IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

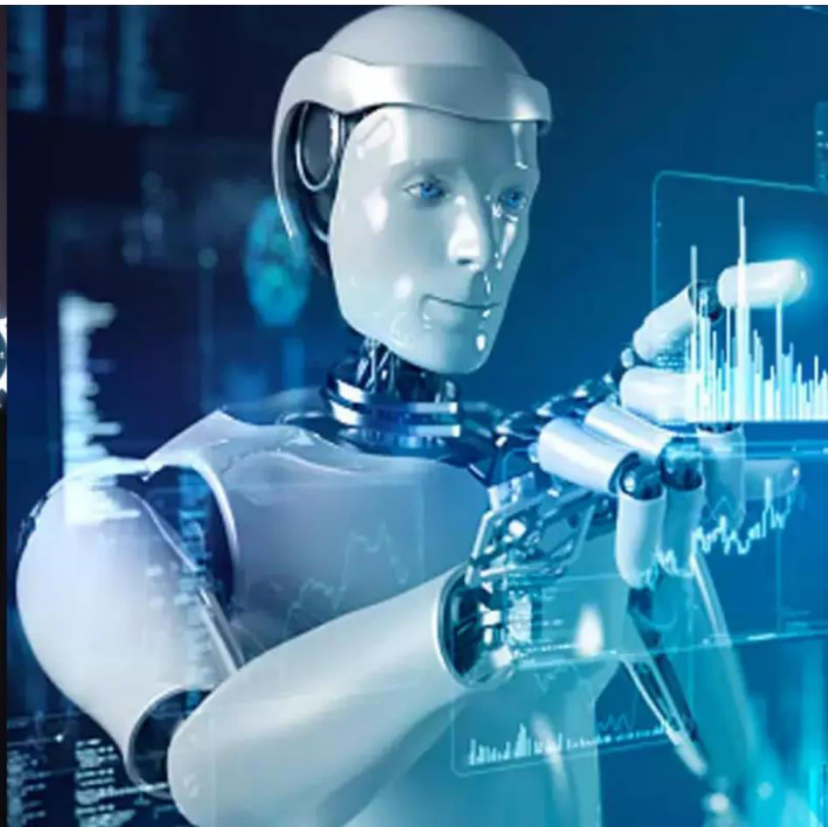
ISSN(O): 2319-8753
ISSN(P): 2347-6710



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Impact Factor: 9.935

Volume 15, Issue 10, October 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

Explainable Ensemble Learning for Employee Performance Prediction: A Surrogate-Data Benchmark and Governance Framework for GHMC Workforce Analytics

C. Narahari Raghu Prasad[1*], Dr. Jaikumar M. Patil[2]

Research Scholar, Computer Science Engineering, Manasarovar Global University Sehore (MP), India[1]

Professor, faculty of Engineering Department of Computer Science Engineering, Manasarovar Global University Sehore (MP), India[2]

Corresponding Author*

ABSTRACT: Municipal workforce analytics has to balance predictive accuracy against transparency, fairness, privacy, and human oversight. This paper proposes an explainable ensemble-learning framework for the Greater Hyderabad Municipal Corporation (GHMC) and tests its technical pipeline on a surrogate human-resources dataset, not on GHMC employee records. The benchmark merges related HR tables into 3,000 employee records with 42 source variables. After privacy, quality, and leakage screening, 25 predictors remain, and these expand to 221 model-ready features after encoding. An 80/20 stratified split gives 2,400 training and 600 test records across four heavily imbalanced performance classes. CatBoost has the best test macro-F1 among the base learners (0.2734), and five-fold cross-validation gives 0.2861 ± 0.0305. Random Forest has the highest ROC-AUC (0.6305), but its class predictions collapse onto the majority class. Weighted voting and stacking do not beat CatBoost on macro-F1. Class-wise results, bootstrap intervals, feature importance, local explanations, and a first fairness audit all point the same way: the surrogate benchmark is not reliable enough for consequential workforce decisions. The contribution is a leakage-aware, explainable, governance-oriented evaluation framework, not a validated GHMC scoring system. A municipal pilot would first need legally approved GHMC data, defensible role-specific targets, grouped and temporal validation, protected-attribute and proxy audits, probability calibration, explanation-stability testing, and human-centred governance.

KEYWORDS: explainable artificial intelligence; ensemble learning; employee performance; people analytics; municipal governance

I. INTRODUCTION

GHMC delivers a wide range of urban services, and their quality depends on the skills of its workforce, the complexity of the work assigned, the resources employees have, and the institutional processes around them [1]. Performance in this setting has many dimensions. A high task count means little if the work is late, has to be redone, breaks procedure, or leaves citizens poorly served. An employee handling complex cases or emergency duties can look less productive on raw counts than a colleague with routine work. Any workable framework therefore has to separate output volume from quality, timeliness, compliance, workload complexity, and resource conditions.

Performance management is now understood as a continuing process that aligns individual contributions with organisational goals, and a century of appraisal research shows that ratings depend heavily on how the measure is designed, how raters think, the context of the work, and the quality of feedback [2]. HR analytics extends this idea by combining information systems and statistical methods for workforce decisions, although reviews caution that its value depends on a clearly defined managerial problem [3]. For GHMC, a predictive system should therefore act as an auditable service-improvement tool and not as an automatic employee-ranking mechanism.

AI can make HR analysis more consistent at scale, but it brings sparse data, changing jobs, legal constraints, employee acceptance problems, and the difficulty of turning predictions into responsible action [4]. Research on human-AI

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

collaboration argues that algorithms should supply pattern recognition while people keep contextual judgement and moral accountability [5], and that automation tends to raise the demand for expert oversight [6]. The framework here accordingly treats AI as decision support for managers, employees, and HR professionals. It excludes autonomous decisions on promotion, transfer, discipline, contract renewal, or termination.

Public-sector AI also has to meet requirements beyond accuracy: legality, due process, public value, and accountability. Reviews report gains in service automation, forecasting, and resource allocation alongside technical, regulatory, ethical, and organisational risks [7], and they call for context-specific empirical evidence [8]. Studies of public administration find that data, infrastructure, skills, leadership, and routines together decide whether AI creates public value [9]. A GHMC programme would therefore need HR staff, department managers, IT, legal and vigilance officers, and employee and data-protection representatives involved from the start.

Explainability matters more for employee predictions than for most applications. XAI taxonomies differ by scope, model type, and explanation type [10], and explanations should be judged on faithfulness, stability, and usefulness to their audience, not only on predictive fit [11]. NIST names four principles for explainable AI: explanation, meaningfulness, explanation accuracy, and knowledge limits [12]. Its AI Risk Management Framework adds governance, mapping, measurement, and management across the system lifecycle [13]. Under these principles, every prediction about an employee should carry an uncertainty estimate, its main feature contributions, and a requirement for human interpretation.

The paper has two parts. The first is a framework for predicting employee performance and operational efficiency at GHMC with explainable ensembles. It aims to (1) define job-specific, workload-normalised indicators; (2) use administrative and operational data without letting protected characteristics act as shortcuts; (3) compare interpretable, bagging, boosting, and stacked models; (4) produce global and local explanations suited to employees and managers; and (5) build fairness, privacy, contestability, and human review into deployment. The second part is an empirical run of the technical pipeline on surrogate HR data. Because that data does not come from GHMC, the run tests the workflow only. It cannot show that the models work for GHMC employees.

II. RELATED WORK

2.1 Personnel analytics and performance prediction

Direct prediction of employee performance is less developed than attrition research. Hasan et al. combined business analytics with machine learning to predict performance and noted the weakness of subjective appraisal data [14]. Earlier data-mining work supported personnel selection and human-capital decisions where defensible attributes and outcomes existed [15]. Attrition studies compared classifiers on heterogeneous employment records [16], [17], [18], but turnover is a different construct from performance. Fuzzy multi-attribute models were proposed to combine several appraisal criteria instead of relying on one score [19]. In the public sector, work on personnel analytics and on AI in local-government performance management discusses opportunities and caveats [20], [21]. None of these studies addresses the job heterogeneity, workload normalisation, explanation stability, and governance needs of a large municipal corporation.

2.2 Ensemble learning and calibration

Random Forest averages decorrelated trees and is a strong baseline for nonlinear tabular data [22]. Extremely Randomized Trees add split randomisation and lower training cost [23]. Gradient boosting builds additive learners that correct earlier errors [24], and XGBoost adds regularisation for scalable training [25]. LightGBM grows trees leaf-wise on histogram-binned features [26], while CatBoost handles categorical variables and limits target leakage from encoding [27]. Stacking combines out-of-fold predictions through a meta-learner [28]. These methods suit workforce data, which contains nonlinear relationships, missing values, categorical job attributes, and interactions between workload and context.

Predictions that feed managerial review also need trustworthy probabilities, and high accuracy does not guarantee them [29]. Platt scaling [30] and isotonic regression [31], fitted on validation data, can repair calibration. A workforce classifier should therefore report a class together with a calibrated confidence, so that uncertain cases go to a human reviewer.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

2.3 Explainability, fairness, and accountability

LIME approximates a complex model around a single prediction [32]. SHAP assigns additive feature contributions from Shapley values [33], and TreeSHAP computes them exactly and efficiently for tree ensembles, which links local and global views [34]. Accumulated local effects reduce the distortion that correlated predictors cause in partial-dependence plots [35], and Anchors give high-precision if-then rules for regions where a prediction holds [36]. Molnar stresses matching the explanation to the question and the audience [37].

Post hoc explanation has limits. Rudin argues that where an interpretable model is accurate enough, it should be preferred for high-stakes decisions [38], so a complex ensemble should have to beat simple alternatives by a meaningful margin before deployment. Data-driven personnel systems can also reproduce historical inequity without using protected classes directly [39], and bias can enter through data collection, labels, features, objectives, evaluation, and deployment [40]. Accountability then calls for internal documentation, stakeholder review, monitoring, and ways to contest results [41]. The framework applies these points by using protected attributes only in fairness audits, attaching an explanation and an uncertainty estimate to every prediction, and barring fully automated adverse decisions.

III. RESEARCH GAPS, QUESTIONS, AND CONTRIBUTIONS

Four gaps stand out. First, most employee-prediction studies target attrition, not performance or efficiency. Second, generic performance datasets lack municipal service complexity, field conditions, statutory timelines, and zone-level demand. Third, accurate models are often published without calibrated uncertainty, tests of explanation stability, or a route for employees to challenge results. Fourth, accuracy is usually estimated on random splits, which can leak employee-specific or temporal information. Five research questions follow. The surrogate benchmark in Section 7 bears only on the technical side of RQ2 (model comparison on a record-level split) and RQ3 (whether explanations can be generated at all).

RQ1. Which administrative, operational, capability, engagement, and environmental variables are most strongly associated with performance and efficiency at GHMC?

RQ2. Does a calibrated stacking ensemble beat individual interpretable, bagging, and boosting models by a practically meaningful and statistically supported margin under employee-grouped and time-based validation? **RQ3.** Are explanations stable, faithful, and understandable across departments, zones, job families, and performance levels? **RQ4.** Are error rates, calibration, and decision thresholds consistent across legally and ethically relevant subgroups, given adequate sample size and lawful audit access?

RQ5. How can model outputs feed developmental feedback, workload balancing, and service improvement while keeping human oversight, contestability, and employee rights?

The contributions are a dual-outcome formulation that separates performance from efficiency; a feature architecture that normalises output by workload and service demand; a calibrated ensemble with transparent fallback models; an XAI protocol that combines global, local, rule-based, and stability evidence; and a governance design covering privacy, fairness, human oversight, contestability, and monitoring. The empirical benchmark tests the analytical pipeline only.

IV. PROPOSED FRAMEWORK

Figure 1 shows the proposed system and the functional components that the rest of this section describes. Data flow left to right through five groups of components (A to E), while governance sits above all of them and lifecycle monitoring below.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

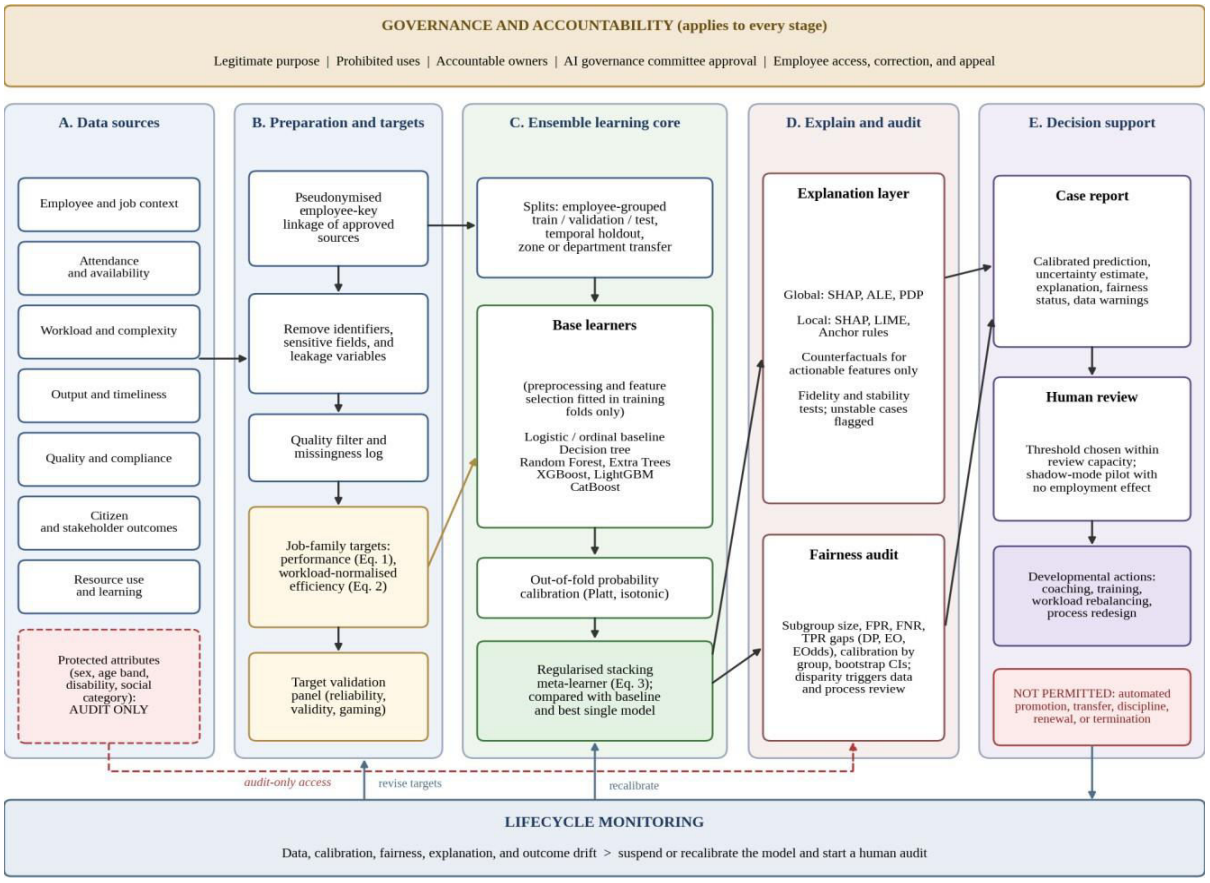


Figure 1. Proposed system architecture for explainable, governed employee-performance analytics: (A) approved data sources, with protected attributes held apart for audit only; (B) privacy-preserving preparation and validated job-family targets; (C) ensemble learning core with calibration and stacking; (D) explanation and fairness audit; (E) human-reviewed decision support limited to developmental actions; governance and lifecycle monitoring span all stages.

4.1 Unit of analysis and outcomes

The preferred unit is the employee-period record, for example employee-month or employee-quarter. Periods capture change over time and allow lagged predictors, and grouping by employee keeps one person out of both training and test folds. Two outcomes are modelled separately for each broad job family (administrative, engineering, sanitation, public health, revenue, town planning, field inspection, and grievance redressal). Performance measures how far expected service objectives are met with acceptable quality and conduct:

$$\text{Performance}_{i,t} = \Sigma_{k=1\ldots K}\, w_k\, z(x_{i,t,k}) \quad (1)$$

Here $x_{i,t,k}$ is the k-th validated indicator for employee i in period t, $z(\cdot)$ standardises it within a peer group, and w_k is a stakeholder-approved weight, which can come from the analytic hierarchy process, expert consensus, or confirmatory factor analysis and should be reviewed each year. Efficiency relates validated output to the time, workload, and resources needed to produce it:

$$\text{Efficiency}_{i,t} = \text{QualityAdjustedOutput}_{i,t} / (\text{WorkHours}_{i,t} \times \text{WorkloadComplexity}_{i,t} \times \text{ResourceFactor}_{i,t}) \quad (2)$$

The denominator stops raw task counts from rewarding easy assignments or penalising hard ones. Quality-adjusted output can draw on first-time-right completion, absence of rework, citizen-service timeliness, and compliance. Extreme values should be winsorised and investigated, not read automatically as excellent or poor performance.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

4.2 Data domains, exclusions, and leakage control

Table 1. Proposed data domains.

Data domain	Illustrative variables	Purpose or restriction
Employee and job context	Job family, grade, tenure band, shift, training history	Sensitive identifiers excluded
Attendance and availability	Authorised hours, approved leave, shift adherence	Adjusted for legitimate leave and duty
Workload and complexity	Cases assigned, backlog, ward or zone demand, severity, emergencies	Denominator for fair comparison
Output and timeliness	Tasks completed, inspections closed, statutory turnaround	Volume and delivery speed
Quality and compliance	Rework, audit observations, first-time-right rate, procedural compliance	Accuracy and reliability
Citizen outcomes	Validated feedback, complaint reopening, escalation	External impact, used cautiously
Resource use and learning	Overtime, equipment use, training completion, collaboration	Efficiency and development
Protected attributes	Sex, age band, disability status, social category	Fairness audit only, never predictors

Free-text supervisor comments, disciplinary narratives, medical data, political affiliation, union activity, biometric templates, and personal addresses should not serve as routine predictors. Where attendance systems are legally adopted, only verified attendance events should reach the analytics layer. Data minimisation, role-based access, encryption, retention limits, and audit logs are mandatory. Source tables are linked through a pseudonymised employee key, and the modelling set must exclude post-outcome information, ratings made after an intervention, and any variable that encodes the target. Missingness is documented as an operational fact rather than simply imputed.

Validation uses three designs. Group or stratified-group splits keep each employee in a single fold. A temporal holdout trains on earlier periods and tests on later ones. Leave-one-zone-out and leave-one-department-out runs check whether the model generalises. Hyperparameter tuning and feature selection stay inside the training folds.

4.3 Ensemble architecture

The candidate set holds ordinal or logistic regression as a transparent baseline and fallback, a constrained decision tree for readable rules, Random Forest and Extra Trees for bagging, and XGBoost, LightGBM, and CatBoost for boosting. Base learners produce out-of-fold predictions, and a regularised logistic or ordinal meta-learner combines them (Eq. 3). Because the meta-learner is trained on predictions for records the base learner never saw, leakage is limited.

$$p_{stack}(x) = g(\alpha_0 + \Sigma_{m=1\ldots M}\ \alpha_m\, p_m(x)) \quad (3)$$

Here $p_m(x)$ is the calibrated prediction of model m and $g(\cdot)$ is the logistic or softmax link. Non-negative or regularised weights can make the stack more stable. If the stack does not clearly beat the transparent baseline under grouped and temporal validation, the simpler model should be used.

4.4 Explanation layer and governance

The explanation layer offers complementary evidence instead of one importance chart: global SHAP summaries, ALE and partial-dependence plots for nonlinear response, local SHAP or LIME for case review, and short Anchor rules. Counterfactual guidance is limited to actionable characteristics such as relevant training, backlog, documentation quality, or turnaround time, and never draws on immutable or protected traits. Each case report shows the prediction, its uncertainty, the reference peer group, the main positive and negative contributions, data timestamps, missing-data warnings, and known model limits. Explanations that change materially under resampling or perturbation are flagged for a human reviewer.

Protected attributes stay outside the development feature store and are available only to an authorised fairness audit, which reports subgroup sizes, false-positive and false-negative rates, calibration, error, and score distributions. A disparity

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

triggers a review of data and process first, since workload allocation, measurement bias, resource differences, or past policy may explain it. Mechanically equalising the numbers is not the response. Outputs guide coaching, training, workload rebalancing, process redesign, and recognition review, and they never decide promotion, transfer, discipline, contract renewal, or termination. Employees can see the data used about them, request corrections, add context, and ask for human reconsideration. An AI governance committee approves use cases, features, thresholds, model versions, monitoring plans, and retirement criteria.

V. PROPOSED RESEARCH METHODOLOGY

5.1 Design, population, and target validation

A longitudinal quantitative design would run alongside qualitative workshops with employees, supervisors, HR specialists, and managers. Data would come only from regular, contractual, and field staff for whom lawful records exist, sampled by job family, grade, department, zone, employment type, and work setting. Two or three departments with good workflow data would form the pilot before any expansion.

Targets should not rest on existing appraisal scores alone, because those can carry rater bias, range restriction, and inconsistent standards. A target-development panel would triangulate operational outcomes, audit results, service-level compliance, supervisor evaluations, and, where suitable, self-reports. It would then check inter-rater reliability, internal consistency, convergent and criterion validity, and incentives for gaming. Rewarding grievance-closure counts without checking quality, for example, invites premature closure. Targets must be finished and documented before any models are compared.

5.2 Training protocol and evaluation

Model development compares raw and workload-normalised targets, alternative feature sets, and class-imbalance strategies. Encoding, imputation, scaling, feature selection, resampling, and calibration are fitted inside training folds only. Hyperparameters are chosen by nested cross-validation over a pre-specified search space, and employee identifiers are used only for grouping and linkage. Evaluation covers discrimination (macro-F1, balanced accuracy, ROC-AUC, PR-AUC); ordinal quality (quadratic weighted kappa, ordinal MAE); regression error for continuous efficiency scores (MAE, RMSE); calibration (expected calibration error, calibration slope, and the Brier score in Eq. 4); operational value (top-risk precision, review workload, decision-curve net benefit); fairness (subgroup F1 or MAE, TPR and FPR gaps, calibration by group); explanation quality (fidelity, stability, sparsity, actionability); and robustness (temporal drift, missingness stress, zone or department transfer).

$$\text{Brier} = (1/N)\ \Sigma_{i=1\ldots N}\ (p_i - y_i)^2 \quad (4)$$

Results are reported as mean and standard deviation over repeated grouped folds with 95% confidence intervals, and paired model comparisons use the Wilcoxon signed-rank test with an effect size. Thresholds follow review capacity and the relative cost of false alerts and missed cases, not the maximisation of one statistic. An ablation plan compares (A) job-context variables only, (B) operational features only, (C) operational features with workload normalisation, (D) the full admissible feature set, (E) the best single model, and (F) the calibrated stack. Sensitivity analyses vary target weights, drop citizen-feedback and attendance variables, change prediction horizons, and compare random, grouped, temporal, and transfer splits. Together these tests show whether apparent gains reflect real signal or leakage, proxies, and split artefacts.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

VI. ALGORITHMIC PROCEDURE

Algorithm 1. Explainable and fair employee-performance ensemble framework.

Input: pseudonymised employee-period dataset D; job families J; indicator weights W; protected attributes A (audit only); candidate models M; explanation methods E.

1. **Govern.** Define the system's purpose, prohibited uses, accountable owners, and the employee appeal process.
2. **Targets.** For each job family j, build a workload-normalised target T_j as a weighted sum of validated indicators.
3. **Data.** Link approved sources through a pseudonymised key, remove identifiers, sensitive fields, and leakage variables, and apply quality filters.
4. **Splits.** Create employee-grouped training, validation, and test sets, a temporal holdout, and a zone or department transfer set.
5. **Base models.** In each training fold, fit preprocessing and feature selection on training data only, train every candidate model, and calibrate its out-of-fold probabilities on validation data.
6. **Stack.** Fit a regularised meta-learner on the calibrated out-of-fold probabilities, then compare the stack, the transparent baseline, and the best single model on all test sets.
7. **Audit.** Compute subgroup gaps in demographic parity, equal opportunity, equalised odds, FPR, and FNR with bootstrap intervals, using protected attributes for the audit only.
8. **Explain.** Generate global (SHAP, ALE, PDP) and local (SHAP, LIME, Anchor) explanations and test their fidelity and stability.
9. **Decide.** After stakeholder review, pick the threshold that maximises utility within human-review capacity H. An approved system enters a shadow pilot with no employment consequences; if pilot criteria are met it is deployed for developmental use only, and otherwise it is revised or rejected.
10. **Monitor.** Track data, calibration, fairness, explanation, and outcome drift. If any measure passes its threshold, suspend or recalibrate the model and start a human audit.
11. **Predict.** For each eligible record, return a calibrated prediction, an uncertainty estimate, an explanation report, a fairness status, and a human-review recommendation.

Output: calibrated prediction, uncertainty, explanation report, fairness status, and human-review recommendation for every eligible record.

VII. RESULTS AND DISCUSSION

7.1 Surrogate data and preprocessing

The benchmark used four related surrogate HR datasets covering employee information, engagement surveys, recruitment, and training and development. None of them is GHMC data. The employee table held 3,000 records and 26 variables, and merging the engagement and training tables on an employee identifier gave 3,000 records and 42 source variables. Recruitment records were left out because no compatible employee-level identifier existed. The target was the available performance score, mapped to four classes: Fully Meets, Exceeds, Needs Improvement, and Performance Improvement Plan (PIP). After removal of direct identifiers, personal data, high-cardinality text fields, constant variables, and selected sensitive attributes, 25 candidate predictors remained and expanded to 221 features after one-hot encoding. An 80/20 stratified split produced 2,400 training and 600 test records.

The classes are badly imbalanced (Fig. 2a). Fully Meets makes up 78.70% of records, Exceeds 12.30%, Needs Improvement 5.90%, and PIP 3.10%. Overall accuracy is therefore a poor guide, and the analysis also reports balanced accuracy, macro-F1, MCC, Cohen's kappa, class-specific recall, and ROC-AUC. Because the split is at record level and not grouped by employee or time, everything below is a preliminary surrogate benchmark, not deployment evidence.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

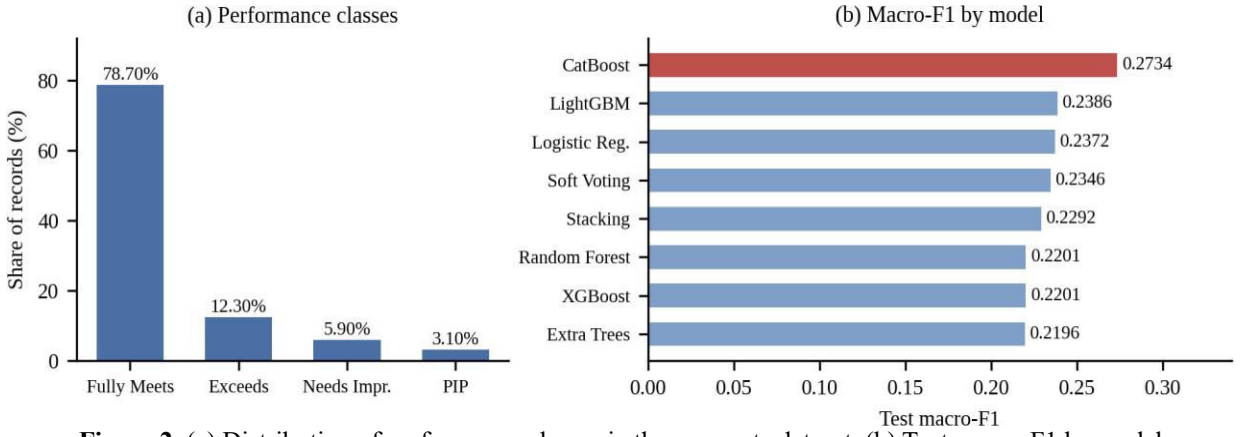


Figure 2. (a) Distribution of performance classes in the surrogate dataset. (b) Test macro-F1 by model.

7.2 Cross-validation performance

Six algorithms were evaluated with five-fold stratified cross-validation. CatBoost had the highest mean macro-F1 (0.2861, SD 0.0305), followed by Logistic Regression (0.2447, SD 0.0160) and LightGBM (0.2354, SD 0.0103). Random Forest, Extra Trees, and XGBoost all sat near 0.22. CatBoost's fold scores ran from 0.2357 to 0.3081, so its result depends visibly on fold composition. It is the strongest model on this measure, but the low absolute score and the fold-to-fold spread rule out any claim of deployment-ready performance.

7.3 Comparative model performance

Table 2 gives the test-set results. CatBoost has the highest macro-F1 (0.2734), so it is the main model for the class-wise and explanation analyses. Its accuracy is 0.5800, balanced accuracy 0.2880, weighted F1 0.6189, MCC 0.0883, and ROC-AUC 0.5870, which is limited separation of the four classes even though it leads on macro-F1. It beats LightGBM by 0.0348 macro-F1 (about 14.6% in relative terms; Fig. 2b). With a low absolute score and no paired significance test across repeated grouped or temporal splits, that gap is not evidence of real predictive capability.

Table.2. Test-set performance of the baseline and ensemble models.

Model	Accuracy	Balanced acc.	Macro F1	Weighted F1	MCC	ROC-AUC
CatBoost	0.5800	0.2880	0.2734	0.6189	0.0883	0.5870
LightGBM	0.7483	0.2521	0.2386	0.6857	0.0056	0.6010
Logistic Regression	0.3717	0.3253	0.2372	0.4630	0.0874	0.6230
Weighted Soft Voting	0.7717	0.2538	0.2346	0.6931	0.0127	0.6217
Stacking Ensemble	0.3600	0.3118	0.2292	0.4530	0.0899	0.5993
Random Forest	0.7867	0.2500	0.2201	0.6927	0.0000	0.6305
XGBoost	0.7867	0.2500	0.2201	0.6927	0.0000	0.6128
Extra Trees	0.7833	0.2489	0.2196	0.6911	-0.0229	0.6231

Random Forest and XGBoost both reach 0.7867 accuracy, with balanced accuracy of 0.2500 and macro-F1 of 0.2201. With four classes, 0.25 balanced accuracy is nominal chance-level recall, and their MCC values are about zero, so the high accuracy comes from the dominant Fully Meets class. Weighted soft voting reaches 0.7717 accuracy but only 0.2346 macro-F1, and stacking reaches 0.3600 accuracy and 0.2292 macro-F1. Neither ensemble passes CatBoost on the pre-specified macro-F1 criterion. Combining learners did not help minority-class recognition here because several components share the same majority-class bias.

Logistic Regression trades things differently. Its accuracy is only 0.3717, yet it has the highest balanced accuracy (0.3253) and a ROC-AUC of 0.6230. Random Forest holds the top ROC-AUC (0.6305) despite majority-biased class decisions. Discrimination, class balance, thresholds, and calibration therefore need joint evaluation, not selection on one headline score.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

7.4 Class-specific performance

Table 3. Class-wise performance of the CatBoost model.

Performance class	Precision	Recall	F1-score	Test support
Exceeds	0.1712	0.3378	0.2273	74
Fully Meets	0.8264	0.6758	0.7436	472
Needs Improvement	0.0652	0.0857	0.0741	35
PIP	0.0455	0.0526	0.0488	19
Macro average	0.2771	0.2880	0.2734	600
Weighted average	0.6765	0.5800	0.6189	600

CatBoost does reasonably on Fully Meets (precision 0.8264, recall 0.6758, F1 0.7436) and poorly elsewhere. Recall is 33.78% for Exceeds, 8.57% for Needs Improvement, and 5.26% for PIP. The two smallest classes are not detected reliably enough for consequential use, and they matter operationally because missed cases could delay training or workload review. Later experiments should try class-weighted objectives, threshold optimisation, balanced resampling inside training folds, cost-sensitive learning, and ordinal formulations. Any synthetic oversampling must stay within training folds and be checked with care on mixed administrative data.

7.5 Confidence intervals

Bootstrap analysis with 500 resamples of the test set produced the intervals in Table 4. The macro-F1 interval runs from 0.2385 to 0.3137, and the balanced-accuracy interval starts at 0.2459, just below the four-class chance level of 0.25. The results are preliminary and do not show that employee performance can be predicted reliably in a municipal setting.

Table 4. Bootstrap 95% confidence intervals for the CatBoost model.

Metric	Observed	Lower 95% CI	Upper 95% CI
Accuracy	0.5800	0.5408	0.6175
Macro F1-score	0.2734	0.2385	0.3137
Balanced accuracy	0.2880	0.2459	0.3380
Weighted F1-score	0.6189	0.5795	0.6539

7.6 Explainability

Table 5 lists the ten highest-ranked variables in the global importance analysis of the CatBoost model. Location code ranks first, followed by job title, tenure, satisfaction, training cost, work-life balance, engagement, managerial role, training duration, and department. These are descriptive associations inside the surrogate data. They are not causal effects and do not justify interventions, and a highly ranked location or role variable may stand in for organisational structure, workload, or other proxies, which needs governance review before any real use.

Satisfaction, work-life balance, engagement, and training variables fit the framework conceptually, but importance does not show that changing them would raise performance.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

Table 5. Ten most influential variables in the CatBoost model.

Rank	Feature	Importance
1	Location code	0.115331
2	Production Technician I job title	0.074378
3	Employee tenure	0.072777
4	Satisfaction score	0.064400
5	Training cost	0.060687
6	Work-life balance score	0.052343
7	Engagement score	0.045507
8	Production Manager job title	0.041240
9	Training duration	0.039908
10	Production department	0.035022

A local LIME explanation for a Fully Meets prediction confirmed that case-level explanations can be produced. In practice they should help reviewers see which recorded factors were associated with a prediction, next to uncertainty and data-quality warnings. They should not serve as causal advice or as a replacement for managerial judgement.

7.7 Fairness and data-governance assessment

The generated fairness summary showed zero differences (0.0000) in selection rate, true-positive rate, and accuracy across the audited age groups. That result is not evidence of fairness. Only age had been designated as a protected attribute in the audit, while model-ready features and explanation outputs still contained variables derived from gender and marital status, so feature screening did not fully implement the stated policy. An exit-date-derived variable also appeared among the influential features, which points to possible temporal leakage, since exit information may be recorded after or in response to a performance outcome.

Before any further benchmark, fields tied to gender, marital status, date of birth, race or ethnicity, disability, religion, and exit timing should be reviewed and removed from predictive inputs where appropriate. Location, department, and job title should be checked for proxy effects and operational legitimacy. A revised audit should give group sizes, class-specific false-positive and false-negative rates, calibration, selection rates at operational thresholds, and uncertainty intervals for every sufficiently large and lawfully auditable subgroup. Parity metrics such as demographic-parity, equal-opportunity, and equalised-odds differences are useful diagnostics, but they should be read against role comparability, target validity, and the consequences of errors, not optimised mechanically.

7.8 Overall discussion

The experiment shows that the workflow runs end to end. It integrates several HR tables, trains diverse models, builds voting and stacking ensembles, computes imbalance-sensitive metrics and bootstrap intervals, produces global and local explanations, and audits governance. These are workflow results, not evidence of operational effectiveness at GHMC.
Predictive strength is weak. CatBoost's MCC is 0.0883, minority-class recall is low, and most of its correct predictions fall in Fully Meets. Ensembling did not fix the imbalance, and the best model depended on the metric: CatBoost on macro-F1, Logistic Regression on balanced accuracy, Random Forest on ROC-AUC. No model can be called best, and the benchmark cannot support personnel decisions. The accuracy trap is plain in Random Forest and XGBoost, which reached about 78.67% accuracy at near-chance balanced accuracy.

A real GHMC study would retrain and re-evaluate the models on legally approved, pseudonymised municipal data with validated measures of service quality, task completion, grievance resolution, attendance, training, workload, response time, resource constraints, and role context. Targets would be co-developed with subject experts and with employee and governance stakeholders. Automated decisions on promotion, discipline, pay, transfer, contract renewal, or termination stay out of scope. The near-term use is developmental decision support with explanation, uncertainty reporting, human review, contestability, and continuing fairness and drift monitoring.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

VIII. LIMITATIONS AND FUTURE RESEARCH

Five limitations matter most. The results come from surrogate HR data, so they do not establish predictive validity for GHMC employees. The stratified record-level split does not test employee-grouped, temporal, department-held-out, or zone-held-out generalisation. The available labels may embed appraisal bias and cannot represent the multidimensional municipal outcomes the framework proposes. Severe class imbalance, uncertain calibration, weak minority-class recall, and the governance problems found in the audit limit how far the benchmark can be interpreted. Finally, global and local explanations describe model behaviour and do not identify causal effects.

Future work should start with legally approved GHMC data and job-family-specific target validation. It should add grouped and temporal evaluation, probability calibration, multi-seed experiments, explanation-stability analysis, proxy and leakage audits, participatory design with employee representatives, and prospective shadow-mode evaluation before any consequential use. Cross-municipality studies could then test transferability, and longitudinal studies could test whether model-supported coaching or workload changes improve service and employee outcomes without creating inequitable effects.

IX. CONCLUSION

This paper presents an explainable ensemble-learning and governance framework for municipal workforce analytics and tests its technical pipeline on a surrogate HR dataset. The framework combines workload-aware targets, leakage-controlled model development, calibrated ensembles, grouped and temporal validation, multi-method explanations, fairness auditing, and human governance. The benchmark confirms that the workflow can be run end to end. It also exposes real weaknesses: macro-F1 stays low, minority-class recall is poor, ensembles do not consistently improve discrimination, and the first fairness audit reveals feature-screening gaps and possible leakage. The empirical results should therefore not be read as a validated GHMC employee-performance model. The framework is best taken as a research protocol for a future, legally authorised GHMC pilot in which targets, data quality, calibration, subgroup performance, explanation stability, and operational value are validated prospectively. Its role is to support human inquiry and developmental interventions, not to automate high-impact employment decisions.

Declarations.
Data and code: the benchmark uses surrogate HR data and no GHMC employee records. Any future GHMC study should document dataset provenance, lawful access, preprocessing code, software versions, random seeds, and a reproducible analysis protocol, subject to privacy and institutional restrictions.

Ethics and governance: the framework is meant for research and developmental decision support and not for fully automated high-impact employment decisions; any municipal use requires lawful data access, purpose limitation, data minimisation, human review, contestability, fairness auditing, and institutional approval.

Conflict of interest and funding: to be completed by the authors as the journal requires.

REFERENCES

1. Greater Hyderabad Municipal Corporation, "About GHMC," GHMC official website, 2026. [Online]. Available: https://www.ghmc.gov.in/
2. A. S. DeNisi and K. R. Murphy, "Performance appraisal and performance management: 100 years of progress?" J. Appl. Psychol., vol. 102, no. 3, pp. 421-433, 2017.
3. J. H. Marler and J. W. Boudreau, "An evidence-based review of HR analytics," Int. J. Hum. Resour. Manag., vol. 28, no. 1, pp. 3-26, 2017.
4. P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: Challenges and a path forward," Calif. Manag. Rev., vol. 61, no. 4, pp. 15-42, 2019.
5. M. H. Jarrahi, "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," Bus. Horiz., vol. 61, no. 4, pp. 577-586, 2018.
6. S. Raisch and S. Krakowski, "Artificial intelligence and management: The automation-augmentation paradox," Acad. Manag. Rev., vol. 46, no. 1, pp. 192-210, 2021.
7. B. W. Wirtz, J. C. Weyerer, and C. Geyer, "Artificial intelligence and the public sector: Applications and

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

challenges," Int. J. Public Adm., vol. 42, no. 7, pp. 596-615, 2019.
8. A. Zuiderwijk, Y.-C. Chen, and F. Salem, "Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda," Gov. Inf. Q., vol. 38, no. 3, Art. 101577, 2021
9. A. van Noordt and L. Tangi, "The dynamics of AI capability and its influence on public value creation of AI within public administration," Gov. Inf. Q., vol. 40, no. 4, Art. 101860, 2023.
10. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion, vol. 58, pp. 82-115, 2020.
11. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.
12. P. J. Phillips et al., Four Principles of Explainable Artificial Intelligence, NISTIR 8312, Natl. Inst. Stand. Technol., Gaithersburg, MD, USA, 2021.
13. National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, Gaithersburg, MD, USA, 2023.
14. M. R. Hasan, R. K. Ray, and F. R. Chowdhury, "Employee performance prediction: An integrated approach of business analytics and machine learning," J. Bus. Manag. Stud., vol. 6, no. 1, pp. 215-219, 2024.
15. C.-F. Chien and L.-F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry," Expert Syst. Appl., vol. 34, no. 1, pp. 280-290, 2008.
16. V. V. Saradhi and G. K. Palshikar, "Employee churn prediction," Expert Syst. Appl., vol. 38, no. 3, pp. 1999-2006, 2011.
17. R. Punnoose and P. Ajit, "Prediction of employee turnover in organizations using machine learning algorithms," Int. J. Adv. Res. Artif. Intell., vol. 5, no. 9, pp. 22-26, 2016.
18. S. N. Khera and Divya, "Predictive modelling of employee turnover in Indian IT industry using machine learning techniques," Vision, vol. 23, no. 1, pp. 12-21, 2019.
19. T. R. Manoharan, C. Muralidharan, and S. G. Deshmukh, "An integrated fuzzy multi-attribute decision-making model for employees' performance appraisal," Int. J. Hum. Resour. Manag., vol. 22, no. 3, pp. 722-745, 2011.
20. W. Cho, S. Choi, and H. Choi, "Human resources analytics for public personnel management: Concepts, cases, and caveats," Adm. Sci., vol. 13, no. 2, Art. 41, 2023.
21. G. Maake and C. M. Schultz, "Using AI in performance management: A global analysis of local government practices," Adm. Sci., vol. 15, no. 10, Art. 392, 2025.
22. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.
23. P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Mach. Learn., vol. 63, no. 1, pp. 3-42, 2006.
24. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189-1232, 2001.
25. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785-794.
26. G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Adv. Neural Inf. Process. Syst. 30, 2017, pp. 3146-3154.
27. L. Prokhorenkova et al., "CatBoost: Unbiased boosting with categorical features," in Adv. Neural Inf. Process. Syst. 31, 2018, pp. 6638-6648.
28. D. H. Wolpert, "Stacked generalization," Neural Netw., vol. 5, no. 2, pp. 241-259, 1992.
29. A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd Int. Conf. Mach. Learn., 2005, pp. 625-632.
30. J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, Cambridge, MA, USA: MIT Press, 1999, pp. 61-74.
31. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2002, pp. 694-699.
32. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 1135-1144.
33. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst. 30, 2017, pp. 4765-4774.
34. S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat. Mach. Intell., vol. 2, no. 1, pp. 56-67, 2020.
35. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," J. R. Stat. Soc. B, vol. 82, no. 4, pp. 1059-1086, 2020.
36. M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in Proc. AAAI

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510106|

Conf. Artif. Intell., vol. 32, no. 1, 2018, pp. 1527-1535.

37. Molnar, Interpretable Machine Learning, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
38. C. Rudin, "Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206-215, 2019.
39. S. Barocas and A. D. Selbst, "Big data's disparate impact," Calif. Law Rev., vol. 104, no. 3, pp. 671-732, 2016.
40. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Comput. Surv., vol. 54, no. 6, Art. 115, 2021.
41. D. Raji et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in Proc. 2020 Conf. Fairness, Accountability, and Transparency, 2020, pp. 33-44.



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details