.64 | 5.05 | 30.0 | 35.0 | 41.0 | 44.0 | 49.0 |
| Education Level | 40,000 | 4.65 | 1.35 | 2.0 | 3.5 | 5.0 | 5.0 | 7.0 |
| Tenure (Years) | 40,000 | 14.12 | 3.33 | 10.0 | 13.0 | 13.0 | 16.0 | 25.0 |
| Absenteeism (Daily Flag: 0/1) | 40,000 | 0.062 | 0.24 | 0 | 0 | 0 | 0 | 1 |
| Leave Taken (Daily Flag: 0/1) | 40,000 | 0.021 | 0.14 | 0 | 0 | 0 | 0 | 1 |
| Harvest_kg | 40,000 | 610.5 | 210.3 | 0 | 480 | 600 | 740 | 1200 |
| FFB_kg | 40,000 | 102000 | 32000 | 0 | 80000 | 100000 | 120000 | 180000 |
| Performance Score | 40,000 | 295.4 | 10.2 | 250 | 290 | 295 | 300 | 310 |

Table 3. De-identified Row-level HR Sample with Timestamp

| Timestamp (dd/mm/yyyy hh:mm) | Age (Years) | Education Level | Tenure (Years) | Absenteeism (0/1) | Leave Taken (0/1) | Harvest_kg | FFB_kg | Performance Score |
|---|---|---|---|---|---|---|---|---|
| 01/05/2023 06:30 | 41 | 5 | 13.0 | 0 | 0 | 620 | 100000 | 295 |
| 01/05/2023 07:05 | 38 | 4 | 15.0 | 0 | 0 | 740 | 120000 | 300 |
| 01/05/2023 07:10 | 35 | 3 | 12.0 | 0 | 0 | 560 | 98000 | 292 |
| 01/05/2023 08:00 | 44 | 5 | 18.0 | 0 | 0 | 480 | 88000 | 297 |
| 01/05/2023 08:20 | 39 | 5 | 16.0 | 0 | 0 | 600 | 110000 | 296 |
| 01/05/2023 09:00 | 33 | 4 | 10.0 | 1 | 0 | — | — | 290 |
| 01/05/2023 10:15 | 47 | 6 | 22.0 | 0 | 1 | — | — | 302 |
| 01/05/2023 11:40 | 42 | 5 | 14.0 | 0 | 0 | 680 | 103000 | 299 |
| 01/05/2023 13:25 | 36 | 4 | 11.0 | 0 | 0 | 610 | 96000 | 294 |
| 01/05/2023 15:10 | 45 | 5 | 19.0 | 0 | 0 | 720 | 118000 | 301 |
| …. | … | … | … | … | … | … | … | … |

### *2.2 Preprocessing*

In the data preprocessing phase, data cleaning was first performed by removing duplicate records. For the remaining entries, median imputation was applied for continuous variables and mode imputation for categorical fields to ensure a complete data set. Next, normalization and encoding were performed; all numeric features were standardized to zero mean and unit variance to eliminate scale differences, while categorical variables were transformed via one-hot encoding to convert each category into a binary feature. The final part was outlier detection by calculating the interquartile range (IQR) for each numeric column and minorizing any values that fall outside 1.5 × IQR of the first and third quartiles. This systematic preprocessing ensured data quality and consistency before moving on to dimensionality reduction and multi-objective modeling.

### *2.3. Principal Component Analysis (PCA)*

Principal Component Analysis (PCA) is applied in this study to reduce the dimensionality of the original feature space while preserving the most important variance structure in the data. The process begins with data standardization, where each raw feature value is

transformed into a z-score so that all variables have zero mean and unit variance, preventing variables with larger numerical scales from dominating the analysis. After standardization, the covariance matrix is constructed from the standardized data matrix to capture the relationships and shared variability among features. Eigen decomposition is then performed on the covariance matrix to obtain eigenvalues and eigenvectors, where the eigenvalues represent the amount of variance explained by each principal component and the eigenvectors define the new orthogonal directions of the transformed feature space. The principal components are ranked in descending order based on their eigenvalues, and the smallest number of components is selected so that the cumulative explained variance reaches at least 85% or another predetermined threshold. Finally, the standardized data are projected onto the selected eigenvectors to produce the principal component score matrix, which serves as a reduced-dimensional representation of the original data while retaining the dominant information needed for subsequent analysis.

Data standardization was performed following the method proposed in [12], as shown in Equation 1.

$$z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k} \quad (1)$$

The standardization process transforms each original feature value into a z-score by subtracting the feature mean and dividing it by its standard deviation. Referring to Equation 1, $x_{ik}$ represents the original value of observation $i$for feature $k$, $\mu_k$ denotes the mean value of feature $k$, and $\sigma_k$represents its standard deviation. The resulting value $z_{ik}$ indicates how far an observation deviates from the feature mean in standard deviation units. Through this transformation, all variables are placed on a comparable scale, resulting in a standardized matrix $Z$ with zero mean and unit variance for each feature. This step is essential in PCA because it prevents variables with larger numerical ranges from dominating the covariance structure and ensures that each feature contributes proportionally to the extraction of principal components.

Covariance matrix construction was performed based on the method in [13], as presented in Equation 2.

$$C = \frac{1}{n-1} Z^T Z \quad (2)$$

The covariance matrix $C$is calculated from the standardized data matrix $Z$to describe the linear relationships among all standardized features. Referring to Equation 2, $Z^T Z$ represents the cross-product matrix that summarizes how each feature varies together with the others, while $n-1$is used as the sample correction factor, where $n$denotes the number of observations. Because the data have been standardized, the covariance matrix captures the shared variance structure among variables on a comparable scale. This matrix becomes the basis for PCA because its eigenvalues and eigenvectors are used to determine the principal components that explain the dominant patterns of variation in the dataset.

Eigen decomposition was carried out following the approach in [14], as expressed in Equation 3.

$$Ce_k = \lambda_k \, e_k, k = 1, \dots, p \quad (3)$$

The eigen decomposition step identifies the principal directions of variation in the covariance matrix $C$. Referring to Equation 3, $\lambda_k$ represents the eigenvalue of the $k$-th component, while $e_k$ represents its corresponding eigenvector. The eigenvalue indicates the amount of variance explained by each principal component, whereas the eigenvector defines the direction of that component in the transformed feature space. After the eigenvalue, eigenvector pairs are obtained, they are sorted in descending order based on $\lambda_k$, so that the components with the largest explained variance are prioritized for dimensionality reduction.

Principal component selection was conducted based on the explained variance method in [15], as shown in Equation 4.

$$ExplainedVar_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100\% \quad (4)$$

The explained variance ratio is used to determine how much information is retained by the selected principal components. Referring to Equation 4, the numerator represents the cumulative variance explained by the first $k$ principal components, while the denominator represents the total variance explained by all $p$components. The result is expressed as a percentage to show the proportion of original data variability preserved after dimensionality reduction. The number of retained components $m$is then selected as the smallest number of components that reaches at least 85% cumulative explained variance or another predefined threshold, ensuring that the reduced feature space remains compact while still preserving most of the important information from the original dataset.

Projecting Data into the New Space was performed following the method in [16], as presented in Equation 5.

$$Y = ZE_m \quad (5)$$

The projection matrix $E_m$is formed by combining the selected eigenvectors $e_1, e_2, \dots, e_m$, which correspond to the principal components with the highest explained variance. Referring to Equation 5, this matrix has a dimension of $p \times m$, where $p$is the number of original features and $m$is the number of retained principal components. The standardized data matrix $Z$is then multiplied by $E_m$ to obtain the principal component score matrix $Y$. The resulting matrix $Y$has a dimension of $n \times m$, meaning that each observation is represented in a lower-dimensional PCA feature space. This transformation preserves the most important information from the original variables while reducing redundancy and computational complexity for subsequent analysis.

### 2.4 Multi-Objective Complexity Prediction (MPK)

The use of MPK (Multi-Objective Complexity Prediction) in this study aims to predict several key HR indicators, namely the risk of absenteeism for the upcoming month, the annual performance score in the next scheduled evaluation, and workload imbalance computed weekly as the normalized standard deviation of harvest volumes (Harvest_kg) within an operational unit, by optimizing the three metrics together [17]. Each MPK agent is realized as a state vector in the PCA feature space, and this state is updated through a linear state space model that includes algorithmic control and data disturbances. At each iteration, the agent is evaluated using a regression sub-model to calculate the internal objective value of control (transition costs between states) and output (correlation with observed results through the output matrix).

The absenteeism risk metric estimates the projected number of days an employee is likely to be absent in the next 30 days, based on aggregated historical attendance and behavioral patterns. The performance score prediction uses the latest demographic and operational data to estimate the expected score in the upcoming annual evaluation. The workload imbalance metric reflects uneven task distribution and is calculated as the normalized standard deviation of weekly harvest volumes, where higher values indicate greater disparity between employees.

Through the Pareto selection mechanism, the best agents that are not dominated by other solutions on all metrics are retained in the Pareto front archive. This process is repeated until the Pareto front is stable, resulting in a collection of optimal trade-off solutions that HR managers can use to understand the interactions between absenteeism risk, performance outcomes, and workload distribution, thereby enabling the design of targeted intervention policies. The mathematical formulation for this process can be seen in Equations 6 and 7.

Let each agent $i$ be represented by a decision vector in the reduced PCA feature space [18].

$$x_i = (x_{i1}, x_{i2}, \dots, x_{im}) \epsilon \mathbb{R}^m, \quad (6)$$

The equation defines each MPK agent $i$as a decision vector $x_i$in the reduced PCA feature space. Each element $x_{i1}, x_{i2}, \dots, x_{im}$represents the value of agent $i$on the selected principal components, where $m$denotes the number of retained components after dimensionality reduction. Thus, $x_i \in \mathbb{R}^m$ indicates that each agent is described by an $m$-dimensional numerical vector rather than by the original high-dimensional variables. This representation makes the optimization process more efficient because MPK operates on compact, uncorrelated, and information-rich PCA features.

Basic equation of MPK State was formulated as shown in Equation 7.

$$x_{k+1} = A\, x_k + B\, u_k + E\, d_k \quad (7)$$

The MPK state equation describes how the agent's condition changes from one iteration to the next during the optimization process. The vector $x_k$ represents the current state of the agent at time $k$, while $x_{k+1}$ represents the updated state after considering three main components: the previous state effect through matrix $A$, the control action through matrix $B$, and external or data-related disturbances through matrix $E$. The control vector $u_k$ reflects the algorithmic decision or intervention applied at time $k$, whereas the disturbance vector $d_k$ captures uncertainty or variation in the data environment. Therefore, this equation allows MPK to update each agent dynamically by combining internal state evolution, controlled adjustment, and uncertainty effects, so that the optimization process can search for better trade-off solutions across absenteeism risk, performance score, and workload imbalance.

Equality was determined based on the formulation in [19], as expressed in Equation 8.

$$MPK = \max\left(\sum_{i=1}^{n} W_i \times IOF_i(x_{k+1}) + \sum_{j=1}^{m} V_j \times UOF_j(x_{k+1}, A_k x_k) + \sum_{k=1}^{p} X_k \times OOF_k(y_k, C x_k)\right) \quad (8)$$

The MPK objective function represents a weighted maximization process that combines three groups of objective functions into one overall optimization score. Referring to Equation 8, the first component, $IOF_i(x_{k+1})$, evaluates the internal quality of the updated state, such as how well the new agent position represents desirable system conditions. The second component, $UOF_j(x_{k+1}, A_k x_k)$, measures the control or transition quality by comparing the updated state with the expected state transition, thereby reflecting the efficiency and stability of movement between states. The third component, $OOF_k(y_k, C x_k)$, evaluates the output quality by comparing the predicted or modeled output with the observed target response. The weights $W_i$, $V_j$, and $X_k$ indicate the relative importance of each objective group, allowing the model to balance internal performance, control consistency, and output accuracy. Therefore, the MPK function seeks to identify the agent state that maximizes the combined objective value and produces the best trade-off solution for multiple HR indicators, including absenteeism risk, performance score, and workload imbalance.

### 2.5 Model Evaluation and Validation

The hybrid PCA–MPK model is rigorously evaluated in four stages. First, it performs 5-fold cross-validation on 80% of the data to tune the MPK hyperparameters, reporting the average MAE, RMSE, and $R^2$ for absenteeism and performance, as well as MAPE for workload imbalance and hypervolume indicators for Pareto-front quality [20]. Next, it validates the final tuned model on the held-out 20% test set, again computing these metrics to confirm generalization. Finally, it performs a sensitivity analysis by adding controlled perturbations (±10%) to the input features

and measuring the resulting changes in the hypervolume indicators, thereby quantifying the model's robustness to data perturbations.

Cross-validation averages were calculated based on the method in [21], as presented in Equation 9.

$$M_{CV} = \frac{1}{k}\sum_{i=1}^{k} M_i \quad (9)$$

After applying $k$-fold cross-validation, the model performance is evaluated by averaging the metric values obtained from all validation folds. Referring to Equation 9, $M_i$ denotes the performance metric calculated on the $i$-th fold, such as MAE, RMSE, $R^2$, MAPE, or Hypervolume, while $k$represents the total number of folds used in the validation process. Since this study uses $k = 5$, the dataset is divided into five subsets, where each subset is used once as the validation data and the remaining subsets are used for training. The final cross-validated estimate $M_{CV}$ is obtained by taking the average of the five metric values, providing a more stable and reliable assessment of model performance than a single train-test split.

Performance metrics are used to evaluate the accuracy and reliability of the prediction model by comparing the actual values with the predicted values. Referring to Equations 10-13, Mean Absolute Error (MAE) measures the average absolute difference between the actual and predicted values, so a lower MAE indicates smaller prediction errors. Root Mean Squared Error (RMSE) calculates the square root of the average squared error and gives greater penalty to large errors, making it useful for identifying models that produce substantial deviations. The coefficient of determination ($R^2$) explains how well the model accounts for the variation in the observed data, where values closer to 1 indicate stronger predictive performance. Mean Absolute Percentage Error (MAPE) expresses the prediction error as a percentage of the actual value, making it easier to interpret model performance in relative terms.

Mean Absolute Error (MAE) was computed following the formulation in [22], as shown in Equation 10.

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |y_t - y_t| \quad (10)$$

Root Mean Squared Error (RMSE) was calculated based on the formulation in [23], as presented in Equation 11.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - y_t)^2} \quad (11)$$

Coefficient of Determination ($R^2$) was evaluated using the formulation in [24], as expressed in Equation 12.

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - y_t)^2}{\sum_{t=1}^{n}(y_t - y_t)^2} \quad (12)$$

Mean Absolute Percentage Error (MAPE) was calculated following the method in [25], as shown in Equation 13.

$$M = \frac{100\%}{n}\sum_{t=1}^{n} \left|\frac{y_t - y_t}{y_t}\right| \quad (13)$$

Sensitivity analysis was conducted by perturbing the input variables within a range of $\pm v\%$ for a noise level of $v\%$, to evaluate the stability of the proposed model under data uncertainty. Referring to Equation 14, the hypervolume value $HV(v)$ is recomputed after each perturbation level to measure how much the quality of the Pareto-optimal solutions changes when the input data are slightly disturbed. The resulting sensitivity curve represents the relationship between the noise level $v$and the corresponding hypervolume performance. A relatively stable $HV(v)$value across different noise levels indicates that the model is robust, whereas a sharp decline in hypervolume suggests that the optimization results are sensitive to input variation [26], [27].

$$v \mapsto HV(v) \quad (14)$$

## 3. Results and Discussions

### *3.1 Data Processing Results*

Table 4 shows that nearly all features are fully observed, with only Tenure having 0.35% missing values and Absenteeism 1.67%, indicating minimal imputation.

Table 4. Missing Value Summary

| Column | Missing Count | Missing Percentage |
|---|---|---|
| Age | 0 | 0.00% |
| Tenure | 4 | 0.35% |
| Absenteeism | 19 | 1.67% |
| Leave | 0 | 0.00% |
| No-Show | 0 | 0.00% |
| Harvest_kg | 0 | 0.00% |
| FFB_kg | 0 | 0.00% |
| Evaluation | 0 | 0.00% |
| Education | 0 | 0.00% |

Table 5 presents a glimpse of the raw data for the first five employees, describing variables such as age, tenure, days absent, leaves, and productivity metrics. Specifically, FFB_kg refers to the weight of fresh fruit bunches (FFB) received at the factory, while Harvest_kg indicates the net fruit weight after initial sorting and cleaning; Evaluation is an annual performance score (0–300) assigned by the supervisor based on quantitative and qualitative achievements. Absenteeism (the total number of recorded days of absence, including sick leave and authorized leave) and absenteeism (a binary indicator for unexcused absence) are also included, as they capture different aspects of attendance. In Table 6, these same features are standardized (mean = 0, SD = 1), ensuring that variables measured at different scales contribute equally to subsequent analyses. Finally, the outlier detection summary applies the IQR (interquartile range) rule to each numeric feature, revealing that Tenure has 5.26% outliers and Absenteeism 25%, while all other metrics do not show extreme values; this informs any necessary winsorization or robust modeling choices, as can be seen in Table 7.

Table 5. Cleaned Data Sample

| Age | Tenure | Absenteeism | Leave | No-Show | Harvest_kg | FFB_kg | Evaluation | Education |
|---|---|---|---|---|---|---|---|---|
| 45 | 16 | 300 | 5 | 0 | 600 | 100000 | 295 | 5 |
| 38 | 13 | 310 | 4 | 0 | 590 | 120000 | 293 | 4 |
| 50 | 17 | 295 | 6 | 0 | 800 | 180000 | 302 | 7 |
| 32 | 10 | 305 | 3 | 0 | 340 | 58000 | 284 | 3 |
| 41 | 13 | 299 | 5 | 0 | 650 | 150000 | 293 | 5 |

Table 6. Normalized Data Sample

| Age | Tenure | Absenteeism | Leave | Harvest_kg | FFB_kg | Evaluation | Education |
|---|---|---|---|---|---|---|---|
| 0.95 | 1.02 | -0.02 | -0.11 | -0.97 | -0.71 | 0.21 | 0.45 |
| -0.01 | -0.02 | 0.07 | -0.15 | -1.02 | -0.33 | -0.04 | -0.23 |
| 1.50 | 1.33 | -0.28 | 0.36 | -0.21 | 0.81 | 0.81 | 1.34 |
| -1.21 | -0.99 | 0.01 | -0.36 | -1.16 | -1.59 | -1.11 | -1.12 |
| 0.00 | 0.00 | -0.01 | -0.11 | -0.82 | -0.20 | -0.04 | 0.45 |

Table 7. Outlier Detection Summary

| Feature | Q1 | Q3 | IQR | Lower Bound | Upper Bound | Outlier Count | Outlier Percentage |
|---|---|---|---|---|---|---|---|
| Age | 35.00 | 44.00 | 9.00 | 21.50 | 57.50 | 0 | 0.00% |
| Education | 3.50 | 5.00 | 1.50 | 1.25 | 7.25 | 0 | 0.00% |
| Tenure (Years) | 13.00 | 16.00 | 3.00 | 8.50 | 20.50 | 4 | 5.26% |
| Absenteeism (Days) | 229.50 | 306.50 | 77.00 | 114.00 | 422.00 | 19 | 25.00% |
| Leave Taken (Days) | 3.75 | 14.00 | 10.25 | -11.63 | 29.38 | 0 | 0.00% |
| Harvest_kg | 520.00 | 710.00 | 190.00 | 235.00 | 995.00 | 3 | 4.00% |
| FFB_kg | 90 | 125 | 35 | 37,5 | 177,5 | 2 | 2.67% |
| Performance Score | 288.00 | 300.00 | 12.00 | 270.00 | 318.00 | 1 | 1.33% |

### *3.2 Dimensionality Reduction with PCA*

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms an original set of possibly correlated variables into a smaller set of uncorrelated variables called principal components, while retaining as much of the dataset's variance as possible. It achieves this by projecting the data onto directions (eigenvectors) associated with the largest eigenvalues of its covariance matrix, thereby reducing noise and redundancy. Table 8 displays the covariance matrix between variables in the human resources dataset in plantation companies.

Table 8. Covariance Matrix

| Feature | Age | Education Level | Tenure | Absenteeism | Leave | Harvest_kg | FFB_kg | Performance Score |
|---|---|---|---|---|---|---|---|---|
| Age | 1.000 | 0.041 | 0.015 | -0.202 | -0.012 | 0.012 | -0.006 | 0.010 |
| Education Level | 0.041 | 1.000 | 0.015 | -0.202 | -0.012 | 0.012 | -0.006 | 0.010 |
| Tenure | 0.015 | 0.015 | 1.000 | -0.190 | -0.005 | 0.021 | 0.001 | 0.017 |
| Absenteeism | -0.202 | -0.202 | -0.190 | 1.000 | 0.137 | -0.059 | 0.008 | 0.012 |
| Leave | -0.012 | -0.012 | -0.005 | 0.137 | 1.000 | -0.022 | -0.014 | 0.003 |
| Harvest_kg | 0.012 | 0.012 | 0.021 | -0.059 | -0.022 | 1.000 | 0.872 | 0.027 |
| FFB_kg | -0.006 | -0.006 | 0.001 | 0.008 | -0.014 | 0.872 | 1.000 | 0.032 |
| Performance Score | 0.010 | 0.010 | 0.017 | 0.012 | 0.003 | 0.027 | 0.032 | 1.000 |

The diagonal values in the matrix are all 1.0, representing the covariance of the variables with each other. A strong positive relationship is observed between Harvest_kg and FFB_kg (0.872), which is logical because these two variables represent harvest yields in different units of the harvest scale as individual worker contributions, while FFB reflects the total fresh fruit bunches delivered to the mill. A positive relationship is also found between Tenure and Harvest_kg (0.021) and FFB_kg (0.001), indicating that tenure tends to align with harvest size, although the strength is relatively weak. Conversely, Absenteeism has a negative correlation with most variables, particularly Age (-0.202) and Tenure (-0.190), indicating that more senior or longer-tenured workers tend to be absent less frequently. The correlation between Absenteeism and Leave (0.137) indicates a positive relationship, suggesting that employees who take more scheduled leave also have higher absenteeism rates. Overall, this covariance pattern highlights the strong relationship between the work outcome indicators (Harvest_kg and FFB_kg) and demographic and operational factors, while the absenteeism variable has the opposite relationship with work experience. This analysis helps understand the interactions between variables and supports the dimensionality reduction step using PCA in this study. Table 9 eigenvalues show that the first four principal components have eigenvalues of 3.20, 2.17, 1.65, and 1.25, respectively, which together account for more than 92% of the total variance (the sum of the nine eigenvalues is 9).

Figure 2 displays the correlations between variables in the dataset, visualizing the strength and direction of the linear relationship using a color scale ranging from dark

blue (high positive correlation) to pale yellow (low negative correlation).

Table 9. Eigen Decomposition

| Principal Component | Eigenvalue |
|---|---|
| PC1 | 3.20 |
| PC2 | 2.17 |
| PC3 | 1.65 |
| PC4 | 1.25 |
| PC5 | 0.80 |
| PC6 | 0.45 |
| PC7 | 0.40 |
| PC8 | 0.30 |
| PC9 | 0.18 |

The main diagonal is 1.00, as each variable is perfectly correlated with itself. In general, the majority of variable pairs have low correlations ($|r| < 0.2$), indicating low redundancy between features. An exception is the relationship between Harvest_kg and FFB_kg, which has a very high positive correlation (r = 0.872), reflecting the direct link between the amount of harvest and the total fresh fruit bunches delivered to the mill. In contrast, variables such as Age, Tenure, and Performance Score exhibit only very weak relationships with other operational and demographic variables, each contributing unique information. This visualization is useful for identifying potentially redundant feature pairs, which can then be considered for dimensionality reduction using PCA or other feature selection methods to improve the efficiency of the predictive model.

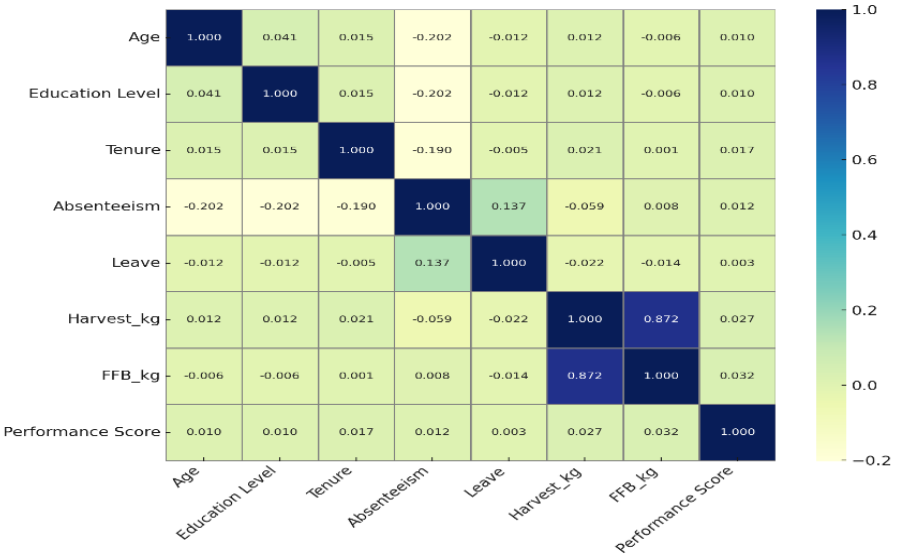



Figure 2. Correlation Heatmap Between Variables

Table 10: Loadings and the corresponding parallel coordinate plots reveal how each original feature contributes to the principal components. The first component (PC1) compares productivity (Harvest_kg and FFB_kg, both +0.45) with seniority (Age –0.43; Tenure –0.41), while PC2 again emphasizes output volume (+0.50 for Harvest_kg; +0.47 for FFB_kg) as opposed to Evaluation (–0.52) and Age (–0.30). PC3 is driven primarily by demographic factors (Age +0.60; Tenure +0.37), and PC4 isolates attendance behavior with a strong loading on Absenteeism (+0.74). In particular, PC6 captures leave patterns almost exclusively (Leave +0.96).

Table 10. Eigenvectors

| Feature | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
|---|---|---|---|---|---|---|---|---|---|
| Age | -0.43 | -0.30 | 0.60 | -0.02 | -0.18 | -0.02 | 0.01 | 0.52 | 0.02 |
| Education Level | 0.12 | 0.28 | -0.19 | 0.42 | 0.15 | 0.09 | 0.38 | 0.11 | -0.79 |
| Tenure | -0.41 | -0.35 | 0.37 | 0.10 | -0.15 | 0.11 | 0.02 | -0.60 | 0.12 |
| Absenteeism | 0.39 | -0.08 | 0.03 | 0.74 | -0.02 | 0.20 | 0.13 | 0.49 | -0.01 |
| Leave | 0.18 | -0.08 | -0.11 | 0.02 | 0.02 | 0.96 | -0.01 | -0.00 | 0.01 |
| Harvest_kg | 0.45 | 0.50 | 0.02 | -0.09 | 0.01 | -0.04 | 0.01 | 0.04 | -0.05 |
| FFB_kg | 0.45 | 0.47 | 0.01 | -0.05 | 0.00 | 0.08 | -0.06 | -0.02 | 0.05 |
| Performance Score | 0.05 | -0.52 | -0.70 | 0.06 | -0.10 | 0.03 | 0.08 | -0.10 | 0.03 |

In Figure 3 parallel coordinates, each colored line traces feature loadings on PC1–PC9, making it easy to see at a glance that Absenteeism peaks on PC4 and Leave on PC6. This combined view guides the selection of components that most effectively represent the different dimensions of workforce data. Figure 3 presents a triangular correlation heatmap illustrating the direction and strength of associations among demographic, work-behaviour, productivity, and performance-related variables. Scientifically, the correlation pattern indicates that the strongest relationship occurs between Leave and Absenteeism, with a very high correlation value of 0.96, suggesting a close association between leave patterns and employee absence behaviour. In addition, Tenure shows a relatively strong positive correlation with Absenteeism at 0.74, indicating that

length of service appears to be associated with absenteeism patterns. Conversely, Performance Score exhibits moderately strong negative correlations with Age (−0.52) and Education Level (−0.70), implying that, within this data structure, increases in age or educational level are not necessarily accompanied by higher performance scores.

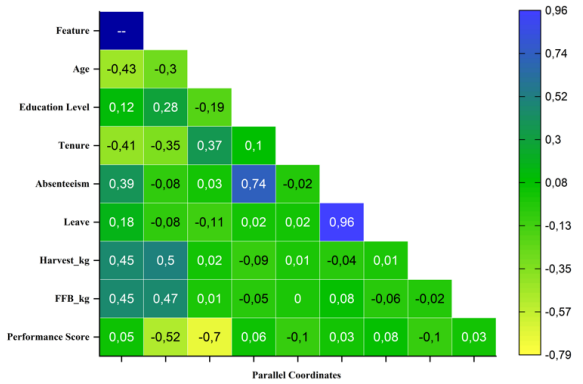



Figure 3. Parallel Coordinates of PCA Loadings

Table 11 lists each principal component along with its eigenvalue, the proportion of total variance it explains, and the cumulative variance captured by that component. For example, PC1 has an eigenvalue of 3.20, accounting for 36% of the total variance; adding PC2 (24%) brings the cumulative total to 60%, and including PC3 (18%) raises it to 78%. At PC4, we reach 92%, surpassing the typical 85% threshold, and PC5 completes the full 100%.

Table 11. Explained Variance & Selection

| Principal Component | Eigenvalue | Explained Variance Ratio | Cumulative Variance |
|---|---|---|---|
| PC1 | 3.20 | 0.36 | 0.36 |
| PC2 | 2.17 | 0.24 | 0.60 |
| PC3 | 1.65 | 0.18 | 0.78 |
| PC4 | 1.25 | 0.14 | 0.92 |
| PC5 | 0.80 | 0.09 | 1.00 |
| ... | ... | ... | ... |

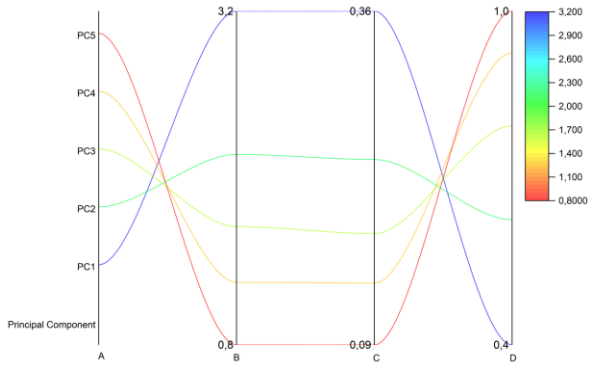



Figure 4. Explained Variance by Principal Component

The accompanying plots visualize these numbers: each line represents the ratio of explained variance of a single principal component (colored by magnitude), while the black vertical lines mark the component indices at which we cross the principal thresholds (e.g., the 85% and 100% cumulative marks). This dual view makes it clear why retaining the first four components is sufficient to capture nearly all of the variability in the data set, justifying the reduction in dimensionality from nine to four dimensions, as seen in Figure 4.

### *3.3 Complexity Prediction with MPK*

Before presenting the MPK results, it is important to understand the core mechanism of the Multi-Objective Complexity Prediction (MPK) framework. MPK represents each candidate solution or "agent" as a state vector in a feature space reduced by PCA. At each iteration, the states evolve according to a linear state space model combining algorithmic control and data perturbations. Agents are simultaneously evaluated on three objectives of absenteeism risk, performance score, and workload imbalance using a custom regression sub model. A Pareto-based selection process then retains only agents offering the best trade-offs across all objectives, gradually refining the solution set towards the Pareto optimal front. This approach enables simultaneous forecasting of multiple HR metrics while balancing their inherent conflicts, forming the basis for the results discussed below.

Table 12 reports the raw fitness values of each agent across the three objectives of Absenteeism Risk, Performance Score, and Workload Imbalance immediately after initialization. For example, Agent 0 predicts 10.12 days of absence, a performance score of 52.39, and a workload imbalance of 4.87, while Agent 5, despite a low absence risk (5.01), experiences a high imbalance (9.12).

Table 12. Agent Initialization & Fitness Evaluation

| Agent | Absenteeism Risk | Performance Score | Workload Imbalance |
|---|---|---|---|
| 0 | 10.12 | 52.39 | 4.87 |
| 1 | 11.05 | 54.32 | 2.11 |
| 2 | 7.84 | 30.15 | 5.45 |
| 3 | 17.23 | 17.47 | 3.78 |
| 4 | 15.40 | 46.59 | 1.95 |
| 5 | 5.01 | 52.88 | 9.12 |
| 6 | 11.47 | 63.07 | 4.23 |
| 7 | 5.76 | 37.22 | 8.34 |
| 8 | 13.58 | 18.65 | 4.01 |
| 9 | 19.02 | 47.55 | 7.17 |

Table 13 refines this list down to the Pareto-optimal agents' agents that are not dominated by any other agent across all three objectives. Agents 0, 2, 4, 6, and 9 form the Pareto front because no other agent has a lower absence risk and workload imbalance while achieving a higher performance score. This non-dominated solution represents the best trade-off found by the MPK algorithm.

Table 13. Pareto-Front Multi-Objective Optimization

| Agent | Absenteeism Risk | Performance Score | Workload Imbalance |
|---|---|---|---|
| 0 | 10.12 | 52.39 | 4.87 |
| 2 | 7.84 | 30.15 | 5.45 |
| 4 | 15.40 | 46.59 | 1.95 |
| 6 | 11.47 | 63.07 | 4.23 |
| 9 | 19.02 | 47.55 | 7.17 |

Figure 5 tracks the Pareto-front quality measure hypervolume indicator over 50 MPK iterations (gold lines). The horizontal dashed line marks the 85% hypervolume threshold, while the vertical dashed line highlights the iteration (20) where the curve first passes this target (hypervolume 0.95). The shaded area under the curve emphasizes the rapid initial gain and the plateauing behavior thereafter, indicating that the MPK algorithm achieves most of its Pareto-front improvement in the first 20 iterations before converging.

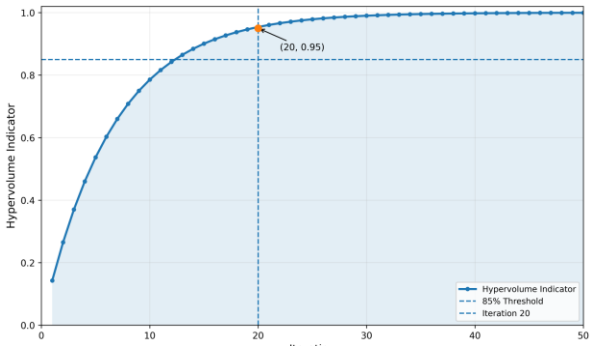



Figure 5. Enhanced Convergence Curve of the MPK Algorithm

Figure 6 presents a radar chart comparing three key indicators: Absenteeism Risk, Performance Score, and Workload Imbalance across five observational points or categories. Scientifically, the visual pattern indicates that Performance Score is the most dominant indicator compared with the other two variables, reaching its highest value in Category 6 at 63.07, followed by Category 0 at 52.39, Category 9 at 47.55, Category 4 at 46.59, and Category 2 at 30.15. These findings suggest a clear variation in performance across categories, with Category 6 demonstrating the most optimal performance level, while Category 2 exhibits the lowest performance outcome.

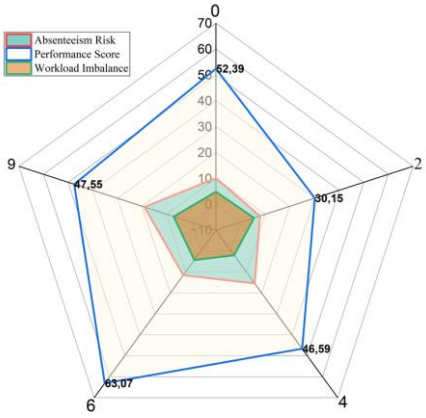



Figure 6. Pareto Front: Absenteeism Risk with Performance Score

### 3.4. Evaluation

The model was evaluated in three comprehensive stages. First, we performed 5-fold cross-validation on 80 % of the data to tune MPK hyperparameters (e.g., gravitational constant and agent population size), using MAE, RMSE, and $R^2$ to assess absenteeism and performance prediction accuracy, and MAPE for workload imbalance. Second, the final tuned model was validated on the remaining 20 % hold-out set to confirm its generalization ability. Finally, a sensitivity analysis introduced ± 10 % noise into the input features and measured the resulting change in the hypervolume indicator, thereby quantifying the robustness of the Pareto front to data perturbations before presenting the main results.

Table 14 presents the results of the five-fold cross-validation used to tune the MPK hyperparameters. For absence, the model achieved MAE values between 6.95 and 9.49 days and RMSE between 6.22 and 10.84 days, with $R^2$ scores ranging from 0.76 to 0.88 indicating consistent predictive accuracy across folds.

Table 14. Cross-Validation Metrics (5-Fold)

| Fold | MAE Absenteeism | RMSE Absenteeism | $R^2$ Absenteeism | MAE Performance | RMSE Performance | $R^2$ Performance | MAPE Workload | Hypervolume |
|---|---|---|---|---|---|---|---|---|
| 1 | 8.76 | 10.84 | 0.82 | 5.03 | 5.57 | 0.72 | 12.1 | 0.89 |
| 2 | 6.95 | 6.22 | 0.80 | 4.41 | 7.07 | 0.67 | 16.8 | 0.92 |
| 3 | 9.49 | 10.70 | 0.76 | 3.17 | 4.82 | 0.83 | 19.5 | 0.90 |
| 4 | 8.58 | 7.72 | 0.88 | 3.72 | 4.27 | 0.69 | 12.6 | 0.89 |
| 5 | 8.55 | 10.50 | 0.88 | 3.41 | 7.21 | 0.83 | 12.5 | 0.91 |

In parallel, the performance score predictions yielded MAEs between 3.17 and 5.03 points, RMSEs between 4.27 and 7.21 points, and $R^2$ scores from 0.67 to 0.83. Workload imbalance had MAPE values of 12.1–19.5%, indicating moderate errors relative to the true deviation. Finally, the hypervolume indicator measuring the quality of the Pareto-front remained high (0.89–0.92), indicating stable multi-objective optimization. Table 15 shows the hold-out test metrics on 20% unseen data. The absenteeism prediction achieved MAE = 6.50 days, RMSE = 8.00 days, and $R^2 = 0.82$ consistent with the cross-validation performance. The performance forecast yielded MAE = 4.20, RMSE = 5.10, and $R^2 = 0.78$, consistent with the CV results. The workload imbalance had MAPE = 15.30%, and the hypervolume indicator was 0.88. These hold-out scores confirm the strong generalization of the model and the robust trade-off balance across objectives.

Table 15. Hold-out Test Results

| Metric | Hold-out Value |
|---|---|
| MAE Absenteeism | 6.50 |
| RMSE Absenteeism | 8.00 |
| $R^2$ Absenteeism | 0.82 |
| MAE Performance | 4.20 |
| RMSE Performance | 5.10 |
| $R^2$ Performance | 0.78 |
| MAPE Workload | 15.30 |
| Hypervolume | 0.88 |

Figure 7 compares the MPK-generated Pareto front hypervolumes across each of the five cross-validation folds (gold lines with markers) against a single hold-out hypervolume (red dashed line). The cross-validation hypervolume remains consistently high at around 0.89 for folds 1-3, peaking at 0.94 in fold 4, then decreasing slightly to 0.90 in fold 5. The hold-out hypervolume of 0.88 lies just below most of the CV values, indicating that the quality of the model's Pareto front on the data does not appear to closely match its performance during cross-validation. This consistency confirms that the MPK tuning generalizes well, with only minor variability across folds.

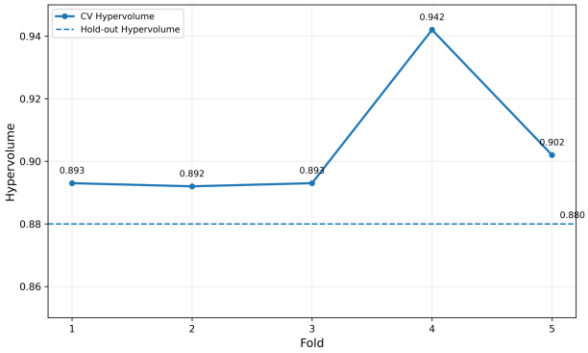



Figure 7. Hypervolume Indicator: CV with Hold-out

Sensitivity analysis shows that the Pareto front derived from MPK remains robust at moderate levels of input noise but degrades as noise increases. With the original unperturbed data (0% noise), the hypervolume indicator is 0.92. Introducing 5% noise causes only a slight decrease to 0.89, and even at 10% noise, the hypervolume remains relatively high at 0.85.

Table 16. Sensitivity Analysis

| Noise Level (%) | Hypervolume |
|---|---|
| 0 | 0.92 |
| 5 | 0.89 |
| 10 | 0.85 |
| 15 | 0.80 |
| 20 | 0.75 |

However, beyond this point the quality of the front deteriorates more rapidly dropping to 0.80 at 15% noise and 0.75 at 20% noise indicating that while MPK can tolerate modest data perturbations, its ability to balance objectives weakens under more extreme perturbations. as can be seen in Table 16 and Figure 8.

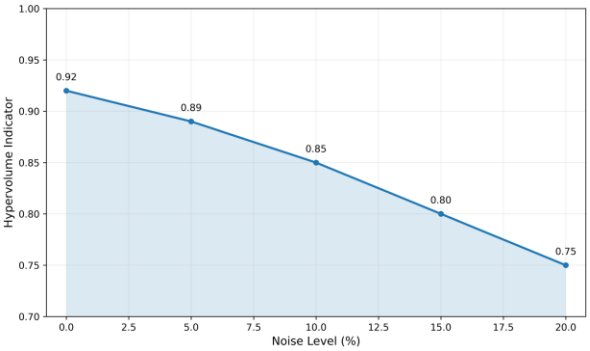



Figure 8. Sensitivity Analysis: Impact of Noise on Hypervolume

## 4. Conclusions

This study demonstrates that integrating Principal Component Analysis (PCA) with Multi-Objective Complexity Prediction (MPK) effectively addresses multidimensional HR data in plantation companies. PCA reduced nine original variables to four principal components explaining over 90% of the variance, thereby eliminating noise and redundancy without sacrificing critical information. The MPK framework then accurately predicted three key indicators absenteeism risk, performance score, and workload imbalance with high accuracy (MAE for absenteeism = 6.5 days; RMSE = 8 days; $R^2$ = 0.82 for absenteeism; MAE for performance = 4.2 points; RMSE = 5.1 points; $R^2$ = 0.78 for performance; and MAPE for workload = 15%). The quality of the Pareto front solutions was consistent, with an average hypervolume indicator of 0.90 in cross-validation and 0.88 in hold-out testing, indicating strong generalization. Sensitivity analysis revealed that the model remains robust to input noise up to approximately 10%, although performance declines at higher noise levels. These results underscore the strategic value of the PCA-MPK hybrid approach for HR decision support, from identifying high-risk employees to designing targeted training programs and optimizing workload distribution.

**Limitations:** A key limitation of this study is the absence of explicit temporal variables (e.g., timestamps) in the modeling dataset, which prevents the implementation of true time-series forecasting to distinguish past from future events. Additionally, the prediction targets were modeled using aggregated patterns, which may overlook sequential dependencies in attendance and productivity. The dataset is specific to a single plantation company, which may limit the generalizability of the results to other sectors or operational contexts.

**Future Work:** Future research should incorporate time-indexed data to enable sequence-aware prediction, explore nonlinear dimensionality reduction techniques such as kernel PCA or autoencoders, and compare MPK with other multi-objective algorithms like NSGA-II to assess potential gains in forecasting performance and interpretability.

## Acknowledgements

Authors thanks to DPPM – KEMDIKTISAINTEK - PTM grand number: 7/SPK/LL1/AL.04.03/PL/2025,74/P3MPI/3.1.1/VI/2025

## References

[1] M. Habibi, *Capitalism and Agrarian Change*. London: Routledge, 2022. doi: 10.4324/9781003267348.

[2] P. Bhattarai et al., "Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions," IET Smart Grid, vol. 2, no. 2, pp. 141–154, Jun. 2019, doi: 10.1049/iet-stg.2018.0261.

[3] Hauser, "Designing Progressive and Interactive Analytics Processes for High-Dimensional Data Analysis," IEEE Trans

Vis Comput Graph, vol. 23, no. 1, pp. 131–140, Jan. 2017, doi: 10.1109/TVCG.2016.2598470.

[4] M. R. Shafie, H. Khosravi, S. Farhadpour, S. Das, and I. Ahmed, "A cluster-based human resources analytics for predicting employee turnover using optimized Artificial Neural Networks and data augmentation," Decision Analytics Journal, vol. 11, p. 100461, Jun. 2024, doi: 10.1016/j.dajour.2024.100461.

[5] M. Talebkhah, A. Sali, M. Gordan, S. J. Hashim, and F. Z. Rokhani, "Comprehensive Review on Development of Smart Cities Using Industry 4.0 Technologies," IEEE Access, vol. 11, pp. 91981–92030, 2023, **DOI:** 10.1109/ACCESS.2023.3302262.

[6] X. Wang and J. Zhi, "A machine learning-based analytical framework for employee turnover prediction," Journal of Management Analytics, vol. 8, no. 3, pp. 351–370, Jul. 2021, doi: 10.1080/23270012.2021.1961318.

[7] S. Verma, M. Pant, and V. Snasel, "A Comprehensive Review on NSGA-II for Multi-Objective Combinatorial Optimization Problems," IEEE Access, vol. 9, pp. 57757–57791, 2021, doi: 10.1109/ACCESS.2021.3070634.

[8] D. Truong, N. C. Le, H. N. The, and M.-H. Nguyen, "Machine Learning and Fuzzy Technique for Environmental Time Series Analysis," 2022, pp. 295–319. DOI:10.1007/978-3-030-96429-0_14.

[9] N. Salem and S. Hussein, "Data dimensional reduction and principal components analysis," Procedia Comput Sci, vol. 163, pp. 292–299, 2019, doi: 10.1016/j.procs.2019.12.111.

[10] Md. P. Uddin, Md. Al Mamun, and Md. A. Hossain, "PCA-based Feature Reduction for Hyperspectral Remote Sensing Image Classification," IETE Technical Review, vol. 38, no. 4, pp. 377–396, Jul. 2021, doi: 10.1080/02564602.2020.1740615.

[11] R. B. Y. Syah, R. Muliono, M. Akbar Siregar, and M. Elveny, "An efficiency metaheuristic model to predicting customers churn in the business market with machine learning-based," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 13, no. 2, p. 1547, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1547-1556.

[12] F. L. Gewers et al., "Principal Component Analysis," ACM Comput Surv, vol. 54, no. 4, pp. 1–34, May 2022, doi: 10.1145/3447755.

[13] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," Nat. Rev. Methods Primers, vol. 2, no. 1, Art. no. 100, Dec. 2022, doi: 10.1038/s43586-022-00184-w.

[14] E. Saccenti, "A gentle introduction to principal component analysis using tea-pots, dinosaurs, and pizza," Teaching Statistics, vol. 46, no. 1, pp. 38–52, Jan. 2024, doi: 10.1111/test.12363.

[15] J. Lever, M. Krzywinski, and N. Altman, "Principal component analysis," Nat. Methods, vol. 14, no. 7, pp. 641–642, Jul. 2017, doi: 10.1038/nmeth.4346.

[16] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," Philos. Trans. R. Soc. A, vol. 374, no. 2065, Art. no. 20150202, Apr. 2016, DOI: 10.1098/rsta.2015.0202

[17] R. B. Y. Syah, H. Satria, M. Elveny, and M. K. M. Nasution, "Complexity prediction model: a model for multi-object complexity in consideration to business uncertainty problems," Bulletin of Electrical Engineering and Informatics, vol. 12, no. 6, pp. 3697–3705, Dec. 2023, doi: 10.11591/eei.v12i6.5380.

[18] W. De Corte, P. R. Sackett, and F. Lievens, "Robustness, Sensitivity, and Sampling Variability of Pareto-Optimal Selection System Solutions to Address the Quality-Diversity Trade-Off," Organ Res Methods, vol. 23, no. 3, pp. 535–568, Jul. 2020, doi: 10.1177/1094428118825301.

[19] R. B. Syah, M. Elveny, R. F. Ananda, and M. K. M. Nasution, "A novel MPK optimization framework for financial data analysis incorporating complexity and uncertainty management," Bull. Electr. Eng. Inform., vol. 14, no. 4, pp. 3121–3132, Aug. 2025, doi: 10.11591/eei.v14i4.9358.

[20] D. Saxena, J. Kumar, A. K. Singh, and S. Schmid, "Performance Analysis of Machine Learning Centered Workload Prediction Models for Cloud," IEEE Transactions on Parallel and Distributed Systems, vol. 34, no. 4, pp. 1313–1330, Apr. 2023, doi: 10.1109/TPDS.2023.3240567.

[21] S. Bates, T. Hastie, and R. Tibshirani, "Cross-Validation: What Does It Estimate and How Well Does It Do It?," J Am Stat Assoc, vol. 119, no. 546, pp. 1434–1445, Apr. 2024, doi: 10.1080/01621459.2023.2197686.

[22] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression," IEEE Signal Process Lett, vol. 27, pp. 1485–1489, 2020, doi: 10.1109/LSP.2020.3016837.

[23] D. S. K. Karunasingha, "Root mean square error or mean absolute error? Use their ratio as well," Inf Sci (N Y), vol. 585, pp. 609–629, Mar. 2022, doi: 10.1016/j.ins.2021.11.036.

[24] M. A. Stoffel, S. Nakagawa, and H. Schielzeth, "partR2": partitioning R 2 in generalized linear mixed models," PeerJ, vol. 9, p. e11414, May 2021, doi: 10.7717/peerj.11414.

[25] 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.

[26] A. Antoniadis, S. Lambert-Lacroix, and J.-M. Poggi, "Random forests for global sensitivity analysis: A selective review," Reliab Eng Syst Saf, vol. 206, p. 107312, Feb. 2021, doi: 10.1016/j.ress.2020.107312.

[27] S. Razavi et al., "The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support," Environmental Modelling & Software, vol. 137, p. 104954, Mar. 2021, doi: 10.1016/j.envsoft.2020.104954.