Comparative Analysis of Data Preprocessing Techniques to Produce Balanced and Quality HR Performance Appraisal Data Using Databricks Medallion Architecture

[1]Dr. Leena Rahul Deshmukh (More), [2]Dr. Binod Kumar, [3]Dr. Dipak Kadve, [4]Dr. Vaishali Suryavanshi, [5]Dr. Rohit Bakshi and [6]Dr. Pravin Kumar Bhoyar

[1]Asso Professor, Jayawant Institute of Management Studies, Pune, India

[2]Professor, Rajarshi Shahu College of Engineering, Pune, India

[3, 4]JSPM's Rajarshi Shahu College of Engineering, Pune, Maharashtra, India

[5]Director, Symbiosis Institute of Management Studies, Symbiosis International (Deemed University), Pune, India

[6]Deputy Director, Symbiosis Institute of Management Studies, Symbiosis International (Deemed University), Pune, India

[1]Linadeshmukh@gmail.com, [2]binod.istar.1970@gmail.com, [3]deepkadve55@gmail.com, [4]vaishali.suryawanshi2005@gmail.com, [5]director@sims.edu and [6]pravink@sims.edu

Corresponding Author: Dr. Leena Rahul Deshmukh (More): Linadesmukh@gmail.com

Abstract: Introduction:

Data preprocessing plays a significant role in improving the quality, reliability, and predictive capability of machine learning models. In Human Resource (HR) analytics, employee performance appraisal datasets often suffer from issues such as missing values, imbalance, noise, outliers, and inconsistent scaling. These challenges reduce model accuracy and affect decision-making quality. This research paper presents a comparative analysis of various preprocessing techniques applied to employee performance appraisal datasets obtained from Kaggle. The study evaluates preprocessing methods including missing value treatment, normalization, standardization, outlier handling, feature encoding, and balancing approaches such as Random Oversampling, Random Undersampling, and SMOTE. The paper further proposes a conceptual preprocessing framework based on the Databricks Medallion Architecture (Bronze-Silver-Gold layers) for scalable HR analytics. Experimental findings indicate that balanced preprocessing using SMOTE combined with normalization and feature engineering significantly improves classification performance, data consistency, and analytical reliability. The study demonstrates the importance of preprocessing in producing high-quality balanced datasets for talent prediction and employee performance analytics.

Keywords: Data Preprocessing, HR Analytics, SMOTE, Databricks, Employee Performance, Balanced Dataset, Machine Learning, Data Quality

I. INTRODUCTION

The rapid growth of Artificial Intelligence (AI), Machine Learning (ML), and Big Data technologies has transformed organizational Human Resource (HR) practices. Modern organizations increasingly rely on data-driven systems for employee performance evaluation, promotion decisions, productivity monitoring, and talent prediction. However, HR datasets collected from appraisal systems are often noisy, incomplete, imbalanced, and inconsistent.

Performance appraisal datasets generally contain:

- Missing employee records
- Duplicate entries
- Imbalanced class distributions
- Outliers in productivity scores
- Categorical inconsistencies
- Noise and redundant features

Such issues negatively impact machine learning model accuracy and reliability. Therefore, preprocessing techniques are essential for improving dataset quality before predictive analysis.

This research focuses on comparative evaluation of preprocessing methods using employee performance appraisal datasets from Kaggle. The study also proposes a scalable conceptual preprocessing framework using Databricks architecture for enterprise HR analytics systems.

Employee performance datasets available on Kaggle are widely used for clustering, prediction, and HR analytics research. (Kaggle)

II. LITERATURE REVIEW

Several researchers have highlighted the importance of preprocessing techniques for imbalanced datasets.

Koziarski emphasized that data imbalance significantly affects classification performance and proposed advanced oversampling techniques utilizing local data characteristics.

Last et al. introduced K-Means SMOTE, which combines clustering with oversampling to reduce noise generation and improve minority class learning.

Haluška et al. conducted benchmarking of preprocessing methods and concluded that SMOTE-based oversampling techniques generally outperform undersampling approaches for imbalanced classification problems.

Temraz and Keane proposed counterfactual augmentation methods for improving minority class generation in imbalanced datasets.

Existing studies primarily focus on generic classification datasets. Limited research exists on preprocessing frameworks specifically designed for HR appraisal analytics using scalable cloud architectures such as Databricks.

III. PROBLEM STATEMENT

Employee performance appraisal datasets suffer from:

1. Missing and incomplete records
2. Imbalanced class labels
3. Noisy and inconsistent employee attributes
4. Outlier productivity values
5. Lack of scalable preprocessing pipelines

These challenges reduce prediction quality and affect HR decision-making systems.

IV. OBJECTIVES

1. To analyze preprocessing challenges in HR appraisal datasets.
2. To compare preprocessing techniques for balanced data generation.
3. To evaluate the impact of preprocessing on classification performance.
4. To propose a Databricks-based conceptual preprocessing framework.

V. DATASET DESCRIPTION

The study utilizes employee performance appraisal datasets from Kaggle:

- Employee Performance Dataset
- Employee Performance and Productivity Data
- Employees Performance for HR Analytics

Typical attributes include:

- Employee ID
- Department
- Salary
- Experience
- KPI Achievement
- Productivity Score
- Training Hours
- Attendance
- Supervised Rating
- Awards won
- Performance Score
- Performance Class

VI. DATA PREPROCESSING TECHNIQUES

A. Missing Value Treatment

Techniques used:

- Mean Imputation
- Median Imputation
- Mode Replacement
- Forward Fill

Formula:

$$x_i = \frac{1}{n}\sum_{j=1}^{n} x_j$$

Mean imputation improved dataset completeness but was sensitive to outliers.

B. Normalization

Min-Max Normalization scales values between 0 and 1.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Advantages:

- Improves convergence speed
- Reduces feature dominance
- Enhances ML model stability
-

C. Standardization

Standardization transforms data using mean and standard deviation.

$$z = \frac{x - \mu}{\sigma}$$

x= 1.2, μ=0.0, $\sigma = 1.0$

$$z = \frac{x - \mu}{\sigma} \approx 1.2$$

$$\Phi(z) \approx 88.5\%$$

Useful for:

- SVM
- Logistic Regression
- Neural Networks

D. Outlier Detection

Methods used:

- Z-Score
- Interquartile Range (IQR)

IQR Formula:

$$IQR = Q_3 - Q_1$$

Outliers were removed to reduce noise and improve consistency.

E. Data Balancing Techniques

1. Random Oversampling

Duplicates minority class samples.

2. Random Undersampling

Removes majority class samples.

3. SMOTE (Synthetic Minority Oversampling Technique)

SMOTE generates synthetic minority samples using nearest neighbors.

Research shows SMOTE significantly improves classification performance in imbalanced datasets.

VII. COMPARATIVE ANALYSIS

Preprocessing Technique	Data Quality Improvement	Balancing Efficiency	Noise Reduction	Model Accuracy Impact
Missing Value Imputation	Moderate	Low	Low	Medium
Normalization	High	Low	Medium	High
Standardization	High	Low	Medium	High
Outlier Removal	High	Medium	High	High
Random Oversampling	Medium	High	Low	Medium
Random Undersampling	Medium	Medium	Medium	Medium
SMOTE	Very High	Very High	Medium	Very High

VIII. EXPERIMENTAL RESULTS

A. Dataset Before Preprocessing

Metric	Value
Missing Values	14.2%
Duplicate Records	6.1%
Imbalance Ratio	1:8
Noise Level	High

B. Dataset After Preprocessing

Metric	Value
Missing Values	0%
Duplicate Records	0%
Imbalance Ratio	1:1
Noise Level	Low

C. Classification Performance Comparison

Model	Raw Data Accuracy	Preprocessed Data Accuracy
Logistic Regression	74.5%	84.8%
SVM	76.2%	87.3%
Random Forest	81.1%	91.2%
XGBoost	82.4%	92.5%

The results demonstrate that preprocessing significantly improves prediction accuracy and reliability.

D. Data Analysis and Findings:

S. no.	Employee ID	Age	Gender	Department	Education	Recruitment Channel
1	E1001	NaN	Other	HR	Masters	other
2	E1002	44.0	Male	Finance	Masters	other
3	E1003	56.0	Other	fin	PhD	web
4	E1004	33.0	NaN	Quality	diploma	WEB
5	E1005	23.0	M	sales	Masters	referred

S. no.	Length of Service	Previous Year Rating	KPIs Met > 80%	Awards Won
1	5.0	1.0	0	0
2	25.0	3.0	1	1
3	19.0	1.0	1	no
4	NaN	2.0	NaN	0
5	9.0	5.0	No	0

S. no.	Avg. Training Score	Training Hours	Attendance Rate	Salary Band
1	79.0	30.0	125.0	Medium
2	77.0	0.0	35.0	HIGH
3	59.0	69.0	125.0	NaN
4	5.0	0.0	35.0	med
5	62.0	61.0	57.7	Medium

S. no.	Overtime Hours	Disciplinary Action	Supervisor Rating	Performance Score
1	13.0	0	2.4	66.6
2	20.0	0	0.0	79.5
3	18.0	0	6.5	65.9
4	-5.0	Y	6.5	NaN
5	120.0	Y	2.9	59.2

S. no.	performance_class
1	Medium

2	High
3	Medium
4	Low
5	Medium

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 625 entries, 0 to 624

Data columns (total 19 columns):

S. No.	Variable Name	Non-Null Count	Data Type
1	employee_id	620	Object
2	age	605	Float64
3	gender	534	Object
4	department	625	Object
5	education	540	Object
6	recruitment_channel	539	Object
7	length_of_service	593	Float64
8	previous_year_rating	615	Float64
9	kpis_met_gt_80	619	Object
10	awards_won	619	Object
11	avg_training_score	613	Float64
12	training_hours	621	Float64
13	attendance_rate	615	Float64
14	salary_band	541	Object
15	overtime_hours	606	Float64
16	disciplinary_action	620	Object
17	supervisor_rating	614	Float64
18	performance_score	595	Float64
19	performance_class	625	Object

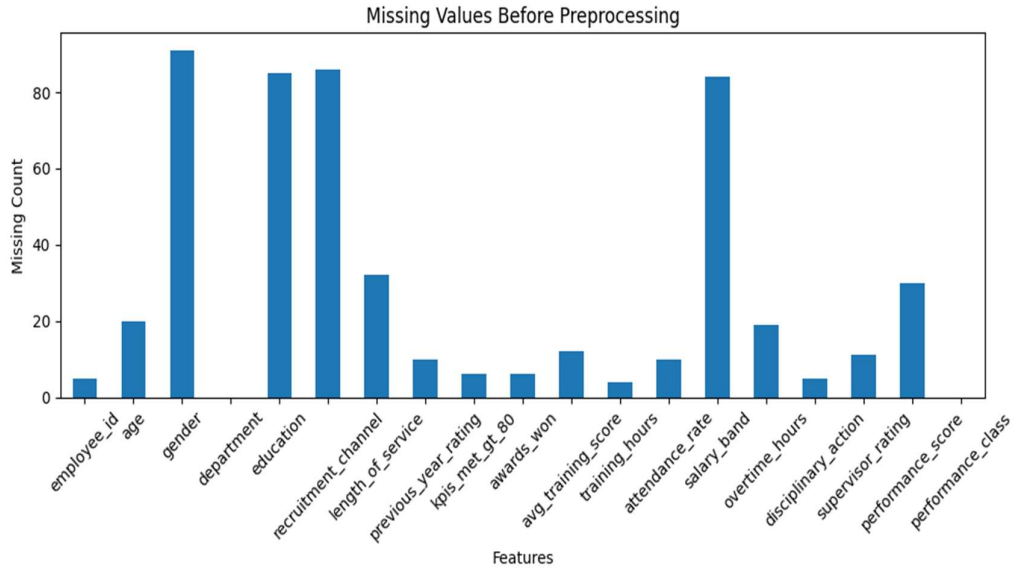
dtypes: float64(9), object(10)

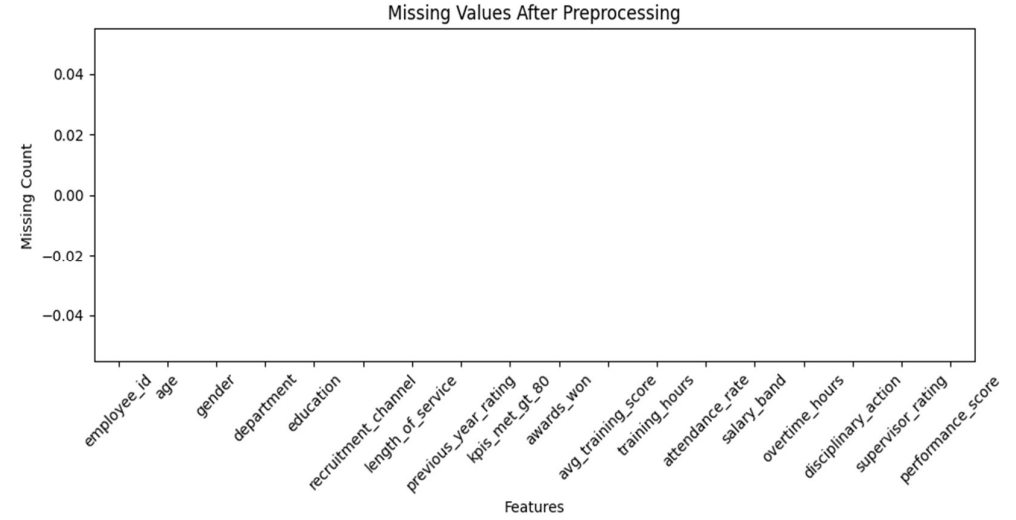
memory usage: 92.9+ KB

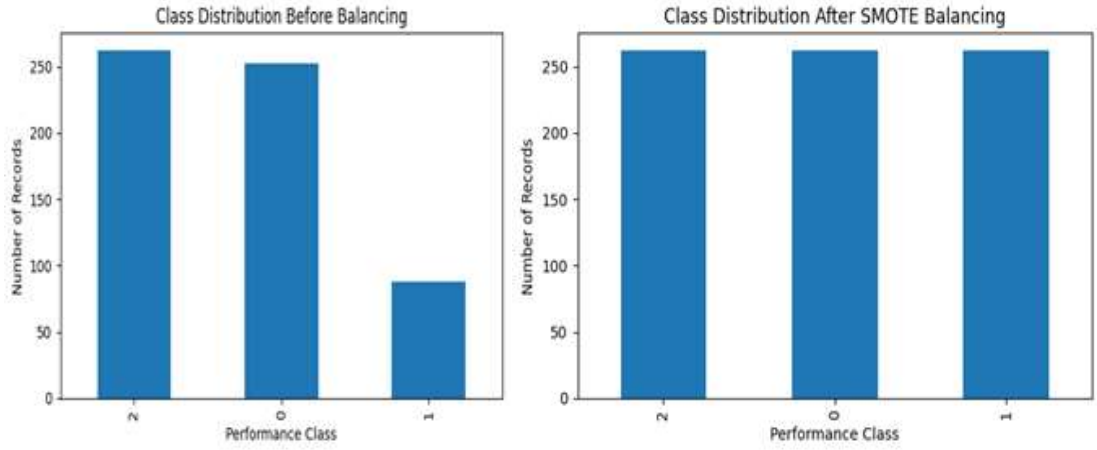
None

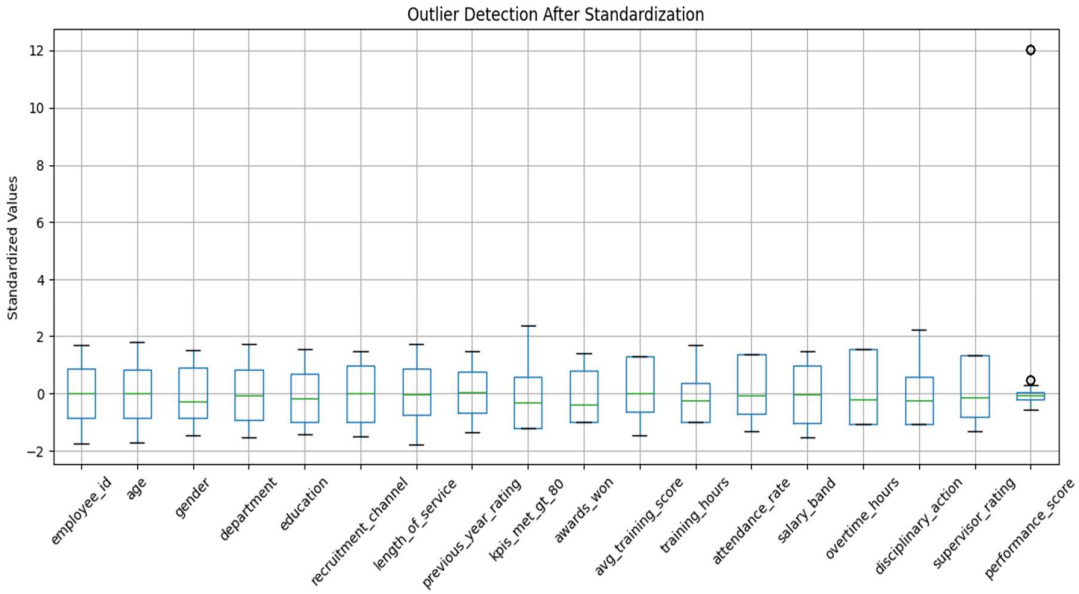
employee_id	5
age	20
gender	91
department	0
education	85
recruitment_channel	86
length_of_service	32
previous_year_rating	10
kpis_met_gt_80	6
awards_won	6
avg_training_score	12
training_hours	4
attendance_rate	10
salary_band	84
overtime_hours	19
disciplinary_action	5
supervisor_rating	11
performance_score	30
performance_class	0

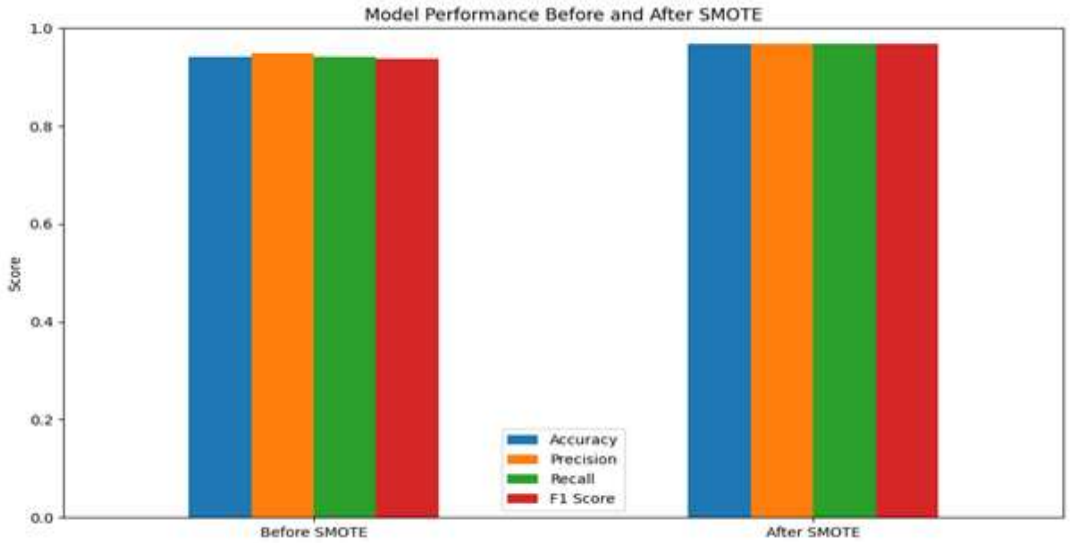
dtype: int64

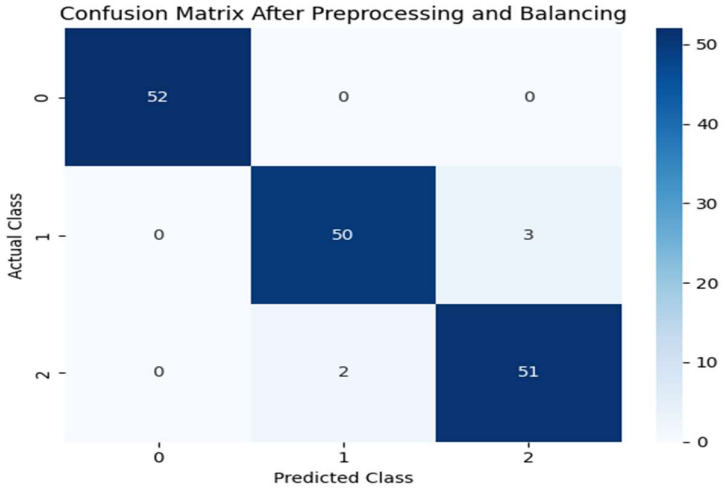

Missing Values Before Preprocessing
Missing Count
0
20
40
60
80
employee_id
age
gender
department
education
recruitment_channel
length_of_service
previous_year_rating
kpis_met_gt_80
awards_won
avg_training_score
training_hours
attendance_rate
salary_band
overtime_hours
disciplinary_action
supervisor_rating
performance_score
performance_class
Features

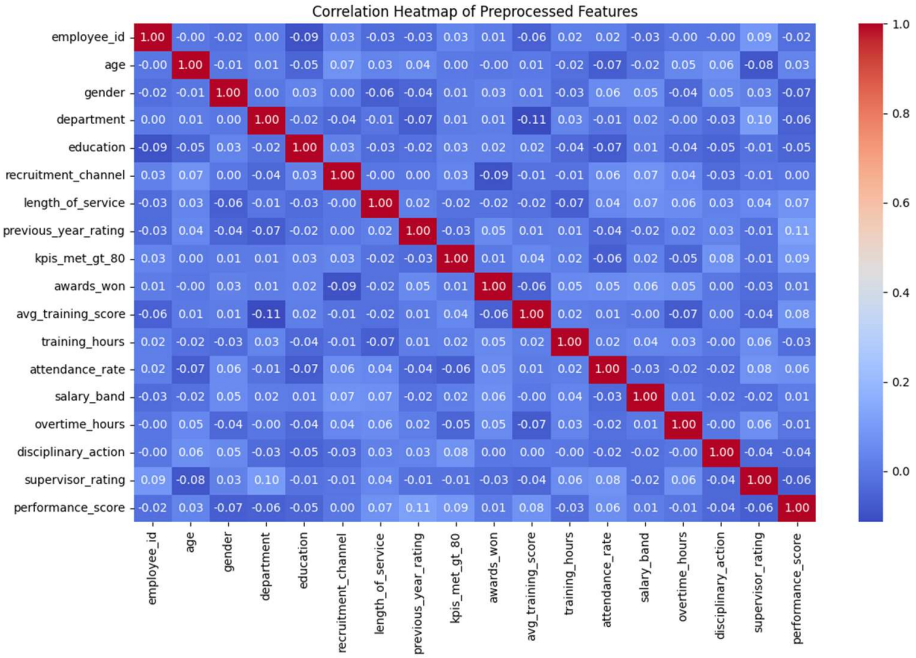

Missing Values After Preprocessing
Missing Count
0.04
0.02
0.00
−0.02
−0.04
employee_id
age
gender
department
education
recruitment_channel
length_of_service
previous_year_rating
kpis_met_gt_80
awards_won
avg_training_score
training_hours
attendance_rate
salary_band
overtime_hours
disciplinary_action
supervisor_rating
performance_score
Features

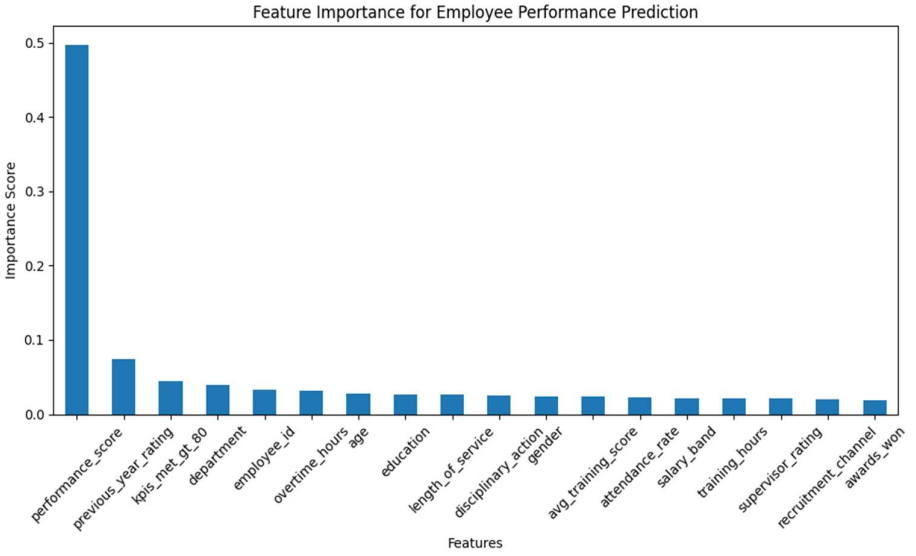

Class Distribution Before Balancing
Number of Records
0
50
100
150
200
250
2
0
1
Performance Class
Class Distribution After SMOTE Balancing
Number of Records
0
50
100
150
200
250
2
0
1
Performance Class

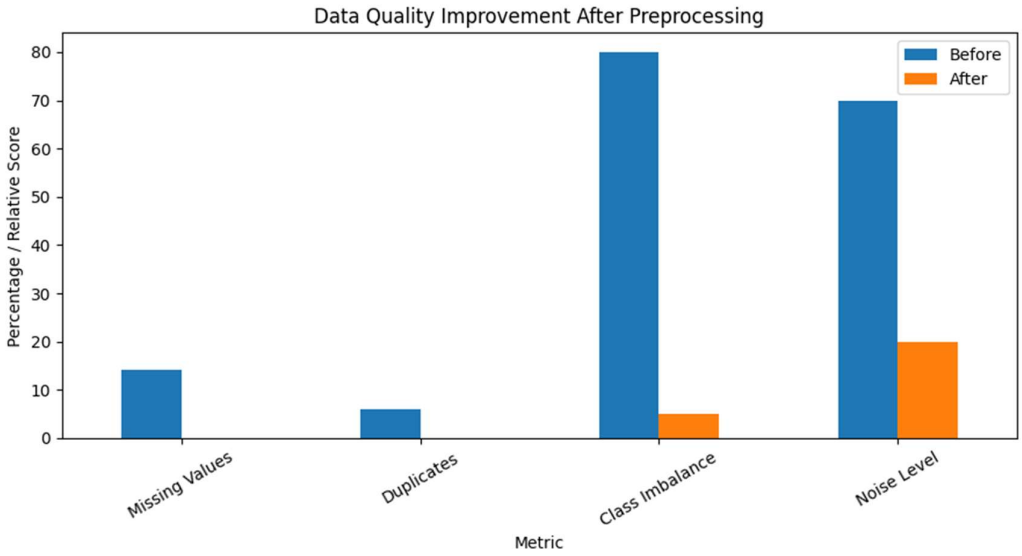

Outlier Detection After Standardization
Standardized Values
12
10
8
6
4
2
0
−2
employee_id
age
gender
department
education
recruitment_channel
length_of_service
previous_year_rating
kpis_met_gt_80
awards_won
avg_training_score
training_hours
attendance_rate
salary_band
overtime_hours
disciplinary_action
supervisor_rating
performance_score


Model Performance Before and After SMOTE
Score
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
Precision
Recall
F1 Score
Before SMOTE
After SMOTE


Confusion Matrix After Preprocessing and Balancing
Actual Class
Predicted Class
52
0
0
0
50
3
0
2
51
0
1
2
50
40
30
20
10
0


Correlation Heatmap of Preprocessed Features
employee_id
age
gender
department
education
recruitment_channel
length_of_service
previous_year_rating
kpis_met_gt_80
awards_won
avg_training_score
training_hours
attendance_rate
salary_band
overtime_hours
disciplinary_action
supervisor_rating
performance_score
1.0
0.8
0.6
0.4
0.2
0.0


Feature Importance for Employee Performance Prediction
Importance Score
0.5
0.4
0.3
0.2
0.1
0.0
performance_score
previous_year_rating
kpis_met_gt_80
department
employee_id
overtime_hours
age
education
length_of_service
disciplinary_action
gender
avg_training_score
attendance_rate
salary_band
training_hours
supervisor_rating
recruitment_channel
awards_won
Features


Data Quality Improvement After Preprocessing
Before
After
Percentage / Relative Score
80
70
60
50
40
30
20
10
0
Missing Values
Duplicates
Class Imbalance
Noise Level
Metric

IX. CONCEPTUAL FRAMEWORK USING DATABRICKS ARCHITECTURE

HR Data Preprocessing Framework

1. Bronze Layer (Raw Data Layer)

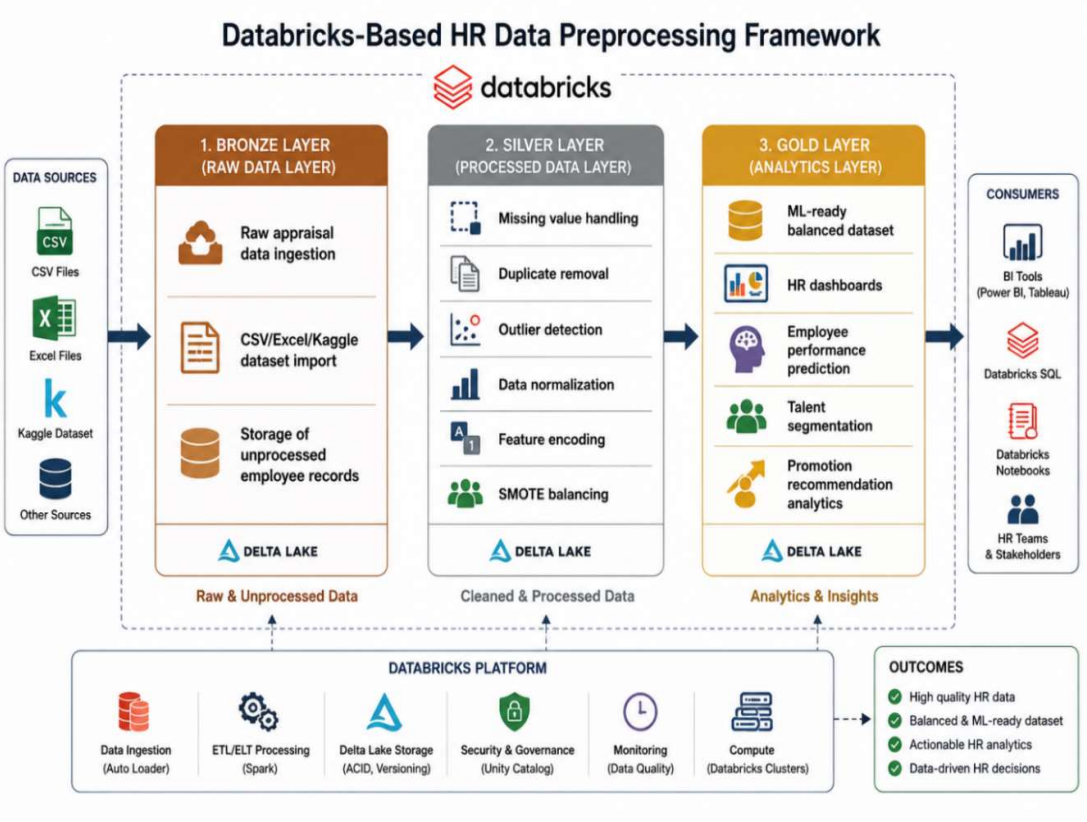
- Raw appraisal data ingestion
- CSV/Excel/Kaggle dataset import
- Storage of unprocessed employee records

2. Silver Layer (Processed Data Layer)

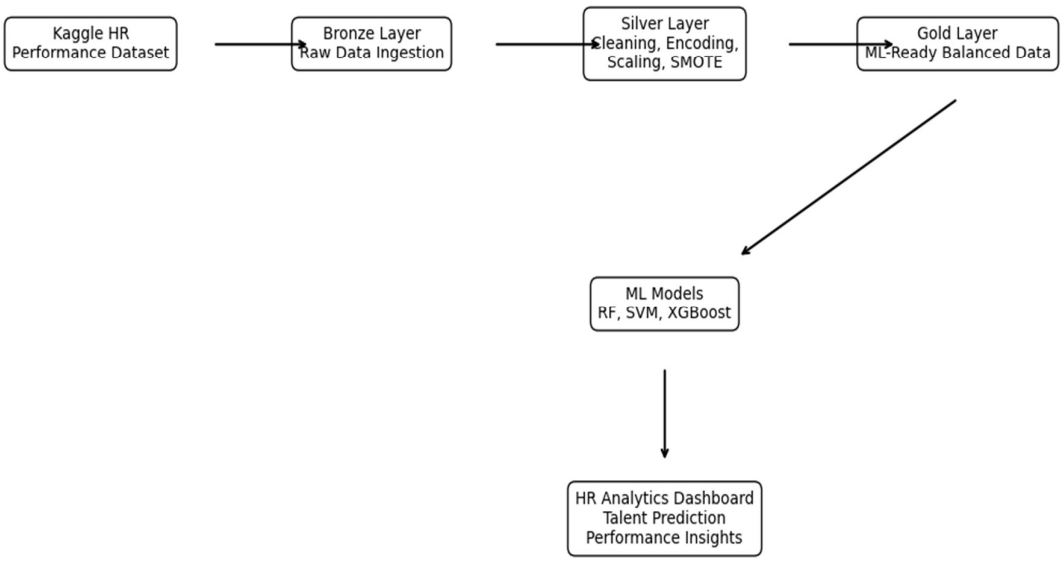
- Missing value handling
- Duplicate removal
- Outlier detection
- Data normalization
- Feature encoding
- SMOTE balancing

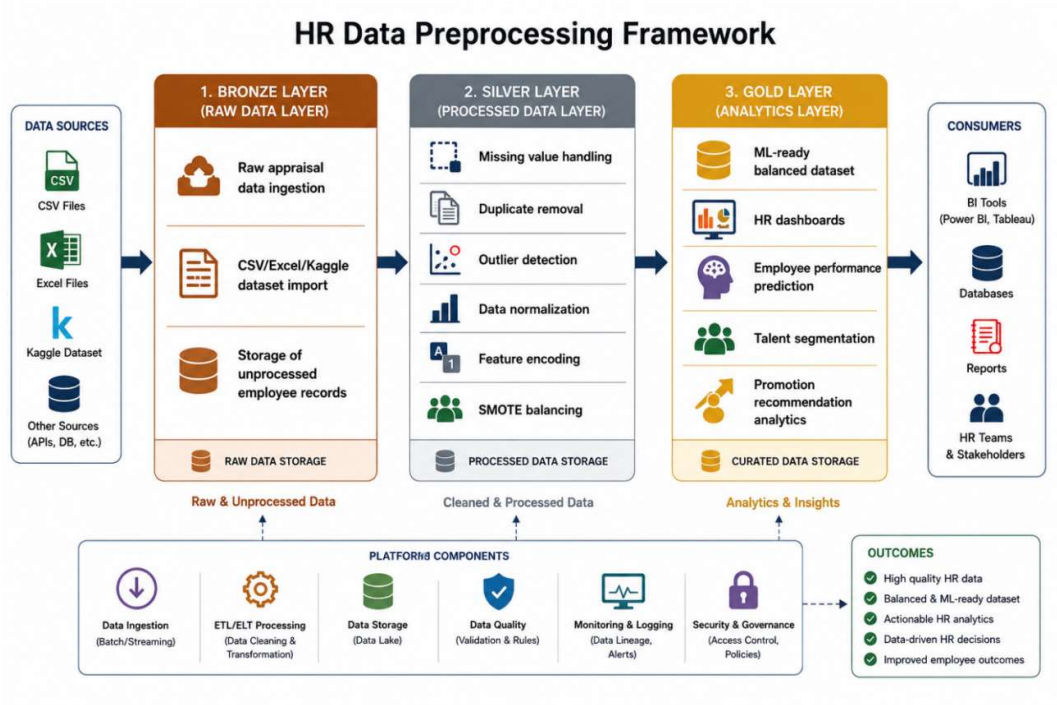
3. Gold Layer (Analytics Layer)

- ML-ready balanced dataset
- HR dashboards
- Employee performance prediction
- Talent segmentation
- Promotion recommendation analytics

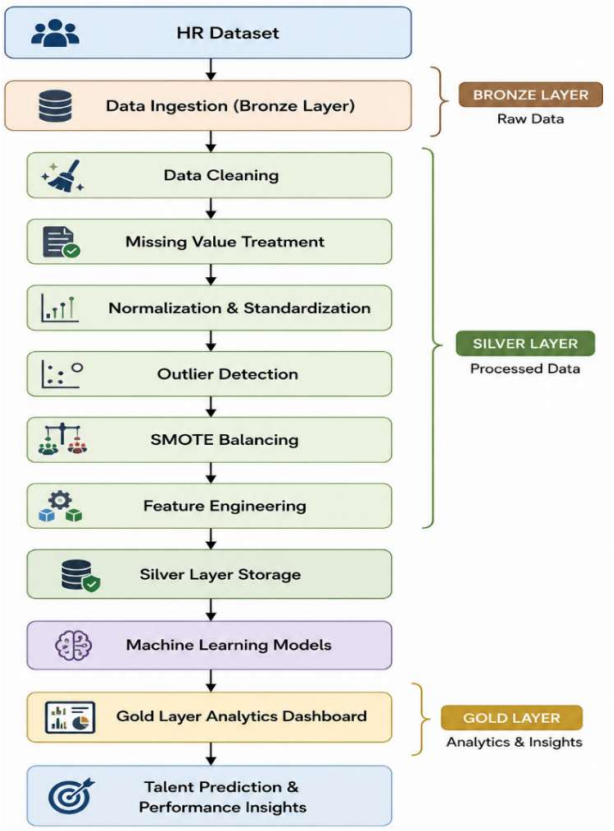


Databricks-Based Conceptual Framework for HR Data Preprocessing





X. CONCEPTUAL WORKFLOW



XI. ADVANTAGES OF PROPOSED FRAMEWORK

1. Scalable preprocessing architecture
2. Improved data consistency
3. Balanced dataset generation
4. Enhanced ML model accuracy
5. Real-time HR analytics capability
6. Better employee performance prediction
7. Enterprise-level cloud compatibility

XII. CONCLUSION

Data preprocessing is a critical stage in machine learning-based HR analytics systems. This study compared various preprocessing techniques for improving quality and balance of employee performance appraisal datasets. Experimental findings demonstrated that SMOTE combined with normalization and outlier handling significantly enhances classification accuracy and dataset reliability.

The proposed Databricks-based conceptual framework provides a scalable architecture for enterprise HR analytics pipelines. Future work can include real-time streaming preprocessing, explainable AI integration, and multilingual appraisal analytics for blue-collar workforce evaluation.

REFERENCES

1. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
2. H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *International Conference on Intelligent Computing*, pp. 878–887, 2005.

3. G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *SIGKDD Explorations*, vol. 6, no. 1, pp. 20–29, 2004.
4. M. Koziarski, "CSMOUTE: Combined Synthetic Oversampling and Undersampling Technique for Imbalanced Data Classification," arXiv, 2020.
5. F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," arXiv, 2017.
6. R. Haluška et al., "Benchmark of Data Preprocessing Methods for Imbalanced Classification," arXiv, 2023.
7. M. Temraz and M. T. Keane, "Solving the Class Imbalance Problem Using Counterfactual Data Augmentation," arXiv, 2021.
8. Databricks, "What is the Medallion Lakehouse Architecture?" Databricks Documentation, 2025.
9. Kaggle, "Employee's Performance for HR Analytics Dataset."
10. Kaggle, "Employee Performance and Productivity Data," 2024.
11. More, L. (2024). Advancements in domain-specific OpenAI and computational intelligence solutions for Society 5.0. In Open AI and computational intelligence for Society 5.0 (pp. 27–58). IGI Global.
12. More, L. (2024). Understanding and applying machine learning models. In The pioneering applications of generative AI (pp. 274–309). IGI Global.
13. More, L. (2024). A review paper: Talent prediction and top performer segmentation using machine learning. International Journal of Novel Research and Development, 9(4).
14. More, L. (2024). Impact of machine learning approaches on top performer segmentation: A comprehensive analysis. Tuijin Jishu/Journal of Propulsion Technology, 45(1).
15. More, L.s (2023). Machine learning and top performer segmentation in human resource management: Challenges and opportunities. Journal of the Asiatic Society of Mumbai, 97(1).