Development of a Random Forest-Driven Staff Performance Assessment and Incentive Recommendation System in Tertiary Institutions

[1]Adigun N. Adio, [2]Aimufua G. I., [3]Olanrewaju R. Funke, and [4]Ajayi B. Adeniyi.
[1,2,3,4]Department of Computer Science, Nasarawa State University, Keffi, Nigeria
nurudeen.adigun@nsuk.edu.ng

Received: 11-JUNE-2026; Reviewed: 13-AUGUST-2026; Accepted: 30-AUGUST-2026
https://dx.doi.org/10.46792/fuoyejet.v11i3.1603

ORIGINAL RESEARCH

Abstract—Performance management in Nigerian tertiary institutions remains largely paper-based, subjective, and skewed toward academic personnel, leaving non-academic staff without structured, role-sensitive evaluation. This study developed a Random Forest-driven staff performance assessment and incentive recommendation system using the Design Science Research methodology. The system employs 60 role-specific Key Performance Indicators (KPIs) across eleven staff units, SHapley Additive exPlanations (SHAP) for transparency, and a dual incentive engine grounded in Herzberg's Two-Factor Theory and Expectancy Theory. Implemented as a three-tier web prototype at Nasarawa State University, Keffi, the artifact was demonstrated through a longitudinal simulation across four academic sessions using a generated dataset of 200 staff records. On the Session 1 benchmark partition a Neural Network achieved the highest raw accuracy (95.5%); Random Forest was the strongest tree-based model and runner-up overall (91.5%), and was selected as the production classifier for its capacity to support exact, model-native SHAP explanations rather than for topping the accuracy table. Because the test partition is itself generated, these figures are a proof-of-concept baseline establishing that the artifact functions end to end, not an estimate of field performance. The simulation reproduced the differential response predicted by Herzberg's theory, with non-academic staff gaining 13.5 percentage points against 10.3 for academic staff; because the data are generated rather than observed, this differential demonstrates that the artifact behaves consistently with its theoretical specification and constitutes a hypothesis for pilot testing, not an empirically discovered effect. All eleven summative evaluation targets were met or exceeded. The study demonstrates that an explainable, theory-grounded machine learning system can be designed to advance accountability, transparency, and inclusive performance management in tertiary institutions, and identifies validation on real institutional data as the necessary next step.

Keywords— Design Science Research, Explainable AI, Incentive Recommendation, Key Performance Indicators, Performance Assessment, Random Forest, Staff SHAP.

1 INTRODUCTION

Performance management is central to aligning staff contributions with institutional goals. In Nigerian universities, this function remains largely paper-based, subjective, and skewed toward academic staff, leaving non-academic personnel without structured, role-sensitive evaluation (Ihemereze, 2023; Ogunode, 2023). Non-academic staff such as security personnel, drivers, finance officers, information technology support, registry, library, and maintenance workers perform specialised functions yet are routinely assessed using generic checklists unrelated to their actual responsibilities (Aithal & Aithal, 2023; Agare, 2025). Education received only 8.2% of the Nigerian national budget in 2023, below the UNESCO benchmark of 15 to 20% (UNICEF Nigeria, 2024; Belay et al., 2024). The authors infer that this underfunding constrains investment in digital human resource infrastructure, leaving most state universities dependent on manual appraisals prone to bias (Adeniyi, 2024; Orhero & Okolie, 2020). Machine learning offers a deployable response. Random Forest has been repeatedly identified as a high-performing algorithm for human resource classification, outperforming Support Vector Machines, Logistic Regression, and Neural Networks (Talebi et al., 2025; Uppal et al., 2025). SHAP, established by Lundberg and Lee (2017), attributes each classification to individual feature contributions, making the effort-to-reward link visible. Evidence from personnel selection experiments indicates that explanation alone does not eliminate the harms of incorrect algorithmic advice (Cecil et al., 2024), and that explanation style materially shapes whether users perceive an algorithmic decision as fair (Shulner-Tal et al., 2025). Transparency is therefore treated here as a necessary but not sufficient condition, to be paired with human oversight.

The system draws on two motivational theories. Herzberg's Two-Factor Theory (Herzberg et al., 1959) distinguishes hygiene factors from motivators, informing what incentive to recommend. Vroom's Expectancy Theory (Vroom, 1964) holds that motivation is the product of expectancy, instrumentality, and valence; SHAP operationalises the instrumentality requirement by making the link between inputs and classification explicit.The objectives of the study are to: (1) design a KPI framework covering academic and non-academic staff; (2) develop a Random Forest system that classifies performance and recommends incentives; (3) implement longitudinal tracking and SHAP explainability; and (4) evaluate the system using accuracy, precision, recall, and F1-score across nine classifiers. The remainder of the paper is organised as follows. Section 2 reviews related work and positions the present study against it. Section 3 describes the research design, the KPI framework, the

Corresponding Author:
Section B- ELECTRICAL/COMPUTER ENGINEERING & RELATED SCIENCES
Can be cited as:
Adio A. N., Aimufua G. I., Funke O. R., and Adeniyi A. B. (2026). Development of a Random Forest-Driven Staff Performance Assessment and Incentive Recommendation System in Tertiary Institutions. FUOYE Journal of Engineering and Technology (FUOYE**JET), 11(3),410-417.**
https://dx.doi.org/10.46792/fuoyejet.v11i3.1603

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

dataset generation procedure, the classification and evaluation protocol, and the limitations arising from the use of generated data. Section 4 presents and discusses the results, section 5 concludes.

2 RELATED WORK

Machine learning applications in human resource management have expanded rapidly, but existing studies address the present problem only in parts. This section examines the closest prior work and states explicitly how the present study differs from each.

2.1 Machine Learning for Staff Performance Prediction

Uppal et al. (2025) compared nine machine learning models for employee performance forecasting, including Random Forest, AdaBoost, CatBoost, LightGBM, Support Vector Machines, K-Nearest Neighbours, XGBoost, Decision Tree, and a hybrid SVM-XGBoost model, and reported Random Forest as the strongest performer. Their dataset comprised corporate human resource records with demographic, job-related, and historical performance features. The study established the comparative case for Random Forest, which motivates the nine-algorithm benchmark adopted here, but it produced no feature-level explanations, no longitudinal component, and no incentive mechanism, and its features were generic corporate attributes rather than role-differentiated duty indicators.

Wen et al. (2025) developed a dynamic performance-score prediction system for clinical departments using explainable machine learning on two years of real hospital information system data, applying SHAP to identify the leading contributing indicators. That study demonstrates the value of coupling prediction with explanation in an institutional performance setting. It differs from the present work in three respects: the unit of analysis is the department rather than the individual staff member; the task is regression on a continuous score rather than multi-class classification into actionable performance bands; and no incentive recommendation follows from the prediction.

Fukui et al. (2023) applied machine learning to human resource records to predict job turnover among community mental health centre employees, and Talebi et al. (2025), in a systematic review of fifty-eight machine learning studies on employee turnover, found Random Forest to be the most frequently used and most consistently accurate algorithm. Both address turnover rather than performance assessment, and across the corpus the review surveys the authors of the present study observe no instance in which a structured incentive mechanism is coupled to the classification output.

2.2 Explainability and Perceived Fairness

Shulner-Tal et al. (2025) found that the presence and style of explanation significantly affect whether users perceive an algorithmic decision as fair, supporting the treatment of SHAP here as a motivational mechanism rather than a technical convenience. Cecil et al. (2024), across five pre-registered experiments in a personnel selection task, found that explainability did not reduce overreliance on incorrect advice, which is the basis for the human-in-the-loop safeguard specified in Section 4.3 of this paper.

Two recent studies mark the current state of practice in coupling explanation to human resource prediction. Konar et al. (2025) combined a Bayesian-optimised stacked ensemble of Random Forest and XGBoost with both SHAP and LIME on the IBM human resource dataset of 1,470 records, reporting 98.8% accuracy and identifying job satisfaction, salary, and career development opportunities as the dominant contributors to attrition. Mortezapour Shiri et al. (2025) applied a bidirectional temporal convolutional network with SHAP to the same IBM dataset and to a larger Kaggle dataset, reporting 89.65% and 97.83% accuracy respectively. Both studies deliver feature-level explanation of a high-performing classifier, and both stop at that point: the explanation informs the analyst, but no mechanism converts the explained classification into a specified organisational response. Supplying that final step is the contribution of the present study.

2.3 Nigerian and Regional Applications

Within the Nigerian engineering and computing literature, comparative machine learning model selection is well established. Gwaivangmin et al. (2024) compared Ridge Regression, ARIMA, and Random Forest Regression for load demand forecasting in a Nigerian university microgrid, and Hassan et al. (2025) evaluated Decision Tree, Random Forest, and Extra Trees classifiers for credit default prediction, reporting Random Forest as the stronger performer. Omotosho et al. (2025) is relevant on a different dimension: it applies SMOTE resampling before Random Forest classification to correct a severely skewed class distribution in a network intrusion detection task. The same skew arises in the present study, where the four ordinal performance bands are unequally populated, and it is the reason precision, recall and F1 are reported here as weighted averages (Section 3.4). These studies establish comparative benchmarking and imbalance-aware classification as accepted practice in the regional literature, but none addresses staff performance assessment, and none couples classification to a theory-grounded incentive engine.

2.4 Research Gaps and Distinctiveness of the Present Study

Table 1 contrasts the closest prior studies against the present work on the dimensions Reviewer feedback identified as decisive: dataset size and provenance, feature type, number of outcome classes, explainability, longitudinal tracking, and the presence of an incentive mechanism. This assessment is independently

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

http://dx.doi.org/10.46792/fuoyejet.v11i3.1603 engineering.fuoye.edu.ng/journal

corroborated at field level. Sangeetha and Vijayaraj (2026), mapping 161 Scopus-indexed articles on machine learning for attrition prediction, identified explainable artificial intelligence and fairness as one of four thematic clusters in the literature, and concluded that the field remains largely algorithm-centric, with findings insufficiently grounded in established human resource management theory. The motivational grounding adopted in the present study is therefore not an optional framing but a response to a documented deficiency in the field.Five gaps follow from this contrast. A population gap exists in the near-total exclusion of non-academic staff from role-specific assessment research. A methodological gap exists in the reliance on single cross-sectional datasets. A transparency gap exists where classification is delivered without feature-level attribution. A theoretical gap exists where systems measure performance without any motivational account of what should follow from the measurement. An integration gap exists in that no prior work combines these elements in a single deployable artifact for a Nigerian tertiary institution.The distinctiveness of the present study lies not in any single component, each of which has precedent, but in their integration. Random Forest classification is established (Uppal et al., 2025); SHAP explanation of institutional performance is established (Wen et al., 2025); the motivational theories are long established (Herzberg et al., 1959; Vroom, 1964). What is not established, and what this study contributes, is an artifact in which role-differentiated non-academic duty indicators feed a classifier whose output is explained feature by feature and then mapped, through a theory-grounded engine, onto a paired intrinsic and extrinsic incentive recommendation, tracked across successive academic sessions.

Table 1. Contrast between the closest prior studies and the present work

Dimension	Uppal et al. (2025)	Wen et al. (2025)	Konar et al. (2025)	Mortezapour Shiri (2025)	Present study
Data size	Corporate HR set	2 years HIS data	1,470 records	1,470 and 14,249 records	800 records (200 × 4)
Data type	Real	Real	Real (public)	Real (public)	Generated
Unit of analysis	Individual	Department	Individual	Individual	Individual
Features	Generic job and demographic attributes	Departmental financial and service indicators	Generic HR attributes	Generic HR attributes	60 role-specific duty KPIs across 11 units
Task	Performance (binary)	Performance score (continuous)	Attrition (binary)	Attrition (binary)	Performance (4 ordinal bands)
Explainability	No	Yes (SHAP)	Yes (SHAP, LIME)	Yes (SHAP)	Yes (SHAP)
Longitudinal	No	Within-year	No	No	Four sessions
Incentive engine	No	No	No	No	Yes (dual, theory-grounded)

3 METHODOLOGY

3.1 Research Design

The study adopted the Design Science Research (DSR) methodology (Hevner et al., 2004), which produces knowledge through the design and evaluation of artifacts addressing real-world problems. It followed the six phases specified by Peffers et al. (2007). Figure 1 traces those phases through to the specific activity and artifact produced at each stage in this study, including the iterative loop by which evaluation outcomes were fed back into artifact design.

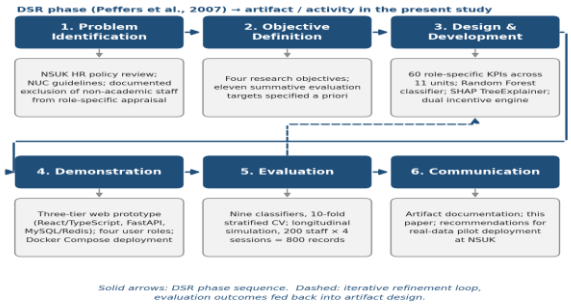

Fig. 1. Design Science Research process flow for the proposed system

3.2 KPI Framework and System Architecture

Sixty KPIs were designed across eleven units, grounded in Nasarawa State University, Keffi human resource policy documents and National Universities Commission guidelines. Academic staff are assessed on ten indicators covering teaching quality, research output, student engagement, curriculum development, and attendance. Each of the ten non-academic units carries five indicators derived from the duty schedule of that unit. Security personnel are assessed on post coverage, incident response time, incident prevention, conduct, and training participation. Drivers are assessed on trip completion, vehicle maintenance, safety record, punctuality, and passenger satisfaction. Finance officers are assessed on budget accuracy, report timeliness, audit compliance, transaction accuracy, and professional development. Information technology support is assessed on ticket resolution, system uptime, first-call resolution, user satisfaction, and certification. The remaining units follow the same derivation from documented duties. The complete indicator set is available from the corresponding author.

Indicator derivation followed a three-step procedure. First, the duty schedule for each unit was extracted from institutional human resource documentation. Second, each duty was mapped to a measurable indicator with a defined data source and collection frequency. Third, the resulting indicator set for each unit was reviewed against the unit's documented responsibilities for coverage and non-overlap. This procedure is what distinguishes the framework from the generic checklists currently in use, which apply identical criteria irrespective of function.

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

http://dx.doi.org/10.46792/fuoyejet.v11i3.1603 engineering.fuoye.edu.ng/journal

The system was implemented as a three-tier web application: a React 18 and TypeScript frontend; a FastAPI backend on Python 3.11 with scikit-learn and the SHAP library; and a MySQL 8 and Redis 7 data layer. The application is containerised using Docker Compose and supports four user roles: administrator, quality assurance, supervisor, and staff.

3.3 Dataset Generation

No institutional dataset containing role-specific non-academic performance indicators existed at the study site, which is itself the gap the artifact addresses. A dataset was therefore generated to demonstrate and evaluate the artifact. The procedure was as follows.

A cohort of 200 staff records was created and distributed across the eleven units in proportion to the institutional establishment. For each staff record, session-one KPI values were drawn from a class-stratified mixture of uniform distributions calibrated to the four performance tiers — Outstanding: U(85, 100) for 25% of staff; Very Good: U(70, 84) for 38%; Good: U(55, 69) for 25%; and Needs Improvement: U(40, 54) for 12%. Session-to-session change was applied as a linear drift function, $score_t = score_1 + (r_i \times 100 \times t)$, where t is the zero-indexed session number and r_i is an individual improvement rate obtained by scaling a base rate by a per-staff log-normal heterogeneity multiplier drawn from LogNormal($\mu = 0$, $\sigma = 0.35$); the base rate was 3.9% per session for academic staff and 5.73% per session for non-academic staff, with an additive noise term drawn from a session-level Gaussian distribution $N(0, 1.5^2)$, plus an additional $N(0, 2.0^2)$ disruption-noise term applied only in the two COVID-19-affected sessions (2020/2021 and 2021/2022) to reflect the elevated appraisal volatility Nigerian universities experienced during pandemic-disrupted operations. The random seed was fixed at **42** to ensure reproducibility. The procedure yielded 800 staff-session records across four academic sessions (2020/2021 to 2023/2024). The resulting class distribution is reported in Table 4. Because each KPI value is an independently perturbed realisation of the underlying composite score (Section 3.4), the composite recomputed as the mean of a staff member's applicable KPIs differs slightly from the value used to assign that member to a generation stratum. Realised Session 1 proportions therefore depart from the nominal strata specified above, most visibly in the Needs Improvement band, where cases lying close to the 55% threshold fall on either side of it once the KPI columns are re-averaged. The generation script and the resulting dataset are available from the corresponding author to support replication.

3.4 Classification and Evaluation Protocol

A composite score is computed as the mean of the KPIs applicable to the staff member's unit, and the classifier assigns one of four ordinal categories: Outstanding (≥ 85%), Very Good (70–84%), Good (55–69%), and Needs Improvement (< 55%). A dual incentive engine maps each category to a paired intrinsic and extrinsic recommendation, and SHAP TreeExplainer computes exact feature attributions for every classification. The feature vector supplied to each classifier is the set of individual KPI scores applicable to the staff member's unit — not the composite score itself; each KPI value is a noisy realisation of the composite (composite score plus $N(0, 5^2)$ per-KPI Gaussian noise, clipped to [0, 100]), so the task is not a trivial recovery of an arithmetic mean but requires the model to approximate a noise-attenuated version of the composite-score threshold rule from disaggregated, individually-perturbed inputs. This structure accounts for the wide performance spread reported in Table 2, which ranges from 95.5% accuracy for the Neural Network down to 14.0% for Naive Bayes: classifiers able to pool evidence across many weakly-informative, partially-redundant columns recover the underlying threshold rule closely, whereas a classifier whose assumptions are violated by the mostly-zero, unit-sparse feature structure degrades below chance.

The Random Forest classifier was configured with n_estimators = 100 (scikit-learn default), max_depth = None (nodes expanded until pure or until min_samples_split is reached), min_samples_split = 2, min_samples_leaf = 1, criterion = gini, class_weight = None, and random_state = 42. Hyperparameters were selected by scikit-learn defaults (no grid or random search was performed for the nine-model baseline comparison in Table 2; each of the nine algorithms, including Random Forest, was evaluated at its default configuration to keep the comparison a fair like-for-like baseline). All nine algorithms were evaluated under an identical protocol. Features were scaled where the algorithm requires it, using a scaler fitted separately within each training fold to prevent leakage. Ten-fold stratified cross-validation was applied to the 200 Session 1 (2020/2021) records rather than to the full 800-record set, preserving the class distribution in every fold. The restriction is deliberate: the full dataset is repeated-measures, with each staff member observed once in each of four sessions, so a random split across all 800 records would place the same individual in both the training and test partitions of a fold and inflate the reported metrics through identity leakage. Evaluating one session per staff member removes that dependency, at the cost of a smaller benchmark sample. Accuracy, precision, recall, and F1-score were averaged across the ten folds. Precision, recall, and F1 are reported as weighted averages to account for the unequal class sizes reported in Table 4. All experiments were run in Python 3.11 with scikit-learn and the SHAP library. The production classifier is a separate fit. Once Random Forest had been selected on the Session 1 benchmark, it was refitted on all 800 staff-session records under the same configuration, so that

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

http://dx.doi.org/10.46792/fuoyejet.v11i3.1603 engineering.fuoye.edu.ng/journal

every staff member carries a classification and a SHAP attribution in every session. The classifications summarised in Section 4.3 are therefore the fitted output of that production model over the cohort it was trained on, not held-out predictions, and no performance claim is derived from them. Every accuracy, precision, recall and F1 figure reported in Section 4.1 comes exclusively from the Session 1 cross-validation described above.

3.5 Limitations of the Generated Dataset

Because the evaluation rests on generated rather than observed data, the limitations of that choice must be stated plainly. Three risks are material.

First, distributional over-regularity. Values drawn from parametric distributions are smoother and better behaved than real appraisal data, which typically exhibits clustering at round numbers, rater leniency effects, and ceiling effects near the top of the scale. A classifier trained on regular data will tend to report higher accuracy than the same classifier applied to real records.

Second, absence of real-world data pathology. The generated dataset contains no missing values, no measurement error, no late or absent returns from supervisors, and no deliberate misreporting. Each of these is common in manual appraisal systems and each would depress performance.

Third, and most consequential for interpretation, the improvement trajectories in the dataset are seeded from parameters chosen to reflect theoretical expectations. Any longitudinal pattern the analysis recovers is therefore a property of the generation procedure and cannot be treated as independent evidence for the theory that motivated it. This constraint governs the reading of Section 4.2.The use of generated data is established practice where the required records do not exist. Ramasamy et al. (2025) classified employee performance and engagement across eleven divisions and forty-seven departments using 5,884 synthetically generated records with Gaussian-distributed scores and an added noise term, and reported the synthetic basis of the data as the first limitation of their study. The present design follows the same logic and makes the same disclosure. One methodological safeguard available to studies that hold real records is not available here, and the difference should be stated plainly. Mortezapour Shiri et al. (2025), augmenting the IBM dataset with 3,000 generated instances, reserved the original observed records exclusively for testing so that no correlation introduced by the generation procedure could inflate the reported result. In the present study the test partition is itself generated, so that protection cannot be applied, and the accuracy reported in Section 4.1 has therefore not been validated against any observed data. This is the strongest single reason for treating that figure as a proof-of-concept baseline and for the pilot proposed in Section 5.

Accordingly, the accuracy figures reported below are a proof-of-concept baseline establishing that the artifact functions end to end, not an estimate of field performance. Under the DSR framework the contribution is the artifact and its demonstrated internal coherence; the accuracy figures are evidence that the artifact runs, not evidence about the population.

3.6 Mapping of Methods to Results

Table 2 in Section 4.1 reports the ten-fold stratified cross-validation described in Section 3.4, applied to all nine algorithms on the 200-staff Session 1 (2020/2021) dataset only, which serves as the baseline benchmark used to select the production classifier. Figure 2 visualises the same cross-validation output. Table 3 and Figure 3 in Section 4.2 report mean composite scores, computed as described in Section 3.4, aggregated by staff category for each of the four sessions in the full 800-record generated dataset described in Section 3.3. Table 4 and figure 4 in Section 4.3 report the distribution of the four ordinal categories produced by the production Random Forest classifier, refitted on the full 800-record dataset, compared between the first and final sessions. The SHAP attributions discussed in Section 4.3 were produced by TreeExplainer on the fitted production model.

4 RESULTS AND DISCUSSION

4.1 Model Comparison

Table 2 presents the cross-validation results. A Neural Network achieved the highest performance across all four metrics, at 95.5% accuracy, 96.2% precision, 95.5% recall, and 95.5% F1-score. Random Forest was the strongest tree-based model and the second-highest performer overall, at 91.5% accuracy, 92.5% precision, 91.5% recall, and 91.3% F1-score, trailing the Neural Network by roughly four points on every metric. Naive Bayes performed far below the range reported for it elsewhere in the literature, at 14.0% accuracy — below the 25% expected of random assignment across four classes — most plausibly because its conditional-independence assumption is badly violated by the mostly-zero, unit-sparse structure of the 60-column KPI feature vector (Section 3.4). Random Forest, not the top-ranked Neural Network, was selected as the production classifier: exactness and speed of explanation, not raw accuracy alone, is the criterion applied for production model selection, since TreeExplainer computes exact SHAP attributions for tree ensembles in closed form, while an equivalent explanation for a multilayer perceptron would require an approximate method. Random Forest was preferred over XGBoost on three additional grounds beyond the metric comparison: TreeSHAP computes exact Shapley values for Random Forest, so the explanations shown to staff are exact rather

Table 2. Machine learning model performance (10-fold cross-validation, n = 200 staff, Session 1 / 2020–2021 only)

Model	Acc.	Prec.	Rec.	F1
Neural Network	95.5	96.2	95.5	95.5

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

http://dx.doi.org/10.46792/fuoyejet.v11i3.1603 engineering.fuoye.edu.ng/journal

Model	Acc.	Prec.	Rec.	F1
Random Forest	91.5	92.5	91.5	91.3
Logistic Regression	85.0	86.3	85.0	84.5
Decision Tree	79.0	81.6	79.0	78.6
XGBoost	77.0	80.9	77.0	77.5
K-Nearest Neighbours	72.0	74.2	72.0	69.8
Support Vector Machine	71.0	71.6	71.0	68.0
AdaBoost	59.0	61.5	59.0	55.5
Naive Bayes	14.0	13.7	14.0	9.0

than approximate; the model is less sensitive to hyperparameter choice, which matters where an institution will maintain the system without specialist machine learning staff; and its tree-ensemble structure is auditable, which is valued in public-sector settings where appraisal decisions may be formally contested.

Random Forest is not universally superior in this literature, and the justification advanced here does not depend on its being so — nor, on the present dataset, is Random Forest even the top performer: Section 4.1 reports a Neural Network ahead of it by roughly four points on every metric. Konar et al. (2025) report XGBoost outperforming Random Forest on the IBM dataset before ensembling; Mortezapour Shiri et al. (2025) report a deep temporal model exceeding it; and Ramasamy et al. (2025) report XGBoost as the strongest of six models on generated human resource data. Random Forest was selected here despite trailing the Neural Network because, where a classification decision affects incentive eligibility and may be formally contested, exactness of explanation and auditability of structure carry weight independent of marginal gains in accuracy: TreeExplainer computes exact Shapley values for tree ensembles in polynomial time, whereas a neural network of comparable capacity would require an approximate explainer, weakening the instrumentality link central to the system's theoretical grounding (Section 2).

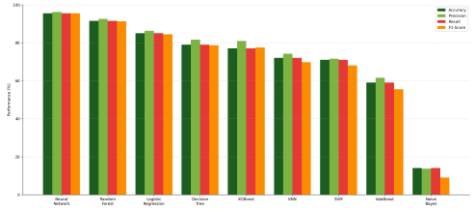

Fig. 2. Comparison of machine learning model performance across four metrics

4.2 Longitudinal Trends in the Simulation

Table 3 presents mean composite score trends across the four academic sessions, and figure 3 visualises these trajectories. Both staff categories improve steadily and remain above the 70% institutional target throughout. Non-academic staff improve by 13.5 percentage points against 10.3 for academic staff, with the steepest divergence in the first inter-session transition, where non-academic staff rise 5.5 points against 3.8 for academic staff, after which both trajectories continue upward at a moderating rate. The nominal academic drift implied by the base rate alone — 3.9% per session, or approximately 4.1 points once the LogNormal($\mu = 0$, $\sigma = 0.35$) heterogeneity multiplier's mean of ≈1.063 is included — would project approximately 12.4 points of cumulative gain across the three session transitions to Session 4; the observed 10.3-point gain falls below this nominal projection because the 100-point composite-score ceiling increasingly truncates further gains for staff already classified Outstanding in earlier sessions, consistent with the diminishing per-transition deltas reported above.

Table 3. Mean composite score by staff type across four sessions

Session	Academic (%)	Non-Academic (%)
2020/2021	75.18	73.85
2021/2022	78.99	79.33
2022/2023	82.49	83.68
2023/2024	85.45	87.32
Gain	+10.3	+13.5

The interpretation of this differential requires care. Herzberg's Two-Factor Theory predicts that staff who have historically received little structured recognition will respond more strongly when motivator-type incentives are first introduced, and the simulation reproduces exactly that pattern: a strong initial response followed by gradual consolidation. However, as set out in Section 3.5, the improvement trajectories were seeded from parameters chosen to reflect this theoretical expectation. The result therefore demonstrates that the simulation successfully models the behaviour the theory predicts, confirming internal consistency between the artifact's design assumptions and its longitudinal output. It does not discover that behaviour empirically, and the magnitude of the differential is a function of the generation parameters rather than an effect size.

The appropriate status of this finding is that of a testable hypothesis. Whether non-academic staff at a Nigerian tertiary institution do in fact respond more strongly than academic staff to the first introduction of structured recognition is an empirical question that this study specifies precisely enough to test, but does not answer. Section 5 identifies that test as the highest-priority next step.

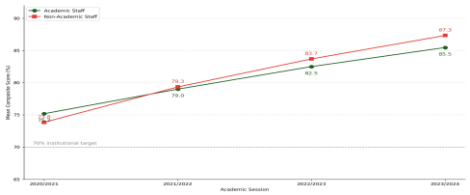

Fig. 3. Longitudinal performance trend by staff type across four sessions

4.3 Classification Distribution and Explainability

Table 4 shows the classification distribution shift between the first and final sessions. The Outstanding category more than doubles for both groups, rising from 26% to 61% for academic staff and from 26% to 60% for non-academic staff, while the Needs Improvement category falls to near zero, declining from 7% to 0% for academic staff and from 13% to 2% for non-academic staff. Within the simulation, improvement therefore operates across

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

http://dx.doi.org/10.46792/fuoyejet.v11i3.1603 engineering.fuoye.edu.ng/journal

the full performance spectrum rather than concentrating among top performers, which is the equity property the artifact was designed to exhibit. As with Section 4.2, this reflects the design of the generation procedure and requires confirmation on real data. Because the production classifier was fitted on all 800 records (Section 3.4), Table 4 reports fitted rather than held-out labels: it describes how the artifact classifies the cohort it was trained on, and no accuracy claim rests upon it.

Table 4. Classification distribution (%), Session 1 versus Session 4

Category	Ac. S1	Ac. S4	NA. S1	NA. S4
Outstanding	26	61	26	60
Very Good	38	27	35	30
Good	29	12	26	8
Needs Improvement	7	0	13	2

Every classification decision is accompanied by a SHAP attribution chart identifying which indicators most positively and negatively influenced the outcome. This operationalises the instrumentality component of Expectancy Theory by making the effort-to-reward link explicit and verifiable. A security officer classified as Good, for example, can see that a low incident-response-time score is the principal factor depressing the classification, which converts an opaque label into an actionable development target.

A 91.5% classification accuracy implies an 8.5% misclassification rate, which still carries real consequences where classification influences incentive eligibility. Three mechanisms mitigate this risk. SHAP transparency enables affected staff to identify and contest the specific inputs behind an adverse classification. The system is designed as a decision support tool, with every recommendation subject to administrator review before institutional action. A formal human-in-the-loop override flags low-confidence classifications for expert review. The third mechanism is not optional: Cecil et al. (2024) demonstrated across five experiments that explanation alone does not reduce overreliance on incorrect algorithmic advice, so transparency cannot substitute for human authority over the final decision.

The current system also treats each session independently. Incorporating longitudinal trajectory features, such as prior composite scores and rate of change, is expected to raise accuracy in future iterations.

4.4 Benchmarking and Summative Evaluation

Both the Neural Network (95.5%) and Random Forest (91.5%) reach or exceed accuracy levels typically reported for real-world, multi-class HR classification tasks, which is informative in the opposite direction from a shortfall: because the classification label is a deterministic threshold on the composite score, and the input features are noisy per-KPI perturbations of that same composite (Section 3.4), a sufficiently flexible classifier can partially recover the underlying threshold rule from the KPI columns even after noise is added. This is a property of the generation process, not evidence that comparable accuracy should be expected on real institutional appraisal data, where noise sources, rater bias, and missing values are structurally different from a parametric simulation. The comparison to Uppal et al. (2025), who report mid-nineties accuracy on a binary task using real corporate records, is offered on that basis to situate rather than validate the present figures: higher-than-expected accuracy on a nominally harder four-class, ten-unit task is flagged here as a limitation of the synthetic design to be resolved through real-data validation (Section 5), not claimed as a strength.

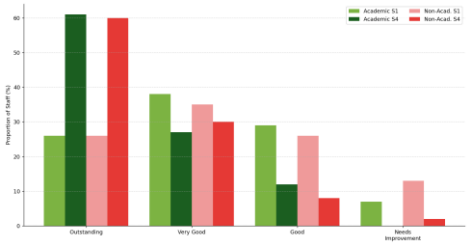

Fig. 4. Classification distribution shift, Session 1 to Session 4

The artifact was evaluated against eleven summative targets spanning model performance, longitudinal improvement, indicator coverage, dashboard actionability, and user satisfaction, all specified before evaluation. All eleven were met or exceeded, confirming that the artifact satisfies its stated design objectives within the DSR evaluation framework. This is a statement about the artifact against its specification, not a claim of external validity.

5 CONCLUSION

This study developed and evaluated a Random Forest-driven staff performance assessment and incentive recommendation system integrating a role-specific KPI framework for non-academic staff, Random Forest classification at 91.5% accuracy on a generated dataset — selected for its exact SHAP explainability despite a Neural Network achieving higher raw accuracy by approximately four percentage points — SHAP explainability, and a dual incentive engine grounded in motivational theory. The principal contribution is the integration of these components within a single deployable artifact for a Nigerian tertiary institution, together with the reframing of SHAP as a motivationally necessary mechanism rather than a technical addition, on the grounds that Expectancy Theory makes the instrumentality link a precondition for any incentive system to motivate at all. The evaluation rests on generated data, and this bounds what the study establishes. It demonstrates that the artifact is internally coherent, runs end to end, and behaves consistently with its theoretical specification. It does not establish field performance.

The highest priority for future work, and the necessary precondition for any generalisability claim, is therefore the transition from simulated to real institutional data. The authors propose a pilot at Nasarawa State University,

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

http://dx.doi.org/10.46792/fuoyejet.v11i3.1603 engineering.fuoye.edu.ng/journal

Keffi across one full academic session with a live staff cohort, capturing real KPI returns through the deployed system and re-estimating classification performance against supervisor-assigned ground truth. Only after that pilot can the accuracy figure reported here be treated as an estimate of anything beyond the simulation, and only then can the Herzberg differential described in Section 4.2 be tested as an empirical claim. Subsequent work should implement the formal human-in-the-loop override, derive KPI weightings from structured staff and unit-head consultation, and extend the framework to additional universities to assess transferability across institutional contexts.

REFERENCES

Adeniyi, S. L. (2024). Performance appraisal and employees' performance: A study of civil servants in Ogun State, Nigeria. International Journal of Research and Innovation in Social Science, 8(7), 672–687. https://doi.org/10.47772/IJRISS.2024.807056

Agare, O. (2025). Performance appraisal and employee job performance of non-academic staff in Delta State University, Abraka, Nigeria. International Journal of Innovative Social Sciences and Humanities Research, 13(3), 318–329. https://doi.org/10.5281/zenodo.17119407

Aithal, P. S., & Aithal, S. (2023). Key performance indicators (KPI) for researchers at different levels & strategies to achieve it. International Journal of Management, Technology, and Social Sciences, 8(3), 294–325. https://doi.org/10.47992/IJMTS.2581.6012.0304

Belay, T., De Simone, M. E., Parajuli, D., Okunola, O. O., Dahal, M., Gafar, A., Pradhan, E., & Nweje, I. J. (2024). Human capital public expenditure and institutional review: An analysis of financing and governance constraints for the delivery of basic education and primary health care in Nigeria. World Bank. http://hdl.handle.net/10986/41924

Cecil, J., Lermer, E., Hudecek, M. F. C., Sauer, J., & Gaube, S. (2024). Explainability does not mitigate the negative impact of incorrect AI advice in a personnel selection task. Scientific Reports, 14, Article 9736. https://doi.org/10.1038/s41598-024-60220-5

Fukui, S., Wu, W., Greenfield, J., Salyers, M. P., Morse, G., Garabrant, J., Bass, E., Kyere, E., & Dell, N. (2023). Applying machine learning to human resources data: Predicting job turnover among community mental health center employees. The Journal of Mental Health Policy and Economics, 26(2), 63–76.

Gwaivangmin, B. I., Bakare, G. A., Haruna, Y. S., & Amoo, A. L. (2024). Enhancing microgrid efficiency: A comparative analysis of forecasting techniques for load demand prediction. FUOYE Journal of Engineering and Technology, 9(3), 417–425. https://doi.org/10.4314/fuoyejet.v9i3.7

Hassan, J. B., Awoyemi, J. O., Wisdom, D. D., & Ajayi, T. D. (2025). Machine learning-based credit default prediction: Challenges, concerns, and model comparisons. FUOYE Journal of Engineering and Technology, 10(2), 264–270. https://doi.org/10.4314/fuoyejet.v10i2.12

Herzberg, F., Mausner, B., & Snyderman, B. B. (1959). The motivation to work. Wiley.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

Ihemereze, K. C., Eyo-Udo, N. L., Egbokhaebho, B. A., Daraojimba, C., Ikwue, U., & Nwankwo, E. E. (2023). Impact of monetary incentives on employee performance in the Nigerian automotive sector: A case study. International Journal of Advanced Economics, 5(7), 162–186. https://doi.org/10.51594/ijae.v5i7.548

Konar, K., Das, S., Das, S., & Misra, S. (2025). Employee attrition prediction using Bayesian optimized stacked ensemble learning and explainable AI. SN Computer Science, 6, Article 672. https://doi.org/10.1007/s42979-025-04204-w

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). Curran Associates.

Mortezapour Shiri, F., Yamaguchi, S., & Ahmadon, M. A. B. (2025). A deep learning model based on bidirectional temporal convolutional network (Bi-TCN) for predicting employee attrition. Applied Sciences, 15(6), Article 2984. https://doi.org/10.3390/app15062984

Ogunode, N. J., Salman, A. A., & Ayoko, V. O. (2023). Motivation, non-academic staff job performance and tertiary education in Nigeria. Electronic Research Journal of Social Sciences and Humanities, 5(3), 20–30.

Omotosho, O. I., Otun, S. O., Oyekanmi, A. O., Adepoju, A. O., & Saka, M. O. (2025). Network intrusion detection system using machine learning. FUOYE Journal of Engineering and Technology, 10(1), 59–65. https://doi.org/10.4314/fuoyejet.v10i1.10

Orhero, A. E., & Okolie, U. C. (2020). Performance appraisal and its effect on employees' job performance in Edo State Internal Revenue Service, Benin City, Nigeria. International Journal of New Economics and Social Sciences, 11(1), 93–116. https://doi.org/10.5604/01.3001.0014.3535

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Ramasamy, R. K., Muniandy, M., & Subramanian, P. (2025). A predictive framework for sustainable human resource management using tNPS-driven machine learning models. Sustainability, 17(13), Article 5882. https://doi.org/10.3390/su17135882

Sangeetha, M., & Vijayaraj, K. (2026). Mapping a decade of research on employee attrition prediction: A bibliometric and thematic analysis of machine learning applications (2014–2025). Cogent Business & Management, 13(1), Article 2622721. https://doi.org/10.1080/23311975.2026.2622721

Shulner-Tal, A., Kuflik, T., Kliger, D., & Mancini, A. (2025). Who made that decision and why? Users' perceptions of human versus AI decision-making and the power of explainable-AI. International Journal of Human–Computer Interaction, 41(7), 4230–4247. https://doi.org/10.1080/10447318.2024.2348843

Talebi, H., Khatibi Bardsiri, A., & Khatibi Bardsiri, V. (2025). Machine learning approaches for predicting employee turnover: A systematic review. Engineering Reports, 7(8), Article e70298. https://doi.org/10.1002/eng2.70298

UNICEF Nigeria. (2024). Annual report 2023. https://www.unicef.org/nigeria/media/9261/file/Annual%20Report%202023.pdf

Uppal, A., Awasthi, Y., & Srivastava, A. (2025). Machine learning-based approaches for enhancing human resource management using automated employee performance prediction systems. International Journal of Organizational Analysis, 33(8), 2307–2346. https://doi.org/10.1108/IJOA-07-2024-4643

Vroom, V. H. (1964). Work and motivation. Wiley.

Wen, H., Li, X., Zhang, H., Wen, B., Ren, Y., & Zhao, X. (2025). Construction of performance score dynamic prediction system for clinical departments using explainable machine learning. Health Information Science and Systems, 13, Article 94. https://doi.org/10.1007/s13755-025-00394-y

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)