Unsupervised Learning Lab

Assignment 2

PCA – Dimensionality Reduction algorithm

Theory of PCA

What is PCA?

PCA (Principal Component Analysis) is a technique that finds the most important patterns in your data and combines many measurements into fewer, more meaningful ones called "principal components."

The Core Idea PCA identifies relationships between your original features, creates new combined features ranked by importance, and lets you keep only the most informative ones, reducing complexity while preserving essence.

Simple Example With credit card data (purchase amount, transaction count, payment amount), PCA might combine these into:

- Component 1 (60%): "Spending Volume"
- Component 2 (30%): "Payment Behaviour"

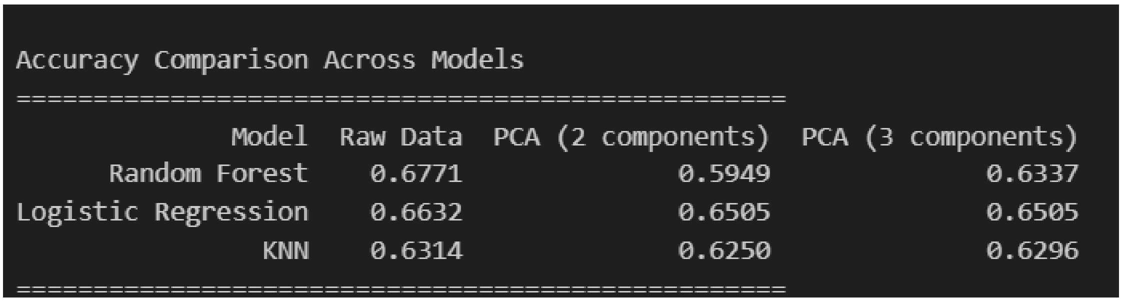
You now track 2 numbers instead of 3 while keeping 90% of the important information.

How It Works

1. Standardize: Scale all features to mean=0, std=1 so no single measurement dominates
2. Find directions: PCA identifies the direction where your data varies most (PC1), then the next best perpendicular direction (PC2), and so on
3. Select components: Keep enough components to capture your target threshold (typically 80-95% of total variance)

Real Example: Iris Dataset With 4 flower measurements, PCA reveals that just 2 components capture 96% of the information, making it easy to visualize and separate the 3 flower types on a 2D plot.

Results obtained



```
Accuracy Comparison Across Models
==================================================
              Model  Raw Data  PCA (2 components)  PCA (3 components)
      Random Forest    0.6771              0.5949              0.6337
Logistic Regression    0.6632              0.6505              0.6505
                KNN    0.6314              0.6250              0.6296
==================================================
```

Inferences from results

Raw data works best for all models - None of the PCA versions improved accuracy. This means the original 17 features contain important information that gets lost when we compress them.

Model-by-Model Breakdown

Random Forest: The Biggest Drop

- **Raw**: 67.7% accuracy (best overall)
- **PCA with 2 components**: 59.5% (big drop of 8%)
- **PCA with 3 components**: 63.4% (partial recovery)

What this means: Random Forest is good at finding complex patterns in many features. When we force it to use just 2-3 combined features, it loses the detailed information it needs.

Logistic Regression: Most Stable

- **Raw**: 66.3% accuracy
- **PCA**: 65.0% with both 2 and 3 components (very small drop)

What this means: Logistic Regression doesn't lose much with PCA. The main patterns in this data might be linear combinations of features, which PCA captures well.

KNN: The Least Affected

- **Raw**: 63.1% (lowest of all)
- **PCA**: 62.5% and 63.0% (almost no change)

What this means: KNN was already struggling with the raw data, so PCA doesn't make it much worse (or better).

Why PCA Didn't Help

1. **All features seem useful** - There's no obvious redundancy to remove
2. **Information loss** - Compressing 17 features into 2-3 loses some details that models need
3. **Non-linear relationships** - Some patterns might not be captured well by linear combinations (what PCA creates)

Key Takeaway

For this credit card dataset, **stick with the raw data**. PCA simplifies things but costs you accuracy - not worth the trade-off. Random Forest with all features gives the best result at 67.7%.