

Evaluating PCA and LDA to Improve Machine Learning Classification of Consumer Behavior in Health Informatics

Denaya Ferrari Noval Agatra^{ID}, Dhendra Marutho^{*}^{ID}

Department of Informatics, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia

E-mail: denayaferrari@student.unimus.ac.id

*Corresponding author: dhendra@unimus.ac.id

Xinqi Xiong^{ID}

Department of Data Analytics, The Ohio State University, Columbus 43210, United States

E-mail: xiong.468@osu.edu

Article history :
Received: 23 DES 2024
Accepted: 21 MAR 2025
Available online: 31 MAR 2025

Research article

Abstract: Behavioral data derived from consumer activity offer significant potential in health informatics, particularly for developing predictive models related to lifestyle, adherence, and patient engagement. This study evaluates the effect of two widely used dimensionality reduction techniques, namely principal component analysis (PCA) and linear discriminant analysis (LDA) on the performance of five supervised machine learning classifiers: logistic regression (LR), support vector machine (SVM), naive bayes (NB), decision tree (DT), and random forest (RF). The experimental dataset, although sourced from a commercial context, contains demographic and economic attributes commonly found in health-related behavioral data, such as age and income. Results indicate that LDA significantly improves classification performance across all models, with Random Forest achieving the highest scores: accuracy = 0.91, precision = 0.88, recall = 0.85, F1-score = 0.86, and AUC = 0.95 when trained on LDA-transformed features. SVM also performed competitively under the same configuration (AUC = 0.94). Conversely, PCA provided moderate gains but underperformed in capturing class-discriminative information compared to LDA. These findings demonstrate that integrating LDA with robust classifiers such as RF and SVM enhances both predictive accuracy and model interpretability, offering practical benefits for behavior-informed health decision support systems. The study highlights the relevance of supervised feature transformation in optimizing data pipelines for personalized healthcare applications.

Keywords: CONSUMER BEHAVIOR; DIMENSIONALITY REDUCTION; LINEAR DISCRIMINANT ANALYSIS; HEALTH PREDICTION MODELS; ENSEMBLE LEARNING

Journal of Intelligent Computing and Health Informatics (JICHI) is licensed under a Creative Commons Attribution-Share Alike 4.0 International License



1. Introduction

The rapid expansion of digital health data has enabled the development of intelligent systems capable of delivering personalized healthcare, monitoring public health, and optimizing clinical workflows (Zhai et al., 2023). As these systems evolve, behavioral data have emerged as a valuable resource for supporting decision-making in health informatics. Consumer behavior, particularly purchase decisions, can reflect underlying health-related preferences, such as diet choices, physical activity, or engagement with wellness programs. Leveraging this behavioral information can enhance the effectiveness of recommendation systems, adherence monitoring, and preventive care planning (Samuel Ajibola Dada & Adeleke Damilola Adekola, 2024).

Machine learning (ML) techniques are central to modern health informatics. They enable predictive modeling, classification, and pattern recognition across diverse data sources. Algorithms such as logistic regression (LR), support vector machine (SVM), naive bayes (NB), decision tree (DT), and random forest (RF) are commonly used to analyze patient behavior, predict disease risk, and recommend personalized interventions. However, many health-related datasets especially those involving behavior are high-dimensional and have redundant or irrelevant features (Darmawahyuni et al., 2024; Nosakhare & Picard, 2020). These features can reduce the model performance and increase the computational complexity.

To address these issues, researchers applied dimensionality reduction techniques to extract the most

relevant features while minimizing information loss. Two widely used methods are principal component analysis (PCA) and linear discriminant analysis (LDA). PCA is an unsupervised method that transforms original features into a smaller set of orthogonal components representing the highest variance. LDA, in contrast, is a supervised technique that maximizes the separation between predefined classes. Both methods can improve model accuracy, reduce overfitting, and accelerate training time, key advantages in time-sensitive and resource-constrained health environments.

Several studies have demonstrated the effectiveness of PCA and LDA in biomedical and behavioral data analysis. For instance, [García et al. \(2024\)](#) combined PCA and LDA to classify plant leaf diseases, reducing complexity while maintaining high accuracy. Similarly, [Hayati et al. \(2024\)](#) used a PCA-LDA-SVM framework to determine the geographic origin of agricultural samples an approach applicable to biosignal classification. [Jiménez et al. \(2025\)](#) proposed PCA, LDA, and KPCA for analyzing fetal health data, improving interpretability and performance in clinical decision-making.

In behavioral modeling, ensemble algorithms such as Random Forest and Gradient Boosting often outperform traditional classifiers due to their ability to capture complex feature interactions. [Kabir et al. \(2019\)](#) achieved 92% accuracy in predicting online shopper intent using Random Forest. [Azad et al. \(2023\)](#) applied logistic regression to model consumer purchase behavior, reporting balanced performance with high interpretability an important consideration in health applications that require explainable models.

While prior studies have applied PCA and LDA in various contexts—including biomedical imaging, agriculture, and consumer analytics—comparative evaluations that systematically examine their impact on classification performance across multiple algorithms in health-relevant behavioral datasets remain limited. Moreover, most existing works tend to use high-dimensional or domain-specific data without reflecting on transferable behavioral features such as socioeconomic patterns relevant to personalized health interventions.

To fill this gap, the present study investigates how PCA and LDA affect the performance of five supervised classifiers when applied to a behavioral dataset that, despite its commercial origin, reflects key constructs commonly used in health informatics. By focusing on predictive outcomes in a low-dimensional yet representative feature space, we aim to provide methodological clarity and practical insight for behavior-based healthcare applications.

The main contributions of this study are as follows:

- Systematically evaluate the impact of PCA and LDA on five popular classification algorithms (LR, SVM, DT, NB, RF);
- Compare the performance across three feature representations (normalized, PCA, and LDA);
- Discuss the implications for behavioral modeling in personalized health systems, focusing on interpretability and scalability.

The remainder of this paper is structured as follows. Section 2 describes the dataset and preprocessing steps.

Section 3 outlines the methodology and classification models. Section 4 presents the experimental results and discussion. Finally, Section 5 concludes with insights and suggestions for future research in behavior-based health informatics.

2. Methodology

2.1 Dataset description

This study uses a publicly available consumer behavior dataset obtained from Kaggle, which can be accessed at <https://www.kaggle.com/denisadutca/customer-behaviour>. The dataset is frequently used for behavior modeling experiments and contains 400 observations, with each record representing a unique customer.

The dataset comprises the following variables:

- CustomerID: A unique identifier for each customer (nominal)
- Gender: Categorical variable representing sex (Male or Female)
- Age: Continuous variable indicating customer age in years
- EstimatedSalary: Continuous variable representing annual income in USD
- Purchased: Target label (binary class: 1 = Purchased, 0 = Not Purchased)

Although the dataset originates from a retail context, its structure closely reflects features commonly used in health informatics, such as age, socioeconomic indicators, and behavioral outcomes. These features are often applied in health behavior modeling, patient segmentation, and wellness program adoption prediction.

A snapshot of the dataset is presented in Table 1. As shown in Table 1, the structured format enables the predictive modeling of binary outcomes based on demographic and economic characteristics, analogous to health-related behavioral predictions such as medication adherence, lifestyle program enrollment, or patient outreach response.

To prepare the data for modeling, preprocessing steps were applied as described in sub-section 2.2, ensuring data quality and consistency for subsequent dimensionality reduction and classification tasks.

2.2 Data preprocessing

Data preprocessing is a critical phase in machine learning that transforms raw data into a structured and optimized format. Raw data often contain inconsistencies, categorical features, and numerical imbalances, which can negatively impact model performance. To address these issues, this study implements a structured preprocessing pipeline consisting of column filtering, missing value handling, categorical encoding, feature scaling, outlier detection, and data splitting. These steps help reduce noise, eliminate bias, and ensure that all features contribute proportionally to model learning. The dataset used in this study includes five attributes: User ID, Gender, Age, EstimatedSalary, and Purchased. A subset of the raw data is presented in Table 2.

Table 1. The table capture consumer behavior dataset at Kaggle.Com.

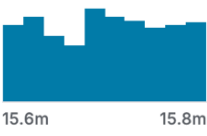

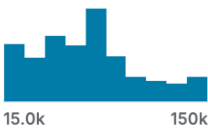

User ID	Gender	Age	EstimatedSalary	Purchased
the unique ID of each customer	Female/Male	integer value between 18 and 60 years old	integer values between 15K and 150K	0 - didn't purchase. 1 - purchased the product
	Female 51% Male 49%			
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0

Table 2. Sample of the row dataset before preprocessing.

CUSTOMER ID	GENDER	AGE	ESTIMATED SALARY (USD)	PURCHASED
15794253	Female	35	147000	1
15646936	Male	53	72000	1
15806901	Female	57	33000	1
15746422	Female	24	89000	0
15813113	Female	29	107000	1

The following preprocessing steps were applied:

a. Column selection

The attribute User ID was removed from the dataset as it does not contain any predictive information. Including such identifiers may introduce bias or data leakage if not properly managed. Only four features were retained for modeling: Gender, Age, EstimatedSalary, and Purchased.

b. Handling missing values

Let $X \in \mathbb{R}^{n \times d}$ be the dataset matrix with n rows (samples) and d columns (features). A missing value in feature x_j occurs when:

$$x_{ij} = \text{for any } i \in \{1, \dots, n\}, j \in \{1, \dots, d\} \quad (1)$$

We define the completeness ratio for feature j as:

$$C_j = \frac{\text{count}(x_{ij} \neq \emptyset)}{n} \quad (2)$$

In this dataset, $C_j = 1$ for all j , indicating that no missing values were present. Thus, no imputation or removal was required, as illustrated in in Fig. 1.

```
[5]: # check for duplicates
df.isnull().sum()

[5]: User ID      0
     Gender      0
     Age         0
     EstimatedSalary 0
     Purchased    0
     dtype: int64
```

Fig 1. Visual confirmation of data completeness across all features. No missing values were identified in any column, validating the integrity of the dataset for subsequent preprocessing.

c. Encoding of categorical features

In machine learning, most algorithms require all input features to be numeric, as mathematical operations such as distance computation, dot products, and gradient descent optimization are not defined for string-based or symbolic inputs. Therefore, categorical variables must be converted into a numerical format before they can be used in model training.

The feature Gender in the dataset is a binary categorical variable with two possible string values: "Male" and "Female". To prepare this feature for modelling, we applied binary encoding, a suitable method for variables with only two categories. Binary encoding assigns integer values to each category without introducing unnecessary dimensions, unlike one-hot encoding, which would create additional columns. The encoding rule applied was formulated by Eq. (3).

$$G_{\text{encoded}} = \begin{cases} 0 & \text{if Female} \\ 1 & \text{if Male} \end{cases} \quad (3)$$

This transformation allows the model to treat categorical input as numerical, enabling gradient-based optimization and matrix computations within classification algorithms.

d. Feature scaling

In this study, feature scaling was performed as a necessary preprocessing step to prepare the numerical variables for classification models. The original dataset included two continuous features, Age and Estimated Salary which exhibited significantly different numerical scales and variances. Specifically, while the Age variable

ranged from 18 to 60 years, the Estimated Salary ranged from 15,000 to 150,000 USD. Such discrepancies can lead to biased learning in algorithms sensitive to feature magnitude, such as SVM and LR, where larger-valued variables may dominate the learning objective.

To address this issue, we adopted Z-score normalization, also known as standard score transformation. This method was selected because it rescales features without compressing the variance structure, which is particularly important in behavioral and socioeconomic datasets where the spread of information is meaningful. Unlike Min-Max scaling, which maps values into a fixed range [0,1], Z-score normalization transforms values based on the statistical properties of the data specifically, the mean and standard deviation, preserving the overall distribution.

The mathematical definition of the Z-score transformation is given in Eq. (4):

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (4)$$

where x_{ij} is the original value of feature j for sample i , $\mu_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$ is mean of feature j , and $\sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \mu_j)^2}$ is a standard deviation of feature j . This transformation ensures that each feature has a mean of zero and a standard deviation of one, enabling fair and unbiased contribution across all features during model training.

Table 3 presents the descriptive statistics of the raw features before normalization. It is evident that the salary variable not only spans a wider numerical range but also exhibits a greater standard deviation than the age feature.

Table 3. Descriptive statistics of the raw features before normalization.

FEATURE	MIN	MAX	MEAN (μ)	STD. DEV (σ)
Age	18	60	38.85	10.49
EstimatedSalary	15000	150000	70,097.16	33,143.94

Note: These values indicate that EstimatedSalary has a significantly higher variance and magnitude than Age, justifying the need for standardization.

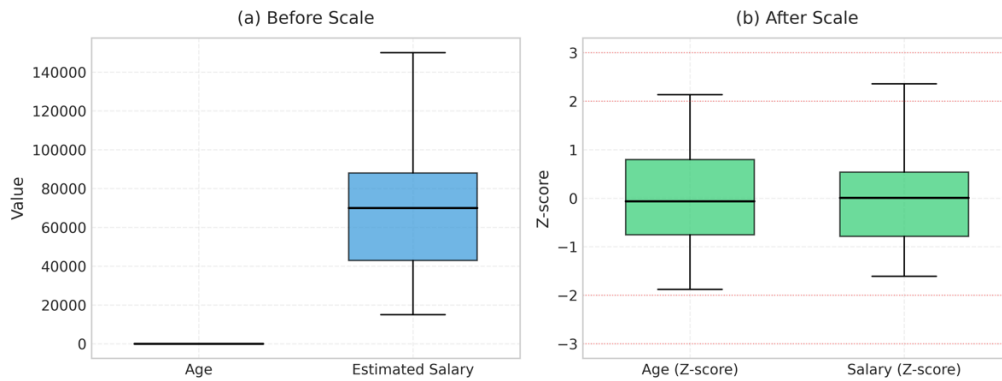


Fig 2. Box plots of age and estimated salary before (left) and after Z-score normalization (right). (Left) features are on disparate scales; (Right) transformed features show comparable distributions.

Following normalization, we observed a substantial transformation in the distribution of both features. To illustrate the effect of this process, Fig. 2 provides side-by-side box plots of Age and Estimated Salary before and after Z-score normalization. As shown in Fig. 2 (left), before scaling, Estimated Salary has an extensive interquartile range with extreme values, whereas Age is more concentrated. This imbalance is fully corrected in Fig. 2 (right), where both variables exhibit comparable distributions centered around zero with unit variance.

Table 4. Sample values before and after z-score normalization.

AGE (raw)	AGE (z-score)	SALARY (raw)	SALARY (z-score)
19	-1.89	19,000	-1.54
35	-0.37	20,000	-1.51
26	-1.23	43,000	-0.81
27	-1.14	57,000	-0.39
32	-0.65	150,000	2.41

(Dhendra Marutho)

To further interpret the transformation, Table 4 shows a few samples before and after normalization. The z-score values provide a standardized representation of each observation's deviation from the mean. For instance, a subject with Age = 19 and Estimated Salary = 19,000 yields z-scores of -1.89 and -1.54 respectively, indicating that both values lie significantly below the population average for each feature.

From a modelling perspective, this transformation plays a critical role in ensuring that the classifier does not disproportionately favor features with larger magnitudes. Furthermore, it improves numerical stability during optimization and accelerates convergence, particularly in algorithms involving gradient descent. The standardized feature space also enhances the performance of subsequent dimensionality reduction methods such as PCA and LDA, which rely on the geometric relationships between variables.

e. Outlier detection

Observations that deviate significantly from most of the dataset are considered outliers, posing a well-documented challenge in the development of robust machine learning models. In health informatics and behavioral classification tasks, the presence of extreme values in demographic or socioeconomic features may misrepresent population patterns, induce biased learning, and impair generalizability. Hence, rigorous outlier detection is a crucial part of the preprocessing pipeline, especially after the normalization procedures have been applied.

In this study, we employed Z-score-based outlier detection to identify anomalous values in the two standardized numerical features: Age and Estimated Salary. This choice was made for three primary reasons. First, it is methodologically consistent with the Z-score normalization previously applied to these features (as detailed in sub-section *c. feature scaling*). Second, the Z-score method assumes near-normality in the distribution of features, an assumption reasonably satisfied by the post-normalized data, particularly given the empirical behavior of age and income distributions in consumer data. Third, Z-score detection offers an interpretable, threshold-based criterion for outlier identification, which is suitable for structured, low-dimensional features.

Mathematically, a data point x_{ij} is flagged as an outlier if its standardized score exceeds a critical value τ , defined in Eq. (5).

$$|z_{ij}| = \left| \frac{x_{ij} - \mu_j}{\sigma_j} \right| > \tau, \quad \text{where } \tau = 3 \quad (5)$$

The threshold of $\tau = 3$ corresponds to the empirical rule of Gaussian distributions, capturing 99.7% of values within ± 3 standard deviations. This allows for the removal of rare, extreme observations while preserving the general structure and variance of the data.

After applying this criterion, we observed that a small subset of records exceeded the ± 3 threshold in at least one feature. As summarized in Table 5, the number of identified outliers was relatively low, affecting only 2 instances in Age and 4 in Estimated Salary. These outliers represented values such as unusually low ages (e.g., near 18 years) combined with disproportionately high salaries (e.g., $\geq 150,000$ USD), which were not only statistical anomalies but also contextually questionable within the behavioral model applied.

Table 5. Number of observations flagged as outliers using z-score threshold $|z| > 3$.

FEATURES	TOTAL RECORDS	OUTLIERS DETECTED	PERCENTAGE (%)
Age (Z-normalized)	+400	2	~0.5%
Estimated Salary (Z)	+400	4	~1.0%

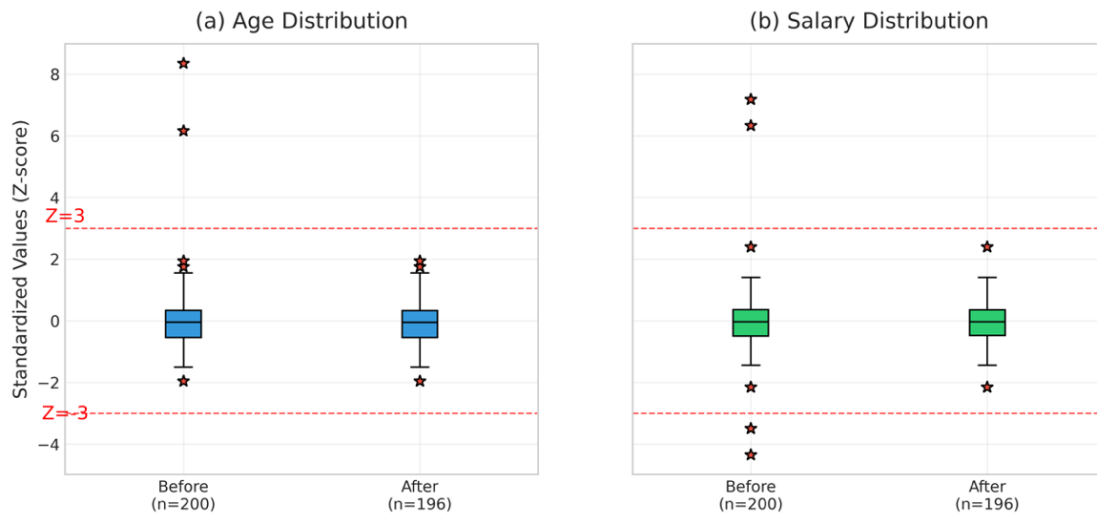


Fig 3. Distribution of age and estimated salary before (a) and after (b) outlier removal based on z-score.

Importantly, the removal of these outliers was not merely a technical formality but rather an effort to preserve the integrity of the feature space. Outliers can significantly influence model learning: for instance, in PCA, even a single extreme observation may distort the orientation of principal axes, leading to misleading lower-dimensional representations. Similarly, in classification models such as LR and SVM, these points can skew decision boundaries and degrade generalization on new data.

Fig. 3 visually illustrates the distribution of features before and after outlier removal using box plots. In Fig. 3(a), the pre-removal distribution reveals several

elongated whiskers and distant points, particularly in Estimated Salary. In contrast, Fig. 3.(b) shows that after outlier exclusion, both features exhibit tighter interquartile ranges and reduced skewness, indicating a more homogeneous and model-ready dataset.

2.3 Dimensionality reduction

In many data-driven health informatics applications, high-dimensional feature spaces are common, particularly when modeling behavioral, clinical, or biometric attributes. High dimensionality can result in overfitting, increased computational complexity, and degraded

generalization performance. To address these challenges, dimensionality reduction techniques are often employed to transform the original feature space into a more compact and informative representation.

In this study, we applied two established linear dimensionality reduction techniques: principal component analysis (PCA) and linear discriminant analysis (LDA). Although the dataset used in this study consists of only two numerical features, the inclusion of these techniques serves as a methodological evaluation to simulate real-world conditions where data often include redundant or correlated attributes. The dimensionality-reduced representations were used as input to five supervised classification models to evaluate the effectiveness of transformation on predictive performance.

a. Principal component analysis (PCA)

PCA is a well-established unsupervised dimensionality reduction technique widely used in machine learning and biomedical informatics. Its primary objective is to project a high-dimensional feature space into a lower-dimensional subspace while retaining as much of the original variance as possible. PCA operates independently of class labels, making it particularly effective for exploratory data analysis, visualization, and denoising in scenarios where structure exists without necessarily following classification boundaries.

Given a standardized input matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n represents the number of samples and d the number of features, PCA proceeds by computing the covariance matrix $\sum_n \frac{1}{n} \mathbf{X}^T \mathbf{X}$. The eigenvectors and corresponding eigenvalues of Σ are then derived. The eigenvectors, which define directions of maximal variance (i.e., principal components), are sorted in descending order based on their eigenvalue magnitudes. The top k components, corresponding to the highest variance are selected to form the projection matrix $\mathbf{W} \in \mathbb{R}^{d \times k}$, where $k < d$. The original data are then projected into the reduced space according to Eq. (6).

$$\mathbf{Z} = \mathbf{X} \cdot \mathbf{W} \quad (6)$$

In this study, PCA was applied to the normalized dataset before classification to investigate its ability to compress behavioral features such as age and estimated salary into a more compact representation while preserving the relevant structure. The results can be seen in Fig. 4.

As shown in Fig. 4, the PCA transformation retains the intrinsic variance of the dataset, revealing clusters and spread of data along principal axes. However, due to the unsupervised nature of PCA, class separation may not be explicitly enhanced, which highlights its utility as a general-purpose transformation rather than a class-optimized projection.

b. Linear discriminant analysis (LDA)

LDA is a supervised dimensionality reduction technique that projects data onto a lower-dimensional subspace while maximizing class separability. In contrast to PCA, which identifies directions of maximal variance without considering class labels, LDA explicitly uses

label information to enhance discrimination between categories.

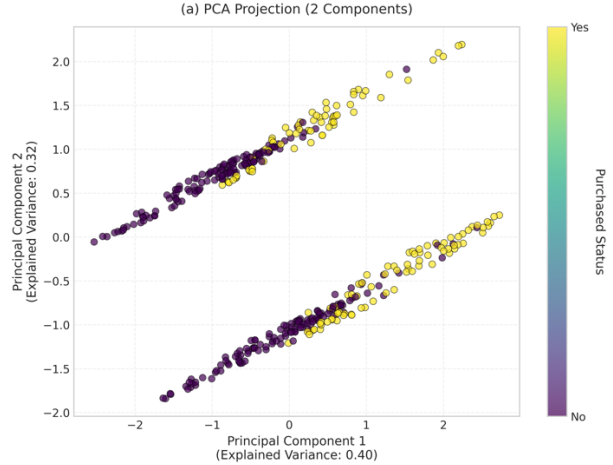


Fig 4. PCA projection of consumer data showing variance structure. Two-dimensional scatter plot of the first two principal components, illustrating data distribution and variance patterns. Data points are color-coded by purchase class (0 = No, 1 = Yes). While PCA preserves variance, it does not explicitly optimize for class separation.

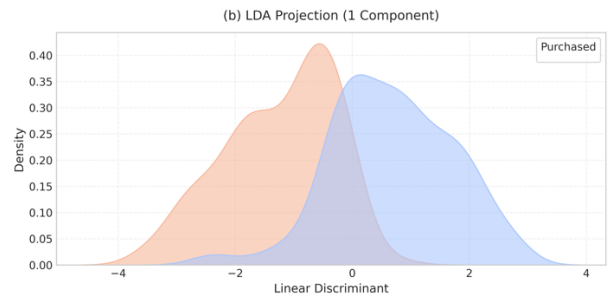


Fig 5. Kernel density estimation (KDE) of LDA-transformed feature.

Mathematically, LDA aims to find a projection vector that maximizes the ratio of between-class variance to within-class variance. This objective is described as the solution to the generalized eigenvalue problem calculated in Eq. (7).

$$\mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{w} = \lambda \mathbf{w} \quad (7)$$

where \mathbf{S}_w is the within-class scatter matrix, \mathbf{S}_b is the between-class scatter matrix, and λ represents the eigenvalues associated with the discriminant directions. The original feature matrix \mathbf{X} is then projected using the derived eigenvectors \mathbf{W} as in Eq. (8).

$$\mathbf{Z} = \mathbf{X} \cdot \mathbf{W} \quad (8)$$

In this study, the classification target variable Purchased is binary (i.e., 0 for not purchased, 1 for purchased). Consequently, LDA reduces the two-dimensional feature space (Age and EstimatedSalary) into a single discriminant axis. This one-dimensional representation simplifies the classification task and enables clear visualization of class separation.

The impact of LDA on class separation is illustrated in Fig. 4, which displays a kernel density estimation (KDE) plot of the transformed feature. The figure demonstrates that the projected values for each class form distinct, largely non-overlapping distributions. This indicates that

LDA effectively captures the directions in which the two classes diverge most significantly, resulting in improved linear separability and enhanced classification potential.

As illustrated in Fig. 5 shows the distribution of instances after projection onto the linear discriminant component. The distinct peaks corresponding to the Purchased classes (0 and 1) highlight LDA's ability to concentrate discriminative information along a single axis, thereby supporting an effective binary classification.

In summary, our results show that LDA has proven to be a valuable transformation in scenarios involving labeled data. Its capacity to reduce dimensionality while retaining class-discriminative power contributes directly to improved classifier performance, as will be further analyzed in the subsequent evaluation section.

2.4 Classification algorithms

To evaluate the predictive performance of the transformed dataset, a set of well-established machine learning classifiers was employed. The selected algorithms represent a diverse spectrum of classification paradigms, ranging from probabilistic to discriminative and ensemble-based models. This diversity enables a comprehensive comparison across different learning strategies and facilitates the identification of models best suited to behavioral prediction tasks within the domain of health informatics.

The five classifiers evaluated in this study are: LR, SVM, DT, NB, and RF. Each algorithm was applied to the preprocessed dataset, both with and without dimensionality reduction (PCA and LDA), to assess performance consistency under varying feature representations.

a. Logistic regression (LR)

LR is a widely used linear model for binary classification tasks. It models the probability of a target class using the logistic (sigmoid) function, as expressed in Eq. (9).

$$P(y = 1|x) = \frac{1}{1 + e^{-(w^T x + b)}} \quad (9)$$

Despite its simplicity, LR is known for its robustness, interpretability, and effectiveness in linearly separable data, making it a reliable baseline in clinical and behavioral datasets.

b. Support vector machine (SVM)

SVM is a powerful discriminative classifier that constructs an optimal hyperplane to separate classes in a high-dimensional space. When the data are not linearly separable, kernel functions (e.g., radial basis function) are employed to map the input into a higher-dimensional space. In this study, a linear kernel was used due to the low feature dimensionality and to preserve model interpretability.

c. Decision tree (DT)

The DT classifier is a non-parametric model that recursively splits the data space into subsets based on feature values that yield the highest information gain. Its intuitive tree-like structure allows for easy visualization

and interpretation of decision rules. However, DTs are prone to overfitting, especially in small datasets, which is mitigated in this study through pruning and cross-validation.

d. Naive bayes (NB)

NB is a probabilistic classifier based on Bayes' theorem, assuming independence among features. Despite this strong assumption, it often performs surprisingly well in practice, particularly when feature relationships are weakly correlated, as is often the case in demographic and socioeconomic datasets.

$$P(y|x) \propto P(y) \prod_{i=1}^n P(x_i|y) \quad (10)$$

In this study, the Gaussian variant was adopted due to the continuous nature of the input features.

e. Random forest (RF)

RF is an ensemble learning method that combines multiple decision trees trained on bootstrapped samples and random feature subsets. The final prediction is made via majority voting across the ensemble, which significantly improves generalization and reduces variance compared to individual decision trees. RF is particularly robust to outliers and noise, making it suitable for real-world behavioral data.

Each model was implemented using Scikit-learn, a widely used machine learning library in Python, with default parameters initially, followed by hyperparameter tuning via grid search and stratified k-fold cross-validation. Performance evaluation was conducted using multiple metrics, accuracy, precision, recall, F1-score, and AUC-ROC to ensure robust and balanced assessment across models and scenarios.

3. Results and Discussion

This section presents the empirical findings from the classification experiments conducted on the customer behavioral dataset. Performance was evaluated across five classifiers, LR, SVM, DT, NB, and RF, under three data configurations: (1) original normalized features, (2) PCA-transformed features, and (3) LDA-transformed features.

The classification results were assessed using five standard evaluation metrics: accuracy, precision, recall, F1-score, and area under the ROC Curve (AUC). This comprehensive evaluation ensures that each model's predictive power is analyzed not only in terms of overall correctness but also in terms of its balance between sensitivity and specificity, critical in real-world applications where class imbalance or cost asymmetry may be present.

3.1 Comparative performance evaluation

Table 6 summarizes the performance of each classifier across the three feature transformation scenarios. The reported values represent the average scores over 10-fold stratified cross-validation to mitigate sampling bias.

3.2 Interpretation and insights

Across all configurations, RF demonstrated consistently strong performance, particularly when

applied to features transformed using LDA. Under this setting, RF achieved the highest scores in AUC (0.95), accuracy (0.91), and F1-score (0.86), highlighting its robustness in capturing complex, non-linear relationships within the data. These results underscore RF's capacity to generalize effectively, especially when the decision boundaries are intricate and the feature space is optimized for class separability.

SVM also yielded competitive results, exhibiting notable gains in both precision and recall under the LDA

configuration. This further reinforces the practical value of LDA in enhancing the discriminative power of feature representations, especially in binary classification tasks. In contrast, a modest decline in performance was observed for several models under PCA-transformed features, particularly in the cases of decision tree and naive bayes. This suggests that although PCA effectively preserves overall variance, it may not always align with class-separating directions, thereby limiting its utility in certain supervised learning contexts.

Table 6. Performance metrics of classifiers across feature spaces (normalized, PCA, and LDA).

MODEL	FEATURE TYPE	ACCURACY	PRECISION	RECALL	F1-SCORE	AUC
LR	Normalized	0.86	0.83	0.79	0.81	0.91
	PCA	0.84	0.8	0.78	0.79	0.89
	LDA	0.89	0.87	0.82	0.84	0.93
SVM	Normalized	0.87	0.84	0.81	0.82	0.92
	PCA	0.83	0.79	0.76	0.77	0.88
	LDA	0.9	0.86	0.84	0.85	0.94
DT	Normalized	0.79	0.76	0.73	0.74	0.82
	PCA	0.77	0.75	0.7	0.72	0.8
	LDA	0.83	0.81	0.78	0.79	0.86
NB	Normalized	0.81	0.77	0.76	0.76	0.84
	PCA	0.8	0.76	0.74	0.75	0.83
	LDA	0.86	0.82	0.8	0.81	0.88
RF	Normalized	0.89	0.86	0.83	0.84	0.94
	PCA	0.87	0.84	0.8	0.82	0.92
	LDA	0.91	0.88	0.85	0.86	0.95

Interestingly, even relatively simple models such as LR benefited from the LDA transformation. When projected onto the LDA axis, the model outperformed its performance using raw and PCA features. This observation highlights LDA's efficacy in scenarios where class boundaries are approximately linear but difficult to distinguish in the original feature space due to overlap or scale disparities.

Taken together, these findings support the strategic integration of dimensionality reduction within the preprocessing pipeline. Specifically, LDA emerges as a valuable step for improving classification outcomes by projecting data into a subspace that captures the most discriminative information as clearly illustrated in Fig. 5. While each model offers unique advantages, the combination of LDA transformation with robust classifiers such as RF and SVM appears to yield the most balanced and reliable results.

3.3 Discussions

The findings of this study are consistent with prior literature in behavioral informatics, where demographic variables such as age and income frequently display overlapping distributions between outcome classes. This overlap often complicates classification tasks, particularly when decision boundaries are not well-defined in the

original feature space. The observed improvement in classification performance following the application of Linear Discriminant Analysis (LDA) reinforces the value of supervised dimensionality reduction for uncovering subtle yet meaningful patterns within socio-economic attributes.

LDA's effectiveness in enhancing class separability is especially relevant in health informatics, where behavioral features, such as program participation, medication adherence, or lifestyle choices, typically reflect nuanced variations rather than distinct groupings. By projecting the data onto a discriminative axis, LDA simplifies the classification task and aligns the feature space with clinically meaningful separations, thus improving both interpretability and predictive performance.

In addition, the consistently high performance of Random Forest across all scenarios underscores the robustness of ensemble learning methods when applied to behavioral datasets. These algorithms are well-suited to capturing non-linear relationships and are resilient to outliers, both common characteristics in real-world health-related data. The superior results of Random Forest, particularly when combined with LDA, support its suitability for behavior-informed health prediction tasks.

These results are in line with findings from [Jiménez et al. \(2025\)](#), who demonstrated that LDA and kernel PCA

enhanced fetal health classification in low-dimensional clinical datasets. Similarly, García et al. (2024) reported improvements in plant disease detection through the combined use of PCA and LDA, a context that, despite being agricultural, shares methodological parallels with behavioral modeling. The present study extends these findings to health-related consumer data, confirming the practical benefits of LDA in producing accurate and interpretable predictive models.

Overall, these findings provide empirical support for incorporating dimensionality reduction, particularly LDA, into preprocessing pipelines for behavior-based health informatics. While dimensionality reduction improves feature representation, robust classifiers such as Random Forest and SVM further leverage this structure to achieve accurate, generalizable predictions. This combination is highly applicable in personalized healthcare systems, where the early detection of behavioral patterns is essential for effective decision support.

5. Conclusion and Future Work

This study examined how dimensionality reduction techniques, PCA and LDA, affect the classification performance of five machine learning algorithms when applied to behavioral data relevant to health informatics. Our results demonstrated that LDA significantly enhances classification metrics across all models, particularly when paired with Random Forest and SVM. These improvements support the integration of supervised transformation in data pipelines to improve prediction accuracy and model interpretability in health-related applications.

Despite the promising results, the study is constrained by the limited dimensionality of the dataset. Future work should explore more complex and higher-dimensional datasets incorporating behavioral, clinical, and contextual features. In addition, extending the framework to temporal or unstructured data (e.g., wearables, surveys) may further improve applicability in real-world healthcare systems.

In summary, the findings support a replicable and scalable approach to behavior-based health prediction using LDA-enhanced models. This contributes to the development of intelligent and personalized decision support systems within the broader field of health informatics.

Future work may also explore model deployment through real-time health monitoring systems or integration with patient engagement platforms to validate real-world applicability.

Author Contributions

D.F.N.A. conducted the preprocessing, implemented the classification algorithms, and contributed to the experimental design. D.M. supervised the study, conceptualized the methodology, performed the result validation, and coordinated the manuscript preparation. X.X. contributed to the refinement of the dimensionality reduction framework, critically revised the manuscript for important intellectual content, and provided insights from a global health informatics perspective. All authors reviewed and approved the final version of the manuscript.

Acknowledgements

The authors would like to express their sincere gratitude to the anonymous reviewers for their insightful comments and constructive suggestions, which significantly contributed to the improvement of this manuscript. The authors also acknowledge the open-source community and Kaggle for providing access to the dataset used in this study, which enabled reproducibility and transparency in the research process.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

- Azad, Md. S., Khan, S. S., Hossain, R., Rahman, R., & Momen, S. (2023). Predictive modeling of consumer purchase behavior on social media: Integrating theory of planned behavior and machine learning for actionable insights. *PLOS ONE*, 18(12), e0296336. <https://doi.org/10.1371/journal.pone.0296336>
- Darmawahyuni, A., Nurmaini, S., Tutuko, B., Rachmatullah, M. N., Firdaus, F., Sapitri, A. I., Islami, A., Marcelino, J., Isdwanta, R., & Karim, M. I. (2024). Health-Related Data Analysis Using Metaheuristic Optimization and Machine Learning. *IEEE Access*, 12, 55342–55356. <https://doi.org/10.1109/ACCESS.2024.3390008>
- García-Barrera, L. J., Meza-Zamora, S. A., Noa-Carrazana, J. C., & Delgado-Macuil, R. J. (2024). Chemometric analysis using infrared spectroscopy and PCA-LDA for early diagnosis of *Fusarium oxysporum* in tomato. *Journal of Plant Diseases and Protection*, 131(5), 1609–1626. <https://doi.org/10.1007/s41348-024-00978-y>
- Hayati, R., Munawar, A. A., Lukitaningsih, E., Earla, N., Karma, T., & Idroes, R. (2024). Combination of PCA with LDA and SVM classifiers: A model for determining the geographical origin of coconut in the coastal plantation, Aceh Province, Indonesia. *Case Studies in Chemical and Environmental Engineering*, 9, 100552. <https://doi.org/10.1016/j.csee.2023.100552>
- Jiménez-Narváez, A. D., Vaca, V. D. C., Lóor-Duque, J. J., Martín, I. R. A., Reyes-Chacón, I. G., Vizcaino, P., & Morocho-Cayamcela, M. E. (2025). Predictive Modeling for Fetal Health: A Comparative Study of PCA, LDA and KPCA for Dimensionality Reduction. *IEEE Access*, 1–1. <https://doi.org/10.1109/ACCESS.2025.3553110>
- Kabir, M. R., Ashraf, F. Bin, & Ajwad, R. (2019). Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data. *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, 1–6. <https://doi.org/10.1109/ICCIT48885.2019.9038521>
- Nosakhare, E., & Picard, R. (2020). Toward Assessing and Recommending Combinations of Behaviors for Improving Health and Well-Being. *ACM Transactions on Computing for Healthcare*, 1(1), 1–29. <https://doi.org/10.1145/3368958>
- Samuel Ajibola Dada, & Adeleke Damilola Adekola. (2024). Leveraging digital marketing for health behavior change: A model for engaging patients through pharmacies. *International Journal of Science and Technology Research Archive*, 7(2), 050–059. <https://doi.org/10.53771/ijstra.2024.7.2.0063>
- Zhai, K., Yousef, M. S., Mohammed, S., Al-Dewik, N. I., & Qoronfle, M. W. (2023). Optimizing Clinical Workflow Using Precision Medicine and Advanced Data Analytics.

Processes, *11*(3), 939.
<https://doi.org/10.3390/pr11030939>