

An Explainable AI Approach to Employee Performance Evaluation Using Predictive Modeling

Arun B Prasad¹, Mohammed Yaseen Puthen Parambil Aboobaker², Dr. Shaik Balkhis Banu³,
Dr. Satish Jha⁴

¹Associate Professor (Economics), Institute of Law, Nirma University, Ahmedabad, 382481, Gujarat, India, arunprasad16@gmail.com

²Level lead / Lecturer, level 4/ CertHE, London Academy For Applied Technology (LAAT)
yaseenaboobacker144@gmail.com

³Assistant Professor, Department of Physiotherapy, Fatima College of Health Sciences, Al Ain, United Arab Emirates, drshaikbalkhis@gmail.com

⁴Dean, Faculty of Hotel Management and Catering Technology, Parul University, Waghodia, Vadodara, Gujarat, 391760, India, satish.jha144400@paruluniversity.ac.in

Abstract: While employee performance analytics can aid in more consistent workforce evaluations, prediction models may exhibit unfair correlations and opaque predictions. The paper focuses on an analysis of the explainable machine learning approach to classification of employee performance grades based on a publicly available synthetic dataset with 100,000 records and 20 features. Four classifiers, namely multinomial logistic regression, decision tree, random forest, and histogram-based gradient boosting were compared by means of a stratified 80:20 train-test split. In a feature ablation study, both with and without inclusion of the monthly salary feature, logistic regression, random forest, and histogram-based gradient boosting showed 100% accuracy, and decision tree – 94.32%. With the removal of salary, all the classifiers showed almost similar performance around the 20% random guess baseline across five near-balanced classes. Additionally, salary proved to be most important in decision trees, contributing an estimated 76.36% towards the features used for classification. This is a clear indication that the predictive ability observed in this case can be a function of the salary feature as a proxy or possibly a leaking feature and not a reflection of the broader characteristics of employees. This is a demonstration of the reason why high accuracy is not enough proof of validity in employee-performance prediction. The data is entirely artificial in nature; hence the findings are diagnostic in nature only.

Keywords: Explainable artificial intelligence, employee performance, predictive modeling, SHAP, feature ablation, target leakage, human resource analytics

1. Introduction

Performance appraisal of employees determines the allocation of training, promotion, career development, recognition, and organization planning. The traditional methods of appraisal consist of manager's judgment, ratings at intervals, achievement of objectives, and contextual observation. While such methods are capable of measuring qualitative elements of the job, they are vulnerable to factors like inconsistent standards, recency effects, inadequate observation, and variation in management expectations. Modeling can provide an approach of examining numerous characteristics of employees, but a model predicting a rating without providing the logic behind it will be hard to analyze by both parties.

Explanation is thus pertinent to the context of people analytics. Explanation enables a stakeholder to understand which input variables affect the model, if the prediction is based on an inappropriate proxy, and if the model differs in its predictions for different groups of employees. However, explanation is not justification. The variable may be influential since it is correlated with the target, embedded into the process of creating the target, or assessed post-hoc after the event. The explanation of the association does not imply that the variable is fair, causative, or appropriate for employment decision-making. Another concern about algorithmic advice in personnel context is the finding that explanation cannot prevent the negative consequences of wrong recommendations [1].

The recent advances in human resource analytics focus on the conflict between the effectiveness and the interpretability of predictive algorithms. According to empirical research with organizational data, ex post explanations of algorithms improve their interpretability while taking into account the fact that interpretation is

possible only when data, context, and the explanation technique are appropriate [2]. Another problem relates to the fact that prediction models can be assessed in a way that corresponds to their purpose. The measures of performance could be the results in the past, which cannot be predicted [3].

This report describes a clearly laid-out experimental process for classifying the performance of employees. It aims to compare four classifiers, estimate the effect of excluding the monthly salary, analyze the correlation between the monthly salary and performance labels, and explain the implications of these findings on an explainable evaluation framework. The main research question posed by this investigation is whether the workforce data provided contains any predictive value, after accounting for a possibly problematic variable. This report neither claims that the fabricated data mimics a real workforce nor that the model can hire people.

2. Literature Review

Explainable artificial intelligence refers to methods for making model behavior explainable by humans. Intrinsically interpretable models, including linear models and shallow decision trees, reveal their internal structure, while post-hoc methods provide an approximation or explanation of the behavior of the fitted model. The method known as SHapley Additive exPlanations (SHAP) is based on the concept of game-theoretic attribution. While these types of explanations might prove to be valuable on the global and local scales, the information provided relates to the model under a specific distribution and background setting, not causal relationships. In HR, the difference is crucial because a high-ranking feature might reflect a legacy company policy or simply be an artifact of the data, not an optimal factor influencing performance [4].

Work on machine learning using real-world HR data points to the fact that prediction accuracy and interpretability do not necessarily go hand in hand. Post hoc explanation techniques will assist in explaining complex model behavior; however, organizations need to evaluate if there is a simpler model that gives enough prediction accuracy, as well as validity and actionability of the explanation [5]. Therefore, accuracy cannot be the only basis for choosing models. Interpretability, provenance of data, future consequences, and decision point need to be considered.

Research into another area pertains to responsible use of AI-produced evaluations. This is critical since an explanation might boost confidence even when the advice is not correct. This becomes even more relevant in the area of employee evaluation, since an incorrect score may determine career prospects, salary, or reputation. In the development of a reasonable system for explaining performance prediction, there must be separation between description and causality, a means to dispute the output and no adverse decision-making on the basis of the model output alone [6].

The research gap that is filled through this paper is more methodological in nature rather than the notion that explainable performance prediction is novel. What the research does is carry out a leakage aware diagnostic comparison on a large synthetic dataset, while explicitly evaluating the impact of a commonly used compensation factor on model performance. Here, low performance is taken as an indicator of what is possible from the data, and not hidden.

3. Methodology

The work flow is intended to isolate the process of testing datasets, comparing models, and diagnosing validity of features. This can be illustrated in Figure 1. It will aim for five-class prediction of Performance_Score with scores ranging from 1 to 5 kept as categorical values.

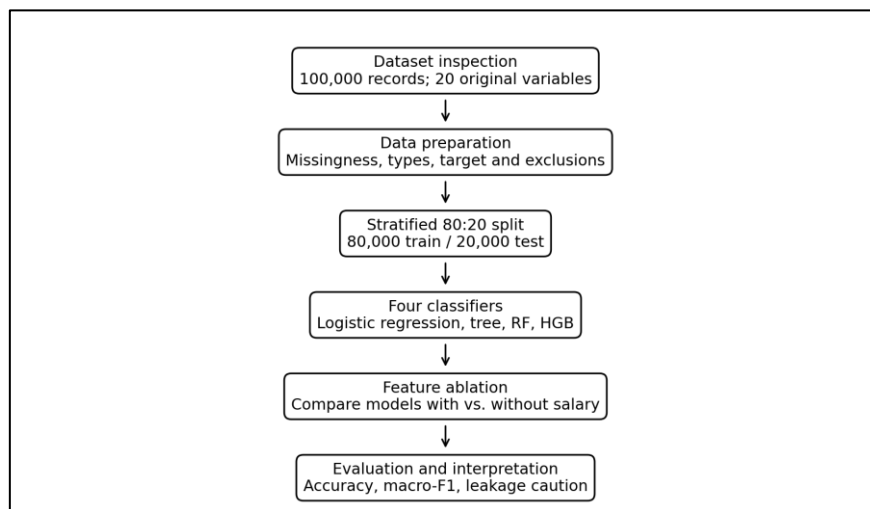


Figure 1. Methodology workflow for data inspection, modeling, feature ablation, and evaluation.

Figure 1 represents the suggested approach for the explainable employee performance assessment through predictive modelling. The first step includes the collection of the dataset and its preparation, including data cleaning, feature selection, and preparation of the target variable `Performance_Score`. The prepared dataset is split into training and testing parts, where multiple models are evaluated in terms of their predictiveness. These models include Logistic Regression, Decision Tree, Random Forest, and HistGradientBoosting. The performance of these models is estimated by such measures as accuracy, balanced accuracy, and macro F1 score. For better interpretability, feature importance analysis is done, and there is an additional comparative experiment where the performance of the model is investigated with and without the salary variable to reveal the dependency of the prediction on certain features.

3.1 Dataset and Variables

The dataset used in this analysis has 100,000 rows and 20 attributes. The list of features present is department, gender, age, job designation, date of hiring, experience, education, performance rating, monthly salary, working hours, number of projects taken care of, overtime hours, sick leave hours, remote work frequency, team size, training hours, employee satisfaction, resignation, and other employee-related features. The dataset has been described by its creators as the synthetic employee performance and productivity dataset [7]. Since the records are synthetic, they do not describe actual experiences of employees.

`Performance_Score` was chosen as the target variable. `Employee_ID` was removed as an identifier variable, which does not contribute meaningfully to the prediction model. The original date of hiring was excluded in the baseline set of features, but a new set of features based on the year and month of hiring was included in the experimental set of features. The `Resigned` variable was removed to ensure that the prediction of post-employment information does not get mixed up with the performance prediction task. One-hot encoding was done for categorical features, while medians were used to fill the null values in the numerical features.

3.2 Data Preparation and Validation

The target variable was analyzed before the modeling process took place. Stratified sampling allocated 80,000 instances in the training set while 20,000 instances were reserved for testing. Preprocessing procedures were trained on the training set and performed on the test set using pipelines, with a random state of 42. The decision tree had a depth of 18 and the minimum number of samples per leaf was set to 10. The random forest utilized 200 trees, square root feature subsampling, and a minimum leaf size of 2. The histogram-based gradient boosting algorithm was run for 180 iterations, learning rate 0.1, and maximum leaf nodes of 31.

3.3 Feature Ablation and Evaluation Metrics

For testing whether the performance was dependent on the compensation, four model families were also trained under two scenarios: (A) with Monthly_Salary feature, and (B) without the Monthly_Salary feature. In both cases, the division into train-test splits, the target variable, and the general preprocessing strategy remained the same. Accuracy, balanced accuracy, macro-F1, and weighted-F1 were calculated. While accuracy determines the proportion of correct classifications, macro-F1 treats the F1 score of all classes equally and thus should be used in case of unbalanced classes.

Feature ablation is the diagnosis procedure: if an ablation of a certain feature leads to a significant decrease in performance, then the feature could contain a lot of target-related information. That does not imply leakage by itself; in order to conclude on the presence of the leakage, one has to know at which point in time the features and the target were measured. As for the documentation of the dataset used in this experiment, it does not provide the needed information on the connection between salary and performance score.

3.4 Explainability and Responsible Use

The original analysis used feature importance based on decision tree impurity to detect features used in the fitted decision tree. This is not SHAP analysis and cannot be called such. SHAP global plot, class-specific SHAP, local SHAP, and SHAP explanation stability analysis are the right next steps following the selection of the right set of predictors. Explanation methods must be used on the final version of the model after leakage screening and assumptions and limitations of such explanations must be reported. Latest research on explainable HR prediction confirms the necessity of separation of global feature selection from local explanation and subgroup behavior assessment instead of single importance plot [8]. The dataset schema was confirmed using the provider's description [9].

4. Results and Discussion

Experimental results are shown to evaluate the predictive ability, interpretability, and reliability of the proposed framework for employee performance evaluation. The analysis starts with a performance score distribution and its connection with the monthly salary. The next step is a comparative analysis of four machine learning models according to the accuracy, balanced accuracy, and macro F1-score metrics. A feature ablation experiment is also performed by comparing performance of the models with and without the salary feature to evaluate its impact on predictions. Lastly, feature importance of a decision tree is analyzed to determine which features are the most influential for making predictions. It should be mentioned that the results are carefully interpreted as the data is artificial and the performance achieved by including salary can be a consequence of its dependence on other features.

4.1 Target Distribution

It is evident that the target is very well-balanced over five classes. Score 1 consists of 20,120 data points (20.12%), score 2 consists of 20,013 data points (20.01%), score 3 contains 19,999 data points (20.00%), score 4 contains 19,940 data points (19.94%), and finally, score 5 contains 19,928 data points (19.93%).

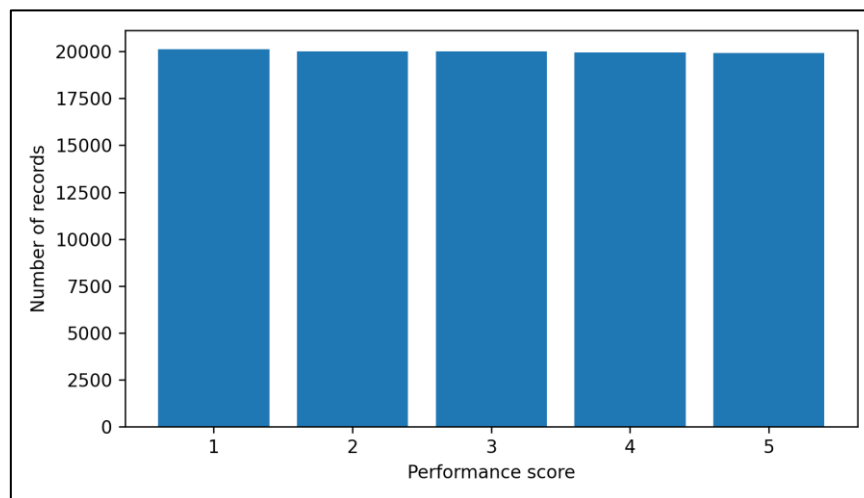


Figure 2. Distribution of the five employee performance score categories.

4.2 Salary and Performance Score

The mean monthly income was higher among all performance categories from 5,422.23 in score 1 to 7,397.67 in score 5. The means for scores 2, 3, and 4 were 5,900.28, 6,409.52, and 6,897.63, respectively. This ordered relation adds some background to the ablation results while proving neither the causal effect nor the appropriateness of salary in measuring employee performance.

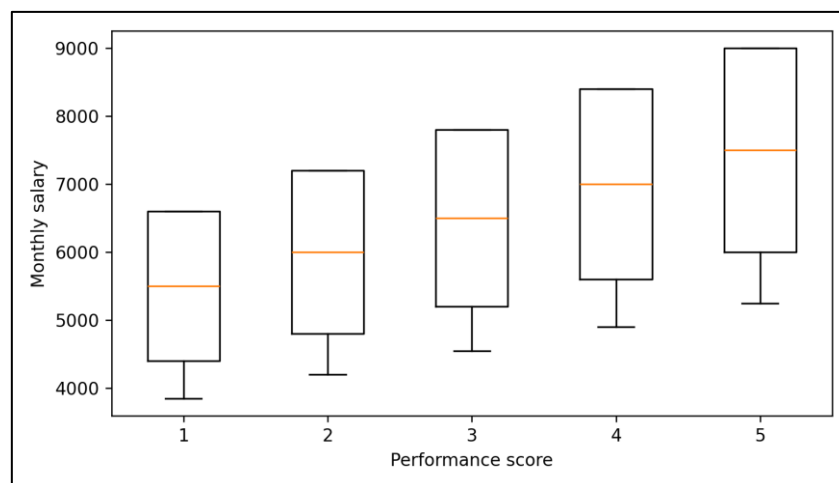


Figure 3. Monthly salary distributions grouped by observed performance score.

Figure 3 shows the correlation between monthly salary and employee performance scores. As is evident from Figure 3, the average monthly salary rises incrementally, from around 5,422 in Score 1 to 7,398 in Score 5, signifying a robust positive correlation between the two variables. It is likely that the two variables (salary and employee performance scores) are correlated and the former can have a great impact on predictive results. Nonetheless, the aforementioned correlation does not imply causation as the data used in the project is artificial.

Performance score	Records	Share	Mean monthly salary
1	20,120	20.12%	5,422.23
2	20,013	20.01%	5,900.28
3	19,999	20.00%	6,409.52
4	19,940	19.94%	6,897.63
5	19,928	19.93%	7,397.67

4.3 Comparative Classification Results

Multinomial logistic regression, random forest, and histogram-based gradient boosting all reached 100.00% test accuracy, balanced accuracy, and macro-F1 on the same level of precision when salary was included. The decision tree got 94.32% accuracy, 94.30% balanced accuracy, and 0.9429 macro-F1. These numbers look good at first glance, but they shouldn't be considered separately from the ablation experiment.

Without monthly salary, the logistic regression got 20.26% accuracy and 0.2000 macro-F1; the decision tree got 19.84% accuracy and 0.1983 macro-F1; the random forest got 20.15% accuracy and 0.2015 macro-F1; and the histogram-based gradient boosting got 19.88% accuracy and 0.1968 macro-F1. They are near the 20% chance level for the five balanced classes. It means that the rest of the attributes carried little information for prediction in this setting.

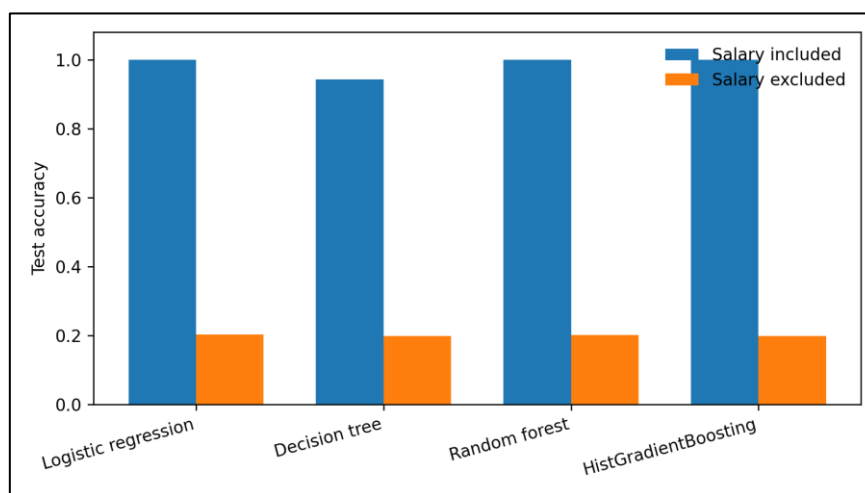


Figure 4. Test accuracy comparison with and without monthly salary.

Figure 4 depicts the comparison of predictive accuracies of the selected machine learning algorithms based on two types of experiments conducted: with the monthly salary as one of the features used and without this feature. In case if the monthly salary was present in the dataset, the accuracy of Logistic Regression, Random Forest and HistGradientBoosting reached 100%, whereas the accuracy of the Decision Tree model made about 94.32%. At the same time, the accuracy of all models drops down to about 20% in case of removal of the monthly salary feature, what corresponds to the expected accuracy of random guesswork in case of classification of five almost equally distributed classes.

Model	Accuracy with salary	Accuracy without salary	Macro-F1 without salary
Multinomial logistic regression	100.00%	20.26%	0.2000
Decision tree	94.32%	19.84%	0.1983
Random forest	100.00%	20.15%	0.2015
Histogram-based gradient boosting	100.00%	19.88%	0.1968

4.4 Feature Importance

In the fit decision tree which contains salary, the feature Monthly_Salary comprised about 76.36% of the total impurity-based feature importance. The top job title indicators were as follows: Technician (7.19%), Consultant (6.45%), Developer (3.84%), Engineer (3.15%) and Manager (3.01%). These figures represent the importance of the variables as utilized by the fit decision tree, and are not causal impacts and cannot be compared with SHAP values. The clear domination of salary can be justified by the drop in performance without salary.

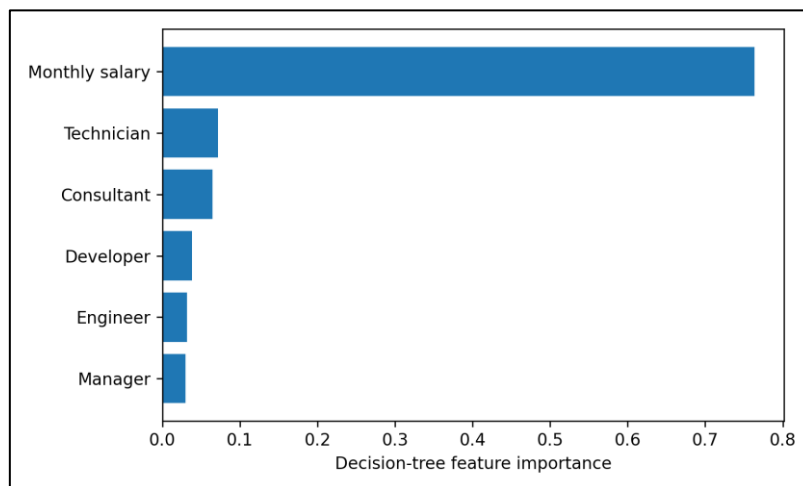


Figure 5. Leading decision-tree feature importances in the salary-included experiment.

The feature importance values generated from the Decision Tree classifier are illustrated in Figure 5. The monthly salary explains about 76.36% of the total impurity-based importance of the model, making it the most influential factor in predicting. On the other hand, the job title explains much lesser, with all other job titles adding up to smaller importance values. This highlights the fact that the model depends on the salary more than anything else, thus failing to learn a balanced set of employee features. While feature importance provides insights into the factors used by the model, they do not establish causality or determine if a particular feature is valid or relevant.

4.5 Interpretation and Implications

The most important result of an empirical nature is not the readiness of the suggested system for implementation, but the fragility of its predictive success. Classification close to perfect in the presence of salary and classification close to chance in the absence of salary show that the target in the dataset is likely encoded through salary or some other administratively related process. In case of salary being dependent on performance evaluations before, then it is wrong to make use of it to predict future performance.

This has direct implications on the explainable AI issue. SHAP can precisely explain how salary impacts the fitted model and still fail to show that salary is an appropriate criterion on which employees should be evaluated. An explanation is an answer to a model behavior question of what caused this outcome according to the fitted model but not an answer to a normative question regarding what criteria should affect performance. A responsible model should set up prediction time, drop all variables after that, compare against the naive benchmark, test for calibration and errors, check subgroup performance when available, and maintain human decision-making with appeals.

5. Conclusion and Future Scope

In this paper, an initial explainable-AI-oriented approach for employee performance classification was developed on a 100,000-record synthetic dataset. Comparison of four classifiers was conducted using a stratified holdout design. Feature ablation was conducted to investigate the dependence of classifier performance on monthly salary. Three out of four models achieved 100% accuracy in prediction with monthly salary present. In the case where it was not used as input, all classifiers performed at about 20% accuracy. Monthly salary is also the most important feature in the decision tree according to its importance calculation based on impurity.

This demonstrates the necessity of interpreting model performance along with feature validity and provenance of data. These findings neither prove that employee performance can be predicted based on other workforce features nor that the model is generalizable to any real organizations. Synthetic nature of data, lack of described target generation process, one-time holdout split, and uncompleted SHAP analysis and subgroup robustness analysis are significant limitations.

The next step should include the verification of the process through which the data were generated, as well as determining which features will be available at the moment of the appraisal decision-making. Then, the leakage-free dataset should be assessed using the repeated stratified cross-validation technique, including suitable baselines, confidence intervals, and independent validation data. Finally, SHAP explanations both globally and locally should be estimated only after proving the appropriateness of the selected predictive model. This should be done together with verifying the stability of SHAP explanations in different splits. Subgroup error analysis and human-centered validation procedures should be included when supported by the data. The purpose should be kept as decision-making support but not as automatic ranking or punishment.

References

1. J. Cecil, E. Lerner, M. F. C. Hudecek, J. Sauer, and S. Gaube, "Explainability does not mitigate the negative impact of incorrect AI advice in a personnel selection task," *Scientific Reports*, vol. 14, article 9736, 2024. doi: 10.1038/s41598-024-60220-5.
2. A. Heidemann, S. M. Hülter, and M. Tekieli, "Machine learning with real-world HR data: mitigating the trade-off between predictive performance and transparency," *The International Journal of Human Resource Management*, vol. 35, no. 14, pp. 2343–2366, 2024. doi: 10.1080/09585192.2024.2335515.
3. M. Lee, G. Lee, K. Lim, H. Moon, and J. Doh, "Machine learning-based causality analysis of human resource practices on firm performance," *Administrative Sciences*, vol. 14, no. 4, article 75, 2024. doi: 10.3390/admsci14040075.
4. S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, pp. 56–67, 2020. doi: 10.1038/s42256-019-0138-9.
5. A. Köchling and M. C. Wehner, "Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development," *Business Research*, vol. 13, pp. 795–848, 2020. doi: 10.1007/s40685-020-00134-w.
6. D. T. Newman, N. J. Fast, and D. J. Harmon, "When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions," *Organizational Behavior and Human Decision Processes*, vol. 160, pp. 149–167, 2020. doi: 10.1016/j.obhdp.2020.03.008.
7. Onyx Data, "Employee Performance and Productivity," *DataDNA Challenge Dataset*, October 2024. Dataset description: 100,000 rows and 20 columns.
8. R. Wagh and R. Kharat, "Explainable AI-Driven Employee Performance Prediction," *Adroid Conference Series: Engineering and Technology*, 2026. doi: 10.63503/acset.89.
9. mexwell, "Employee Performance and Productivity," *Kaggle dataset*, 2023. Dataset source as identified by the public dataset documentation. <https://www.kaggle.com/>.