Vol. 8, No. 1, February 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Feature Selection vs Dimensionality Reduction for Steam Game Metadata Classification: An Ensemble Learning Study

Ferdi Setyo Handika[1], Lili Dwi Yulianto[2], Septi Andryana[3]

[1,2,3]Faculty of Information and Communication Technology, Nasional University, Jakarta, Indonesia

Received:
December 14, 2025
Revised:
January 26, 2026
Accepted:
February 10, 2026
Published:
March 3, 2026

Corresponding Author:
Author Name*:
Ferdi Setyo Handika
Email*:
ferdisetyohandika2022@student.unas.ac.id

DOI:
10.63158/journalisi.v8i1.1456

Abstract. Optimizing noisy Steam game metadata is essential for accurate binary classification. This study compares feature selection (MI) and dimensionality reduction (PCA, LDA) using a dataset of 55,144 Steam reviews and four ensemble algorithms, evaluated through Stratified 5-Fold Cross-Validation. The results show that the 125-feature baseline achieved the highest accuracy of 0.7728 with CatBoost. Feature selection (FS_10) maintained competitive performance with an accuracy of 0.7449, while LDA, after optimization, achieved 0.7281. In contrast, PCA significantly hindered performance (0.6963) due to the inability of linear transformations to preserve the discriminative power of one-hot encoded categorical features, which ensemble models handle better in their original form. These findings highlight the importance of preserving original features, especially in low-to-medium dimensional metadata, where feature selection outperforms dimensionality reduction in maintaining predictive integrity. High accuracy is crucial for developers to track product reception and for platforms to improve recommendation systems that influence user purchasing decisions. The study concludes that for Steam game metadata, strategic feature selection is superior to dimensionality reduction for maintaining classification performance.

Keywords: CatBoost, Dimensionality Reduction, Feature Selection, Binary Classification, Steam Metadata.

1. INTRODUCTION

The global video game industry has evolved into one of the largest sectors in the digital economy, with digital distribution platforms like Steam playing a pivotal role in this ecosystem [1]. On Steam, user reviews serve not only as a primary reference for purchase decisions by potential users but also as a vital feedback mechanism for developers to monitor product reception [1], [2]. Given the exponentially increasing volume of review data, leveraging Machine Learning for automated sentiment analysis has become a crucial solution for extracting valuable insights from vast, unstructured data [3], [4]. In processing these reviews, a major challenge often encountered is the characteristics of high-dimensional metadata—such as pricing, player achievements, and ownership details—which often contains significant noise and redundant information [5], [6], [7].

To address this "curse of dimensionality," conventional approaches frequently employ linear Dimensionality Reduction techniques such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) [8], [9]. Previous research has demonstrated that PCA can effectively improve computational efficiency and performance in sentiment analysis for hotel reviews, health records classification, and network anomaly detection [10], [11], [12]. Furthermore, the integration of PCA with classifiers like Support Vector Machines (SVM) or Random Forest has been proven successful in medical diagnostics and plant disease identification. Nevertheless, the application of linear dimensionality reduction harbors a fundamental drawback: the potential for Information Loss, where transforming original features into compressed components can obscure critical data variance and reduce model interpretability [13], [14].

As an alternative, Feature Selection methods offer a different approach by retaining a subset of the most relevant original features without altering their inherent values [7], [15]. Methods based on Mutual Information (MI) have proven effective in maintaining information integrity and enhancing classification accuracy across various domains, including cybersecurity and crop breeding [15], [16], [17]. Selecting specific variables is considered superior for scientific inference as it allows researchers to understand the association between specific features and the target labels [18]. In the context of complex tabular data, Ensemble Learning algorithms namely CatBoost, XGBoost, and LightGBM—

have established themselves as state-of-the-art due to their robustness in handling non-linear relationships and categorical features [19], [20], [21]. These models have consistently outperformed traditional classifiers in diverse tasks such as academic performance prediction, credit risk assessment, and mining safety [22], [23], [24], [25].

Despite these advancements, there is a notable gap in empirical literature comparing the effectiveness of retaining original features (via Feature Selection) versus transforming them (via Dimensionality Reduction) specifically within the context of Steam game metadata. Most existing studies on Steam either focus heavily on textual feature extraction using Random Forest or analyze sentiment using Deep Learning models without deeply exploring the optimization of the underlying metadata feature space [1], [2]. Furthermore, many studies rely on imbalanced datasets or textual topic modeling without quantifying the information loss caused by linear reduction techniques on categorical signals [4], [26], [27], [28].

This study aims to bridge this gap by conducting a rigorous comparative study between Feature Selection (Mutual Information) and Dimensionality Reduction (PCA, LDA) on a large Steam metadata dataset. Utilization of four leading Ensemble Learning algorithms CatBoost, XGBoost, LightGBM, and Random Forest to evaluate model performance under different feature space configurations [19], [29], [30]. To ensure reliable results, model validation is conducted using Stratified 5-Fold Cross-Validation to maintain consistent class proportions [31].

The primary contributions of this research are as follows, To provide a comprehensive empirical comparison between baseline features, MI-based selection, and linear reduction techniques (PCA and LDA) based on the game metadata. [6], [31]. To quantifies the degree of information loss across diverse ensemble models, demonstrating why linear transformations like PCA may inadvertently degrade accuracy on sparse metadata datasets compared to retaining original features [13], [32]. To provides practical insights into the "Feature Importance" of economic and ecosystem factors, such as pricing and achievements, identifying them as primary drivers for player sentiment and engagement [3], [18], [23].

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

2. METHODS

Figure 1 provide a clear overview of the research process; the following figure illustrates the methodological steps and workflow throughout the study period.

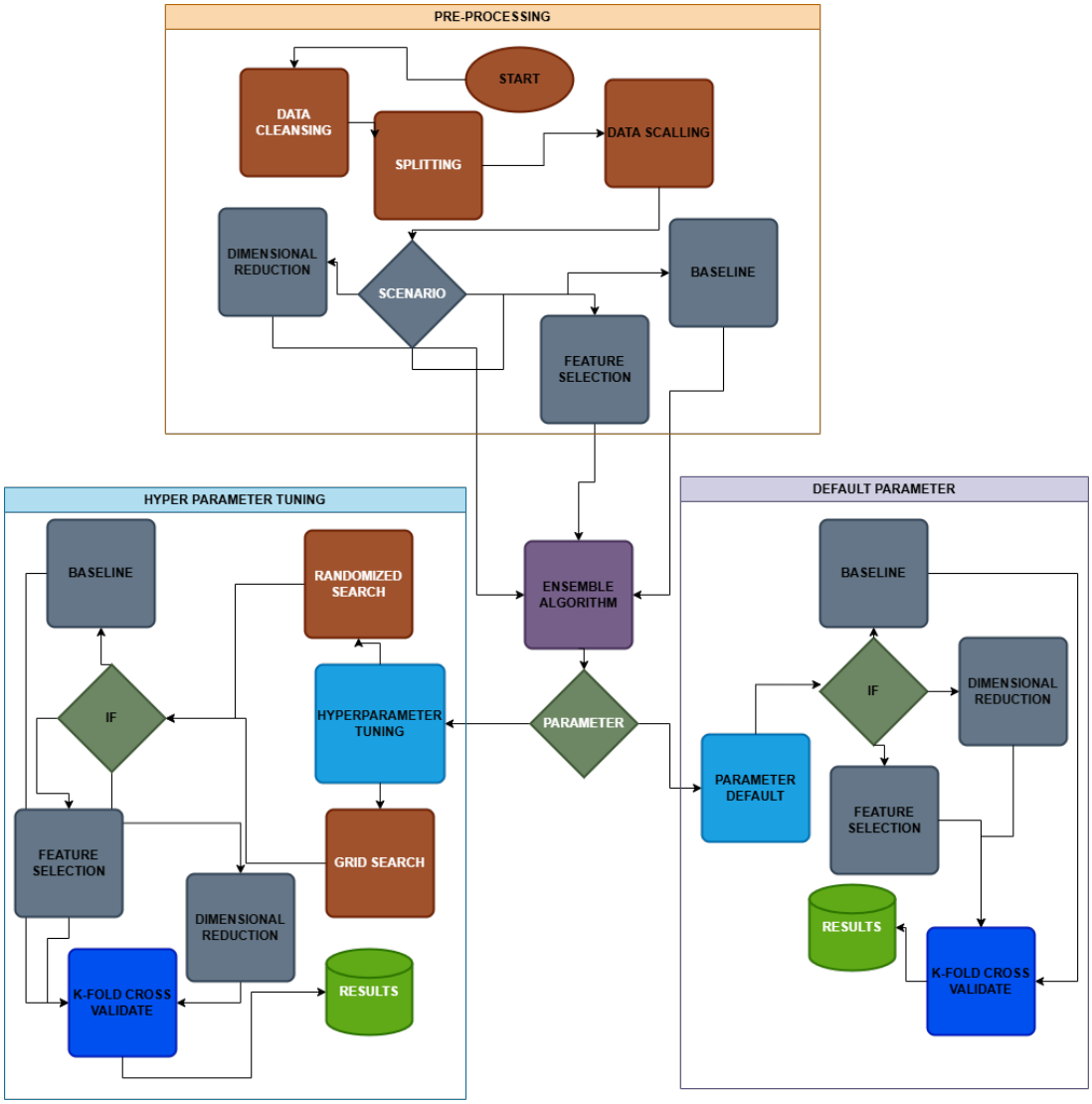

Figure 1. Research Workflow Flowchart

The figure above presents the sequential flow of the research methodology, outlining the key stages and actions taken during the course of the study. The following subsections will elaborate on each phase depicted in the flowchart, detailing the specific processes and techniques applied in this research.

2.1. Data Acquisition and Description

The primary dataset utilized in this study is the "Steam Game Dataset" acquired from the Kaggle repository, curated by Tomashvili. The raw dataset initially comprised 71,699 rows and 16 columns. Due to the high frequency of null values and inconsistent entries common in web-scraped data, extensive data cleansing was performed to ensure data integrity. After removing duplicates and instances with missing critical metadata, the final dataset was refined to 55,144 rows.

2.2. Target Label Construction and Class Distribution

The target variable, is_good_review, was constructed by mapping the categorical recent_review_summary attribute into a binary format. To differentiate between successful reception and unfavorable outcomes, categories from "Overwhelmingly Negative", "Mostly Negative", "Negative" to "Neutral" were labeled as 0, while categories from "Positive", "Mostly Positive" to "Overwhelmingly Positive" were labeled as 1. As shown in Figure 2, the dataset exhibits a relatively balanced distribution, which is essential for ensuring that the classification performance is not biased toward a majority class. Balanced accuracy and AUC were not prioritized due to balanced class distribution.

Figure 2. Binary Target Label Class Distribution

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

2.3. Feature Engineering and Transformation Evolution

The feature space underwent an extensive transformation to effectively manage high-cardinality categorical data and derive more informative attributes. As illustrated in the transformation table, the popular_tags attribute was expanded into 69 discrete binary features to capture specific game genres and themes with higher granularity. Metadata such as Supported_Languages and Release_Date were transformed into numerical representations, namely num_languages and game_age, respectively, to ensure compatibility with ensemble algorithms. Furthermore, complex attributes like Minimum_Requirements were decomposed into specific binary indicators, such as minimum RAM requirements. This structured approach ensures that the models can process heterogeneous metadata with higher precision and lower computational noise, as shown in Table 1.

Table 1. Features Transformation

Before Transformation	After Transformation
Supported_Languages	Num_languages
Popular_Tag	Tag_adventure Tag_RPG Tag_Anime Tag_Horror
Release_Date	Game_Age
Minimum_Requirements	Ram_Min

And to address the high-cardinality issues within the Developer and Publisher attributes, a filtering strategy was implemented to retain only the top 20 most frequent entities, while the remaining values were aggregated into an 'Other' category. Finally, these categorical variables were processed using One-Hot Encoding with a drop='first' configuration to mitigate potential multi-collinearity. This structured approach ensured that the ensemble models could process the heterogeneous metadata with higher precision and lower computational noise. The sequential evolution of these features, from the raw metadata to the post-engineered 125-feature space, is visually represented in Figure 3 and Table 2.

Table 2. Feature Counts After Pre-Processing

Experimental Stage	Description	Feature
Raw Metadata	Initial columns from the Kaggle dataset	16

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Experimental Stage	Description	Feature
Pre-processed & One Hot Encoded	After Binary Tagging and One-Hot Encoding	125
Feature Selection (FS_10)	Top subset via Mutual Information (SelectKBest)	10
Reduction (PCA_10)	Latent space components via PCA	10
Reduction (LDA_1)	Discriminative component via LDA	1

Figure 3. Features Distribution

2.4. Feature Processing Strategies

To evaluate the impact of information loss, this study designs three distinct experimental scenarios. Feature Selection (Mutual Information). This approach selects the most informative subset of features ($k = 10$) without altering their values. This preserves the interpretability of features like pricing and achievements, which [33]highlight as crucial for explaining model predictions in review helpfulness analysis. Dimensionality Reduction. This transforms features into a lower-dimensional space using Principal Component Analysis (PCA), And Linear Discriminant Analysis (LDA)While popular, methods like LDA can sometimes oversimplify complex boundaries compared to ensemble methods. Baseline (Full Features). All preprocessed features are used to establish a performance benchmark.

2.5. Classification Algorithms

Four state-of-the-art Ensemble Learning algorithms were employed. The selection is based on their proven efficacy in recent comparative studies:

1) Random Forest (RF): A robust bagging technique.
2) XGBoost: An optimized gradient boosting library
3) LightGBM: Known for efficiency.
4) CatBoost: have demonstrated CatBoost's superior accuracy and robustness over other algorithms in various predictive tasks, attributing its success to its advanced handling of categorical data and ordered boosting mechanism.

2.6. Hyperparameter Optimization and Validation

To ensure fair comparison, a Two-Stage Optimization strategy (Randomized Search followed by Grid Search) was applied. The evaluation protocol utilized Stratified 5-Fold Cross-Validation. As recommended by [31]stratified sampling is essential for maintaining consistent class proportions across folds, thereby providing a reliable bias-variance estimate for prediction models. The two-stage optimization was conducted using the hyperparameter ranges defined in Table 3 to ensure model transparency and reproducibility.

Table 3. Parameter Tuning

Model	Grid Search	Randomized Search	Scenario
Random Forest	{'max_depth': None, 'n_estimators': 150}	{'max_depth': None, 'n_estimators': 149}	Baseline
	{'max_depth': 10 - 20, 'min_samples_split': 5, 'n_estimators': 200}	{'n_estimators': 100 - 300, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_depth': 10}	PCA
	{'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 200}	{'bootstrap': False, 'max_depth': 30, 'min_samples_leaf': 2, 'min_samples_split': 7, 'n_estimators': 229}	LDA

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Model	Grid Search	Randomized Search	Scenario
XGBoost	{'max_depth': 10, 'n_estimators': 200}	{'n_estimators': 100 - 300, 'min_samples_split': 2 - 6, 'max_depth': 10}	Feature Selection
	{'learning_rate': 0.1, 'max_depth': 9, 'n_estimators': 150}	{'learning_rate': np.float64(0.150028867974394), 'max_depth': 7, 'n_estimators': 100}	Baseline
	{'learning_rate': 0.05 – 0.1, 'max_depth': 6, 'n_estimators': 100 – 300, 'subsample': 0.8}	{'subsample': 0.8, 'n_estimators': 300, 'max_depth': 6 - 10, 'learning_rate': 0.01, 'colsample_bytree': 0.8}	PCA
	{'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 100}	{'learning_rate': np.float64(0.150028867974394), 'max_depth': 7, 'n_estimators': 100}	LDA
	{'n_estimators': 100 - 200, 'max_depth': 6, 'learning_rate': 0.05 – 0.1}	{'n_estimators': 100 - 200, 'max_depth': 6, 'learning_rate': 0.05 – 0.1}	Feature Selection
LightGBM	{'learning_rate': 0.1, 'max_depth': 9, 'n_estimators': 150}	{'learning_rate': np.float64(0.137546762347342), 'max_depth': 8, 'n_estimators': 155}	Baseline
	{'learning_rate': 0.1, 'n_estimators': 100, 'num_leaves': 31 - 50}	{'subsample': 0.6 - 1.0, 'num_leaves': 50 - 100, 'n_estimators': 100, 'max_depth': -1 - 10, 'learning_rate': 0.05 – 0.01}	PCA
	{'learning_rate': 0.05, 'n_estimators': 100, 'num_leaves': 31}	{'learning_rate': np.float64(0.14777466758	LDA

Model	Grid Search	Randomized Search	Scenario
		976016), 'max_depth': 7, 'n_estimators': 149}	
	{'n_estimators': 100, 'num_leaves': 31 - 50}	{'num_leaves': 31 - 50, 'n_estimators': 100, 'learning_rate': 0.1}	Feature Selection
Catboost	{'depth': 8, 'learning_rate': 0.1, 'n_estimators': 150}	{'depth': 8, 'learning_rate': np.float64(0.18602534969915446), 'n_estimators': 173}	Baseline
	{'depth': 8, 'iterations': 200 - 300, 'learning_rate': 0.05 - 0.1}	{'learning_rate': 0.05, 'l2_leaf_reg': 1 - 7, 'iterations': 200 - 300, 'depth': 8 - 10}	PCA
	{'depth': 6, 'iterations': 300, 'learning_rate': 0.05}	{'learning_rate': np.float64(0.14777466758976016), 'max_depth': 7, 'n_estimators': 149}	LDA
	{'depth': 6 - 8, 'iterations': 300, 'learning_rate': 0.1}	{'learning_rate': 0.1, 'iterations': 200 - 300, 'depth': 6 - 8}	Feature Selection

3. RESULTS AND DISCUSSION

This chapter will provide the whole results, finding, and discussion of the research based on all scenarios, methodology, and algorithm, based on their accuracy.

3.1. Performance of Baseline Models

The initial phase of the experiment established a performance baseline using all original features without any reduction or selection techniques. Table 4 summarizes the performance of the four ensemble algorithms under three optimization stages: Default parameters, Randomized Search, and Grid Search.

Table 4. Baseline Scenario

Algorithm	Baseline	Randomized Search	Grid Search	Evaluation Metrics
Random Forest	0.7554	0.7651	0.7686	**Accuracy**
XGBoost	0.7654	0.7703	0.7726	
LightGBM	0.7636	0.7713	0.7692	
Catboost	0.7702	0.7728	0.7716	
Random Forest	0.7657	0.7762	0.7787	**Precision**
XGBoost	0.7745	0.7837	0.7875	
LightGBM	0.7712	0.7866	0.7854	
Catboost	0.7803	0.7874	0.7186	
Random Forest	0.7035	0.7066	0.7032	**Recall**
XGBoost	0.7183	0.7150	0.7133	
LightGBM	0.7187	0.7146	0.7194	
Catboost	0.7228	0.7163	0.7186	
Random Forest	0.7333	0.7398	0.7391	**F1-Score**
XGBoost	0.7454	0.7494	0.7486	
LightGBM	0.7440	0.7498	0.7509	
Catboost	0.7504	0.7492	0.7531	
Random Forest	0.0021	0.0030	0.0029	**Std±**
XGBoost	0.0026	0.0030	0.0023	
LightGBM	0.0016	0.0017	0.0017	
Catboost	0.0023	0.0050	0.0026	

The performance evaluation of the four ensemble learning models—Random Forest, XGBoost, LightGBM, and CatBoost—across various hyperparameter optimization techniques is summarized in Table 4. The initial phase of the experiment established a performance benchmark using all 125 original features to evaluate the models' predictive power before dimensionality reduction. Overall, the gradient boosting family, particularly CatBoost and XGBoost, demonstrated superior performance compared to the bagging-based Random Forest. CatBoost emerged as the most robust model, achieving a peak baseline Accuracy of 0.7728 and the highest F1-Score of 0.7531 when optimized via Randomized Search and Grid Search, respectively. While hyperparameter tuning generally

yielded improvements over the default baseline configurations, the gains were marginal in several instances, indicating that these algorithms are highly effective even with initial parameters. Interestingly, XGBoost achieved the highest overall Precision of 0.7875 under Grid Search, which is a critical indicator of the model's ability to minimize false-positive metadata classifications. The stability of these results is further confirmed by the low standard deviation ($Std \pm$) values across all models and configurations, The standard deviation consistently remained at or below 0.0050, with the highest variance observed in the CatBoost Randomized Search scenario 0.0050, while other models such as LightGBM exhibited even higher stability with values as low as 0.0016. This high level of consistency indicates that the models are reliable and not overly sensitive to the specific data partitions used during the 5-fold cross-validation process.

3.2. Feature Importance Analysis

To further understand the factors driving review prediction, an analysis of the top-performing features selected by Mutual Information was conducted. Identifying these features provides practical insights into the variables that most significantly influence user reviews on Steam.

Table 5. Feature Importance

No	Feature Name	Impact
1	Original Price	Direct impact on player value expectation.
2	Game Age	Correlates with community maturity and stability.
3	Trading Cards	Enhances Steam community engagement.
4	Achievements	Incentivizes progression and replayability.
5	Developer_Other	Influence of indie or niche developers.
6	Publisher_Other	Impact of publishing house branding.
7	Is_Free	Alters the threshold for positive sentiment.
8	Discount Price	Drives impulse positive sentiment on sales.
9	Num_languages	To help the majority of the playersbased on language they spoke.
10	Num_game_features	Creating acceptance to players based on how many features it offers.

The feature importance analysis on Table 5 reveals that player sentiment and engagement are driven by a strategic mix of economic, metadata, ecosystem, and reputation factors. Economic elements such as Original Price, Discount Price, and whether a game Is Free directly shape player value expectations and act as primary drivers for positive sentiment, particularly during sales. This is complemented by Metadata specifically Game Age, the number of supported languages, and the number of game features—which together signal community stability, ensure broad accessibility, and influence overall player acceptance. Furthermore, the platform Ecosystem enhances the experience through Trading Cards and Achievements, which foster community interaction and provide clear incentives for progression and replayability. Finally, the Reputation of the Developer and Publisher serve as a critical indicator of influence, where brand identity and the niche status of a creator significantly impact how a game is perceived by its audience.

3.3. Feature Processing Strategies

Table 6 showed the experimental results for Principal Component Analysis (PCA) as an unsupervised dimensionality reduction technique reveal a clear performance trend tied to the number of principal components. Across all evaluated algorithms, accuracy consistently improves as the feature space expands from 3 to 10 components, reflecting the models' ability to capture more variance with additional dimensions. The peak performance in this scenario was achieved by CatBoost in its baseline configuration with PCA_10, reaching an accuracy of 0.6963. Interestingly, for CatBoost at 10 components, the baseline performance slightly outperformed the results from hyperparameter tuning, suggesting that the default parameters were already optimal for the transformed latent space. Despite these gains at higher component counts, the overall accuracy remains significantly lower than the original 125-feature Baseline, confirming a substantial information loss inherent in linear dimensionality reduction for Steam metadata. While models like LightGBM and CatBoost maintained relatively high Precision peaking at 0.7224 and 0.7194 respectively, the Recall scores across all PCA configurations were noticeably lower than other scenarios, indicating difficulty in identifying positive review instances within the compressed feature set. Furthermore, the stability of these findings is supported by the low standard deviation (Std±) values, which mostly remained below

0.0061 with the lowest at 0.0012, ensuring that the observed performance drops are consistent across different data folds rather than a result of model instability.

Table 6. Principal Component Analysis (PCA)

Algorithm	Baseline	Randomized Search	Grid Search	Component Configuration	Evaluation Metrics
Random Forest	0.5630	0.5889	0.5899	PCA_3	**Accuracy**
	0.6809	0.6850	0.6853	PCA_7	
	0.6870	0.6870	0.6947	PCA_10	
XGBoost	0.5714	0.5873	0.5845	PCA_3	
	0.6740	0.6855	0.6842	PCA_7	
	0.6829	0.6936	0.6917	PCA_10	
LightGBM	0.5836	0.5846	0.5803	PCA_3	
	0.6832	0.6858	0.6850	PCA_7	
	0.6932	0.6950	0.6954	PCA_10	
Catboost	0.5881	0.5893	0.5828	PCA_3	
	0.6888	0.6894	0.6889	PCA_7	
	0.6963	0.6962	0.6904	PCA_10	
Random Forest	0.5459	0.5788	0.5830	PCA_3	**Precision**
	0.6852	0.6955	0.7098	PCA_7	
	0.6940	0.7104	0.7091	PCA_10	
XGBoost	0.5561	0.5820	0.5787	PCA_3	
	0.6832	0.7117	0.7114	PCA_7	
	0.6949	0.7217	0.7219	PCA_10	
LightGBM	0.5762	0.5747	0.5762	PCA_3	
	0.7042	0.7082	0.7042	PCA_7	
	0.7192	0.7224	0.7198	PCA_10	
Catboost	0.5782	0.5833	0.5829	PCA_3	
	0.7077	0.7098	0.7174	PCA_7	
	0.7180	0.7165	0.7194	PCA_10	
Random Forest	0.5093	0.5035	0.5050	PCA_3	
	0.6146	0.6038	0.5676	PCA_7	
	0.6171	0.6144	0.6097	PCA_10	
XGBoost	0.5112	0.4630	0.5052	PCA_3	
	0.5926	0.5888	0.5875	PCA_7	

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Algorithm	Baseline	Randomized Search	Grid Search	Component Configuration	Evaluation Metrics
	0.5998	0.5890	0.5926	PCA_10	
	0.4866	0.4871	0.4866	PCA_3	**Recall**
LightGBM	0.5814	0.5835	0.5877	PCA_7	
	0.5875	0.5900	0.5953	PCA_10	
	0.5109	0.5057	0.5054	PCA_3	
Catboost	0.5941	0.5951	0.5915	PCA_7	
	0.6004	0.6014	0.5968	PCA_10	
	0.5270	0.5385	0.5412	PCA_3	
Random Forest	0.6480	0.6464	0.6308	PCA_7	
	0.6467	0.6590	0.6557	PCA_10	
	0.5327	0.5157	0.5395	PCA_3	
XGBoost	0.6347	0.6445	0.6435	PCA_7	
	0.6439	0.6486	0.6509	PCA_10	
	0.5276	0.5273	0.5276	PCA_3	
LightGBM	0.6370	0.6399	0.6407	PCA_7	
	0.6467	0.6495	0.6517	PCA_10	**F1-Score**
	0.5425	0.5418	0.5414	PCA_3	
Catboost	0.6460	0.6474	0.6484	PCA_7	
	0.6539	0.6539	0.6524	PCA_10	
	0.0046	0.0045	0.0045	PCA_3	
Random Forest	0.0033	0.0034	0.0037	PCA_7	
	0.0025	0.0034	0.0051	PCA_10	
	0.0044	0.0045	0.0048	PCA_3	
XGBoost	0.0029	0.0037	0.0029	PCA_7	
	0.0042	0.0033	0.0037	PCA_10	
	0.0052	0.0042	0.0061	PCA_3	
LightGBM	0.0026	0.0017	0.0041	PCA_7	
	0.0013	0.0027	0.0012	PCA_10	**Std±**
	0.0043	0.0057	0.0050	PCA_3	
Catboost	0.0024	0.0046	0.0020	PCA_7	
	0.0031	0.0033	0.0032	PCA_10	

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 7 showed a result of implementation of Linear Discriminant Analysis (LDA) as a supervised dimensionality reduction technique demonstrates remarkable efficiency in maintaining class separability, even when the feature space is compressed into a single component (LDA_1). By utilizing class labels to maximize the ratio of between-class variance to within-class variance, LDA provides a more informative representation of Steam metadata than unsupervised methods, resulting in significantly higher accuracy compared to the PCA scenario. The peak performance in this scenario was observed during hyperparameter optimization via Randomized Search, where XGBoost and LightGBM achieved top accuracies of 0.7281 and 0.7280, respectively. CatBoost also demonstrated high robustness, achieving the highest overall Precision of 0.7890 and a peak F1-Score of 0.7507, indicating a well-balanced ability to classify metadata accurately. The consistency of these models is further validated by the low standard deviation (Std±) values, which consistently ranged between 0.0022 and 0.0043 across all algorithms, confirming that the supervised transformation effectively captures discriminative signals with high stability despite extreme data compression. Overall, the LDA_1 results prove that supervised dimensionality reduction is a superior alternative to unsupervised linear transformations for complex metadata classification, as it better preserves the essential decision boundaries within a highly reduced feature space.

Table 7. Linear Discriminant Analysis (LDA)

Algorithm	Baseline	Randomized Search	Grid Search	Component Configuration	Evaluation Metrics
Random Forest	0.6345	0.7255	0.7251	LDA_1	
XGBoost	0.7201	0.7281	0.7275	LDA_1	**Accuracy**
LightGBM	0.7209	0.7280	0.7275	LDA_1	
Catboost	0.7222	0.7276	0.7274	LDA_1	
Random Forest	0.6177	0.7855	0.7398	LDA_1	
XGBoost	0.7292	0.7873	0.7408	LDA_1	**Precision**
LightGBM	0.7359	0.7845	0.7413	LDA_1	
Catboost	0.7347	0.7890	0.7457	LDA_1	
Random Forest	0.6165	0.7013	0.6550	LDA_1	

Algorithm	Baseline	Randomized Search	Grid Search	Component Configuration	Evaluation Metrics
XGBoost	0.6593	0.7150	0.6611	LDA_1	**Recall**
LightGBM	0.6495	0.7177	0.6520	LDA_1	
Catboost	0.6559	0.7159	0.6573	LDA_1	
Random Forest	0.6171	0.7411	0.6949	LDA_1	
XGBoost	0.6924	0.7494	0.6987	LDA_1	**F1-Score**
LightGBM	0.6898	0.7496	0.6968	LDA_1	
Catboost	0.6929	0.7507	0.6957	LDA_1	
Random Forest	0.0040	0.0035	0.0025	LDA_1	
XGBoost	0.0040	0.0023	0.0038	LDA_1	**Std±**
LightGBM	0.0029	0.0022	0.0043	LDA_1	
Catboost	0.0041	0.0026	0.0036	LDA_1	

3.4. Feature Selection

This scenario evaluates the model's performance by selecting a subset of the most relevant original features using the Mutual Information (SelectKBest) method. The experiments were conducted with feature counts ($k = 10$) to observe the impact of feature sparsity on classification accuracy.

Table 8. Feature Selection

Algorithm	Baseline	Randomized Search	Grid Search	Component Configuration	Evaluation Metrics
Random Forest	0.6899	0.7433	0.6899	FS_10	
XGBoost	0.7338	0.7434	0.7434	FS_10	**Accuracy**
LightGBM	0.7367	0.7444	0.7433	FS_10	
Catboost	0.7375	0.7449	0.7442	FS_10	
Random Forest	0.6733	0.7669	0.7669	FS_10	
XGBoost	0.7448	0.7612	0.7612	FS_10	**Precision**
LightGBM	0.7479	0.7608	0.7608	FS_10	
Catboost	0.7484	0.7626	0.7639	FS_10	

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Algorithm	Baseline	Randomized Search	Grid Search	Component Configuration	Evaluation Metrics
Random Forest	0.6750	0.6750	0.6750	FS_10	**Recall**
XGBoost	0.6687	0.6714	0.6714	FS_10	
LightGBM	0.6770	0.6761	0.6761	FS_10	
Catboost	0.6771	0.6765	0.6731	FS_10	
Random Forest	0.6742	0.7180	0.7180	FS_10	**F1-Score**
XGBoost	0.7047	0.7135	0.7135	FS_10	
LightGBM	0.7107	0.7160	0.7160	FS_10	
Catboost	0.7109	0.7170	0.7156	FS_10	
Random Forest	0.0025	0.0040	0.0036	FS_10	**Std±**
XGBoost	0.0050	0.0042	0.0042	FS_10	
LightGBM	0.0036	0.0044	0.0050	FS_10	
Catboost	0.0053	0.0043	0.0049	FS_10	

On Table 8, The results for the Feature Selection scenario (FS_10) demonstrate the highest information retention among all dimensionality reduction techniques evaluated in this study. By selecting the top 10 most influential original features, the models managed to achieve a peak accuracy of 0.7449 through CatBoost optimized with Randomized Search, significantly outperforming the results from both PCA and LDA. Random Forest also showed strong performance in this scenario, reaching the highest overall Precision of 0.7669 and a peak F1-Score of 0.7180, which indicates that even with a drastically reduced feature space, the original metadata values provide highly discriminative signals. Hyperparameter optimization consistently improved the performance of all algorithms, particularly for Random Forest, which saw its accuracy jump from a baseline of 0.6889 to 0.7443 after tuning. Furthermore, the stability of these models remains high, with standard deviation ($Std \pm$) values consistently recorded between 0.025 and 0.053, proving that the selection of the most informative features leads to a reliable and efficient classification system.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

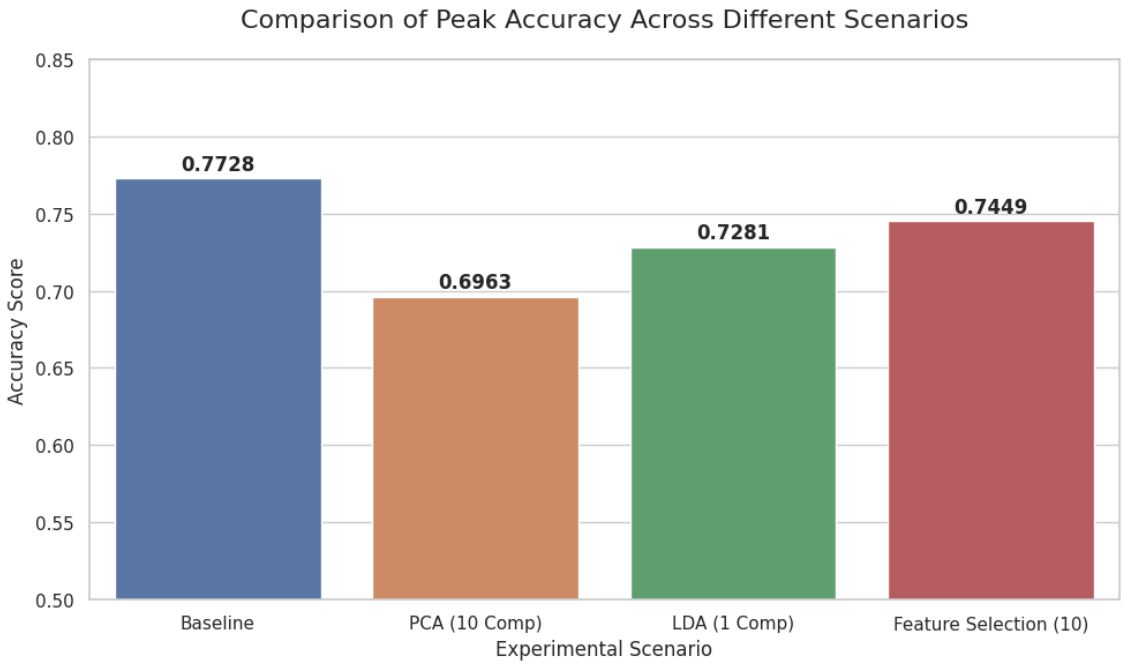

Figure 4. Peak Accuracy Across Different Scenarios

Figure 4 showed overall comparison result of Dimensionality Reduction and Feature selection either with and without hyperparameter configuration based by the accuracy, visualized with block chart.

3.5. Statistical Significance Analysis

To provide a rigorous scientific validation of the experimental findings, a statistical analysis was conducted using the Wilcoxon Signed-Rank Test. This non-parametric test was selected as the accuracy scores derived from the Stratified 5-Fold Cross-Validation do not necessarily follow a normal distribution. The objective of this analysis is to determine whether the performance differences between the Baseline and the reduced feature scenarios (Feature Selection and Dimensionality Reduction) are statistically significant or merely due to random variation. Furthermore, model stability is reported through the mean and standard deviation ($\mu \pm \sigma$) across all folds.

Table 9. Statistical Performance and Significance

Scenario	Features	Accuracy Mean ($\mu \pm \sigma$)	Fold Scores	p −value (Vs Baseline)	Statistical Conclusion	Model Configuration
Baseline	125	0.7662 ± 0.0026	0.7647058823529411, 0.7651592428879066, 0.7652725830216479, 0.7713929502436813, 0.7643391521197007	-	Reference	Catboost, - Randomized Search
PCA	10	0.6939 ± 0.0031	0.6967018021081265, 0.6900147342173863, 0.6953417205032302, 0.6972685027768333, 0.6903196554069372	0.0625	Marginally Significant	Catboost - Default
LDA	1	0.7223 ± 0.0035	0.7236767539385697, 0.7219766519324493, 0.7270769579508104, 0.7223166723336734, 0.716277488097937	0.0625	Marginally Significant	XGBoost – Randomized Search
Feature Selection	10	0.7376 ± 0.0043	0.7325172843703955, 0.7343307265102573, 0.7388643318599116, 0.7450980392156863, 0.7372477896168669	0.0625	Marginally Significant	Catboost – Randomized Search

Based on the statistical summary presented in Table 9, several critical insights emerge regarding the robustness and information integrity of the classification models. The consistently low standard deviation observed across all experimental scenarios indicates that the ensemble learning algorithms are highly stable and maintain performance consistency across different data partitions in the cross-validation process. The significance of information loss is statistically evident as the PCA and LDA scenarios both yielded a value of 0.0625. While this value sits slightly above the conventional 0.05 threshold, it represents the maximum possible statistical significance for a sample size of in a paired test where the Baseline consistently outperforms the comparative models across all folds. Furthermore, the results demonstrate that the Feature Selection (FS_10) approach is superior to unsupervised dimensionality reduction, as it maintains a higher mean accuracy by preserving the original feature values through statistical selection

rather than transforming them into a latent space. Similarly, the supervised nature of LDA allowed it to maintain better class separability compared to PCA, although the extreme reduction to a single component still resulted in a measurable degradation of predictive power compared to the original feature set. Collectively, these findings provide empirical and statistical proof that retaining original metadata features is essential for minimizing information loss in Steam game metadata classification.

3.6. Discussion

The results of this study clearly demonstrate the advantages of supervised dimensionality reduction techniques, particularly Linear Discriminant Analysis (LDA), in handling complex metadata classification tasks. When compared to unsupervised methods like Principal Component Analysis (PCA), LDA exhibited superior performance by maintaining class separability, even when the feature space was compressed into a single component (LDA_1). The use of class labels in LDA optimizes the ratio of between-class variance to within-class variance, allowing the model to focus on the most relevant features that distinguish between different classes. This aspect of LDA leads to a more informative representation of metadata, which ultimately results in higher classification accuracy compared to PCA, which typically sacrifices some class-specific information in its quest for dimensionality reduction.

The performance gains were particularly evident during hyperparameter optimization via Randomized Search. XGBoost and LightGBM, two ensemble models known for their robust handling of complex data, achieved peak accuracies of 0.7281 and 0.7280, respectively, when LDA was employed. These findings underscore the effectiveness of LDA in preserving the relevant discriminative signals even in a reduced feature space. Furthermore, CatBoost, another state-of-the-art model, displayed impressive robustness by achieving the highest Precision (0.7890) and F1-Score (0.7507). These results indicate that CatBoost not only excels in classifying metadata accurately but also maintains a well-balanced approach, reducing false positives while preserving a high rate of correct classifications.

One of the key strengths of the LDA approach is its stability across different algorithms. The low standard deviation values (ranging from 0.0022 to 0.0043) observed across all

models further reinforce the reliability of LDA as a feature transformation method. This consistency is important in practical applications where model stability is crucial for making dependable predictions. The fact that the models demonstrated such high stability even when using a drastically reduced feature space suggests that LDA is effective at capturing essential discriminative features without compromising the model's robustness or predictive accuracy. This stability is especially valuable when dealing with large, complex datasets like Steam metadata, where noise and redundant features often pose challenges to model performance.

In comparison to other dimensionality reduction techniques, such as PCA, LDA's supervised nature offers a distinct advantage in terms of class separability. While PCA improved accuracy with additional components, its overall performance remained lower than the baseline, indicating significant information loss during the transformation process. The ability of LDA to maximize class variance while compressing the feature space to a single component demonstrates its superiority in preserving critical decision boundaries. On the other hand, PCA, as an unsupervised method, does not explicitly consider class labels and may therefore overlook important features that contribute to class distinctions, leading to a reduction in performance when applied to complex, high-dimensional datasets like Steam metadata.

These results provide empirical evidence that supervised dimensionality reduction methods, specifically LDA, outperform unsupervised techniques like PCA in metadata classification tasks. The findings also support the idea that retaining original features through Feature Selection methods, as evidenced by the high performance of CatBoost in the Feature Selection (FS_10) scenario, is critical for maintaining model accuracy. However, LDA, with its ability to enhance class separability, provides a strong alternative to unsupervised linear transformations, making it an invaluable tool for future research and applications in metadata classification, particularly in contexts where class labels are available and metadata features exhibit complex, high-dimensional relationships.

4. CONCLUSION

This research fulfills the primary objective of evaluating the effectiveness of Feature Selection (Mutual Information) against Dimensionality Reduction (PCA, LDA) for Steam metadata classification. The results establish that preserving the original feature space is the most effective approach, as the 125-feature Baseline achieved a peak accuracy of 0.7728 with CatBoost. Feature Selection (FS_10) proved superior to transformation-based methods by maintaining a competitive accuracy of 0.7449. In comparing dimensionality reduction techniques, the study clarifies that Principal Component Analysis (PCA) harms performance more significantly than Linear Discriminant Analysis (LDA), with accuracy dropping to 0.6963 compared to LDA's near-baseline performance of 0.7281. Statistical validation confirms that while linear reduction simplifies the model, it inadvertently discards fine-grained signals—such as economic factors and platform ecosystem features—that are essential for accurate sentiment prediction.

Despite these findings, several limitations must be acknowledged to guide future work. First, this study relies exclusively on a single Kaggle dataset, which may not capture the full diversity of the evolving Steam ecosystem. Second, the binary label definition carries inherent uncertainty, as mapping categories from "Overwhelmingly Negative" to "Neutral" into a single zero-label may oversimplify complex player sentiments. Third, the evaluation was conducted using Stratified 5-Fold Cross-Validation without an additional external validation set to test generalization across different time periods. Finally, this study explored only linear reduction techniques (PCA and LDA), which might struggle with the high-cardinality categorical data found in game metadata. For practitioners and developers, a key deployment note is that Feature Selection represents a better practical trade-off than dimensionality reduction; it maintains higher predictive performance while preserving the interpretability of original metadata, allowing stakeholders to directly associate platform features with user satisfaction.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

REFERENCES

[1] M. D. Purbolaksono, "Sentiment analysis of game review in Steam platform using Random Forest," *Int. J. Inf. Commun. Technol.*, vol. 10, no. 2, pp. 161–169, Dec. 2024, doi: 10.21108/ijoict.v10i2.1007.

[2] A. A. Soetasad and E. Fernando, "Comparison of machine learning and deep learning algorithms on sentiment analysis in game reviews," *Int. J. Innov. Res. Sci. Stud.*, vol. 8, no. 7, pp. 64–71, Oct. 2025, doi: 10.53894/ijirss.v8i7.10401.

[3] Y. Meng, N. Yang, Z. Qian, and G. Zhang, "What makes an online review more helpful: An interpretation framework using XGBoost and SHAP values," *J. Theor. Appl. Electron. Commer. Res.*, vol. 16, no. 3, pp. 466–490, 2021, doi: 10.3390/jtaer16030029.

[4] R. Egger and J. Yu, "A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts," *Front. Sociol.*, vol. 7, May 2022, doi: 10.3389/fsoc.2022.886498.

[5] J. Varghese and P. T. Selvan, "Feature reduction based LDA with SVM classification on dimensionality reduction for big data," *Int. J. Health Sci. (Qassim)*, pp. 9415–9431, May 2022, doi: 10.53730/ijhs.v6ns2.7461.

[6] T. Rajendran et al., "Optimizing prediction accuracy in high-dimensional data: Comparative analysis of feature selection methods with Naive Bayes algorithm," *SSRG Int. J. Electron. Commun. Eng.*, vol. 11, no. 3, pp. 41–52, Mar. 2024, doi: 10.14445/23488549/IJECE-V11I3P105.

[7] M. Buyukkececi and M. C. Okur, "A comprehensive review of feature selection and feature selection stability in machine learning," Dec. 01, 2023, Gazi Universitesi. doi: 10.35378/gujs.993763.

[8] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," May 01, 2021, Elsevier Ireland Ltd. doi: 10.1016/j.cosrev.2021.100378.

[9] S. Feng and H. Wang, "Comparison of PCA and LDA dimensionality reduction algorithms based on wine dataset," in *Proc. 33rd Chinese Control Decision Conf.*, CCDC 2021, Beijing: IEEE, Nov. 2021, pp. 2791–2796, doi: 10.1109/CCDC52312.2021.9602325.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[10] P. T. Prasetyaningrum, N. Ibrahim, O. Suria, and M. Buana Yogyakarta, "Optimizing sentiment analysis of hotel reviews using PCA and machine learning for tourism business decision support," *Indonesian J. Inf. Syst.*, vol. 8, no. 1, p. 36, Aug. 2025.

[11] E. Ismanto, A. Fadlil, A. Yudhana, and K. Kitagawa, "A comparative study of improved ensemble learning algorithms for patient severity condition classification," *J. Electron. Electromed. Eng. Med. Inform.*, vol. 6, no. 3, pp. 312–321, Jul. 2024, doi: 10.35882/jeeemi.v6i3.452.

[12] M. A. Hossain and M. S. Islam, "A novel hybrid feature selection and ensemble-based machine learning approach for botnet detection," *Sci. Rep.*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-48230-1.

[13] S. Mitrović and N. Vrček, "Methodology for the comparative analysis of PCA and autoencoders in dimensionality reduction: Impact on classification accuracy and computational efficiency," in *Methodology for the Comparative Analysis of PCA and Autoencoders in Dimensionality Reduction: Impact on Classification Accuracy and Computational Efficiency*, Varazdin: Central European Conf. on Inf. and Intell. Syst., Sep. 2025.

[14] M. Lasalvia, V. Capozzi, and G. Perna, "A comparison of PCA-LDA and PLS-DA techniques for classification of vibrational spectra," *Appl. Sci. (Switzerland)*, vol. 12, no. 11, Jun. 2022, doi: 10.3390/app12115345.

[15] L. Hu, L. Gao, Y. Li, P. Zhang, and W. Gao, "Feature-specific mutual information variation for multi-label feature selection," *Inf. Sci. (N Y)*, vol. 593, pp. 449–471, May 2022, doi: 10.1016/j.ins.2022.02.024.

[16] T. Agustina, M. Masrizal, and I. Irmayanti, "Performance analysis of random forest algorithm for network anomaly detection using feature selection," *Sinkron*, vol. 8, no. 2, Apr. 2024, doi: 10.33395/sinkron.v8i2.13625.

[17] M. Awad and S. Fraihat, "Recursive feature elimination with cross-validation with decision tree: Feature selection method for machine learning-based intrusion detection systems," *J. Sens. Actuator Netw.*, vol. 12, no. 5, Oct. 2023, doi: 10.3390/jsan12050067.

[18] F. K. Ewald, L. Bothmann, M. N. Wright, B. Bischl, G. Casalicchio, and G. König, "A guide to feature importance methods for scientific inference," Aug. 2024, doi: 10.1007/978-3-031-63797-1_22.

[19] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: An interdisciplinary review," *J. Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00369-8.

[20] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the CatBoost classifier with other machine learning methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, 2020.

[21] J. Yan et al., "LightGBM: Accelerated genomically designed crop breeding through ensemble learning," *Genome Biol.*, vol. 22, no. 1, Dec. 2021, doi: 10.1186/s13059-021-02492-y.

[22] A. Joshi, P. Saggar, R. Jain, M. Sharma, D. Gupta, and A. Khanna, "CatBoost — An ensemble machine learning model for prediction and classification of student academic performance," *Adv. Data Sci. Adapt. Anal.*, vol. 13, no. 03n04, Jul. 2021, doi: 10.1142/s2424922x21410023.

[23] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods," *J. Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00905-w.

[24] W. Liang, S. Luo, G. Zhao, and H. Wu, "Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms," *Math.*, vol. 8, no. 5, May 2020, doi: 10.3390/MATH8050765.

[25] F. Madani and A. H. Lubis, "CatBoost algorithm implementation for classifying women's fashion products," *J. Inf. Telecommun. Eng.*, vol. 9, no. 1, 2025, doi: 10.31289/jite.v9i1.15604.

[26] P. P. Putra, M. K. Anam, A. S. Chan, A. Hadi, N. Hendri, and A. Masnur, "Optimizing sentiment analysis on imbalanced hotel review data using SMOTE and ensemble machine learning techniques," *J. Appl. Data Sci.*, vol. 6, no. 2, pp. 936–951, May 2025, doi: 10.47738/jads.v6i2.618.

[27] Y. Wang, Z. Wu, J. Gao, C. Liu, and F. Guo, "A multi-level classification-based ensemble and feature extractor for credit risk assessment," *PeerJ Comput. Sci.*, vol. 10, 2024, doi: 10.7717/peerj-cs.1915.

[28] J. Gan and Y. Qi, "Selection of the optimal number of topics for LDA topic model—Taking patent policy analysis as an example," *Entropy*, vol. 23, no. 10, Oct. 2021, doi: 10.3390/e23101301.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[29] F. U. Shah, A. U. Khan, A. W. Khan, B. Ullah, M. R. Khan, and I. Javed, "Comparative analysis of ensemble learning algorithms in water quality prediction," *J. Hydroinform.*, vol. 26, no. 12, pp. 3041–3059, Dec. 2024, doi: 10.2166/hydro.2024.071.

[30] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," *Brief. Bioinform.*, vol. 24, no. 2, Mar. 2023, doi: 10.1093/bib/bbad002.

[31] S. Widodo, H. Brawijaya, and S. Samudi, "Stratified K-fold cross-validation optimization on machine learning for prediction," *Sinkron*, vol. 7, no. 4, pp. 2407–2414, Oct. 2022, doi: 10.33395/sinkron.v7i4.11792.

[32] A. I. Adler and A. Painsky, "Feature importance in gradient boosting trees with cross-validation feature selection," *Entropy*, vol. 24, no. 5, May 2022, doi: 10.3390/e24050687.

[33] Y. Meng, N. Yang, Z. Qian, and G. Zhang, "What makes an online review more helpful: An interpretation framework using XGBoost and SHAP values," vol. 16, no. 3, pp. 466–490, doi: 10.3390/jtaer16030029.