# PCA & Feature Selection Guide

## 1. What is PCA?

Principal Component Analysis (PCA) is an unsupervised machine learning technique used to reduce the number of features in a dataset while preserving as much variance (information) as possible.

It transforms the data into a new coordinate system, where the axes (principal components) represent directions of maximum variance.

## 2. Principal Components

- PC1 (Principal Component 1): The direction in which the data varies the most.
- PC2 (Principal Component 2): Perpendicular to PC1, capturing the second-most variance.
- The components are uncorrelated (orthogonal) and ordered by importance.

These components are linear combinations of the original features, and they allow us to compress the data with minimal loss of information.

## 3. Visualizing PCA

The image below shows a dataset with PC1 and PC2 marked. The red arrow is PC1 (maximum variance direction), and the orange arrow is PC2 (perpendicular).

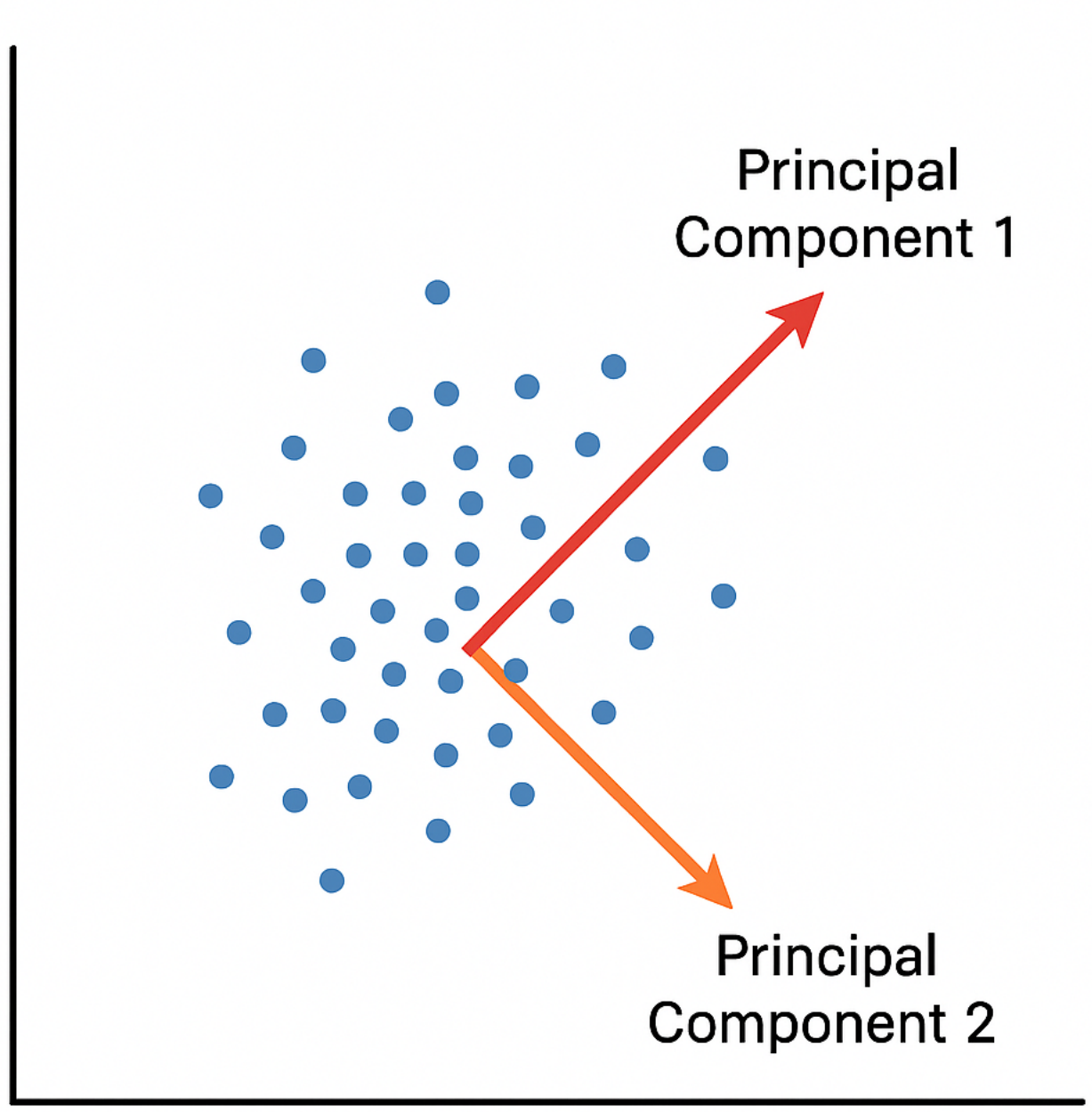


*Figure 1: PC1 and PC2 visualized in a 2D dataset. PC1 captures the most variance.*

### 4. Projecting Onto PC1

This image shows how data is projected onto PC1. This reduces the dataset from 2D to 1D while retaining its main structure.

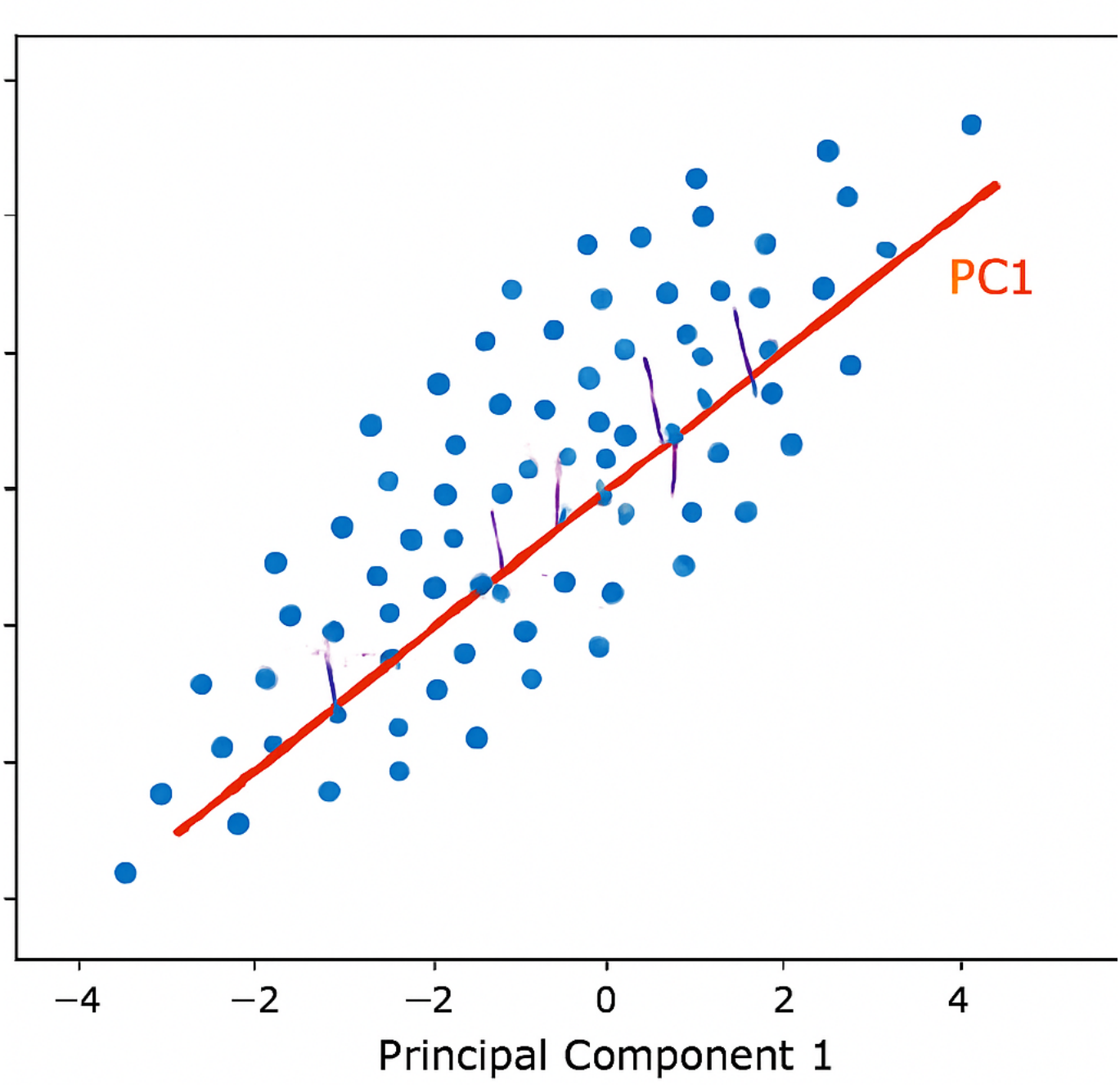


*Figure 2: Blue points are projected onto PC1. This illustrates how PCA compresses data.*

### 5. When is PCA Useful?

PCA can be useful when:

- The dataset has many features with correlation or redundancy.
- You want to reduce noise or compress the dataset.
- You need to visualize high-dimensional data in 2D or 3D.

However, PCA does not use label information and may discard predictive features if their variance is low.

# PCA & Feature Selection Guide

## 6. Supervised vs Unsupervised Learning

- PCA is unsupervised: it only looks at feature variance, not the label.
- In supervised learning, you define the label and select or rank features based on how well they help predict it.

## 7. Supervised Feature Selection Techniques

1. SelectKBest:
- Uses statistical tests to rank and keep top k features.

2. Recursive Feature Elimination (RFE):
- Trains a model and removes the least important features step by step.

3. Feature Importance (from tree models like Random Forest):
- Highlights which features contributed most to predictions.

4. L1 Regularization (Lasso):
- Penalizes less useful features and reduces some weights to zero.

## 8. Where Labels Matter in PCA

This example shows why PCA, being unsupervised, may ignore important features.

We simulate a case where:
- x1 has high variance but is unrelated to the label.
- x2 has low variance but determines the label.

Since PCA keeps directions of high variance, it would focus on x1 and possibly ignore x2. But x2 is essential to distinguish the classes.

This illustrates the risk of using PCA in supervised learning tasks without care.

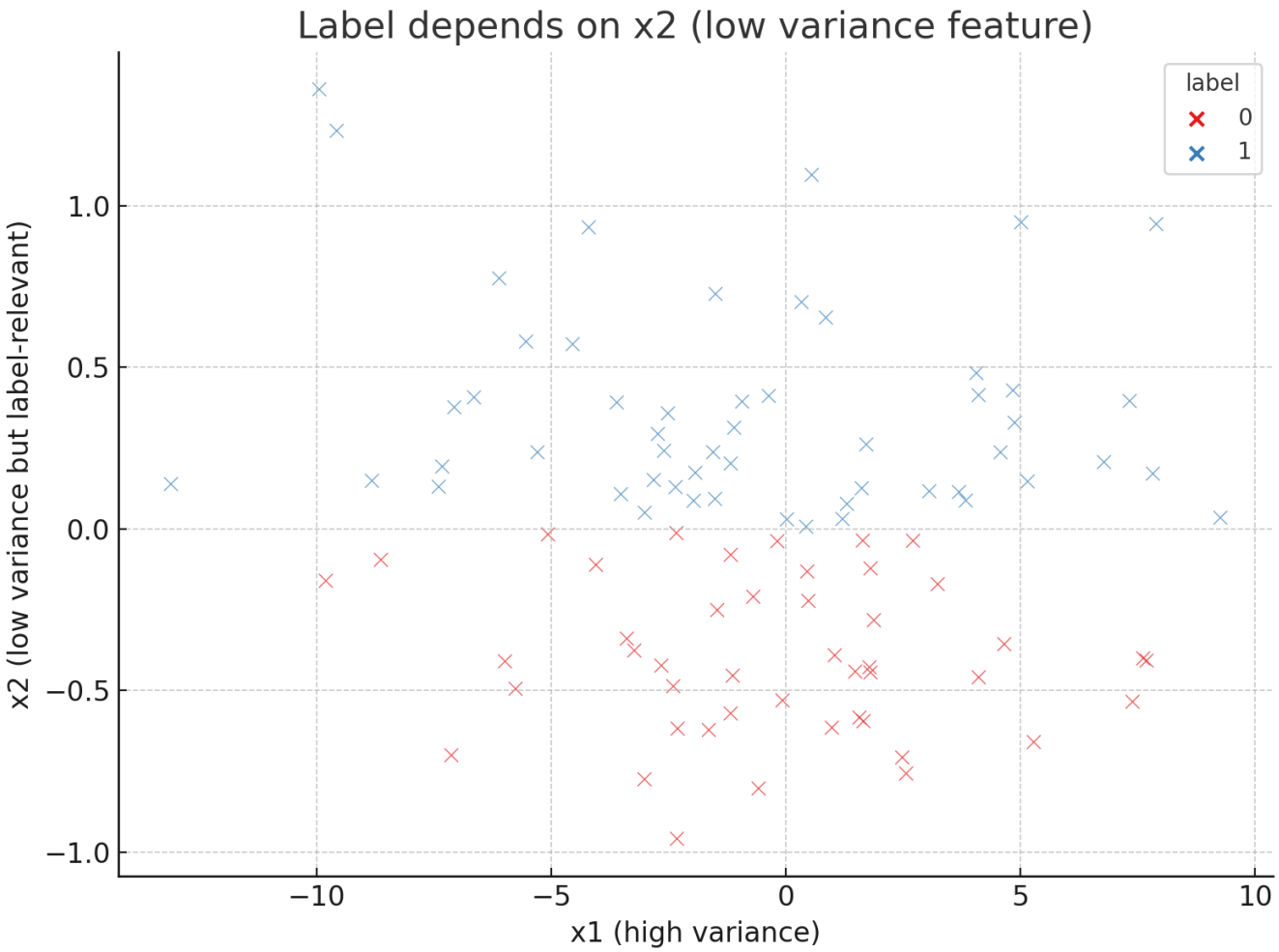


*Figure 3: Although x2 is related to the label, it has low variance and may be ignored by PCA.*

## 9. Summary

- PCA helps reduce feature space using variance and is unsupervised.

- Principal components are new axes aligned with directions of highest data spread.

- Feature selection in supervised learning identifies the best inputs for prediction.

- Use both PCA and feature selection carefully, depending on your goals: compression, visualization, or prediction.

All pipeline code, examples, and visuals are included in the GitHub repository accompanying this document.