- **ACM476 TERM PROJECT PHASE 3**

Melbourne Housing

Kaan Akkök, Hasan Emir Çilesiz

28.05.2025

Content

1. Feature Selection
2. Principal Component Analysis (PCA)

Feature Selection:

Feature Removal

- Address: High cardinality, no predictive power
- SellerG: Seller names cause inconsistency
- Date: No direct correlation with price
- Propertycount: Weak correlation with price
- Postcode: Minimal impact, CouncilArea available

Price Transformation

- Applied log(1+x) transformation (right-skewed distribution)
- Created Price_Category using statistical binning:
  - Cheap: min to (mean-std)
  - Affordable: (mean-std) to (mean+std)
  - Expensive: (mean+std) to max

Target variable:

- Price_Category

Independent variables:

- Suburb
- Rooms
- Type
- Method
- Distance
- Bedroom2
- Bathroom
- Car
- Landsize
- CouncilArea
- Lattitude
- Longtitude
- Regionname

Feature Selection:

- We identified the numerical and categorical features to apply standard scaling and one-hot encoding.
- Rooms, Car, Bedroom, and Bathroom are numerical features but behave like categorical variables. Therefore, we reclassified them from numerical to categorical.
- Then, we applied standard scaling to numerical features and one-hot encoding to categorical features.
- Lastly we applied label encoding to Price Category feature.

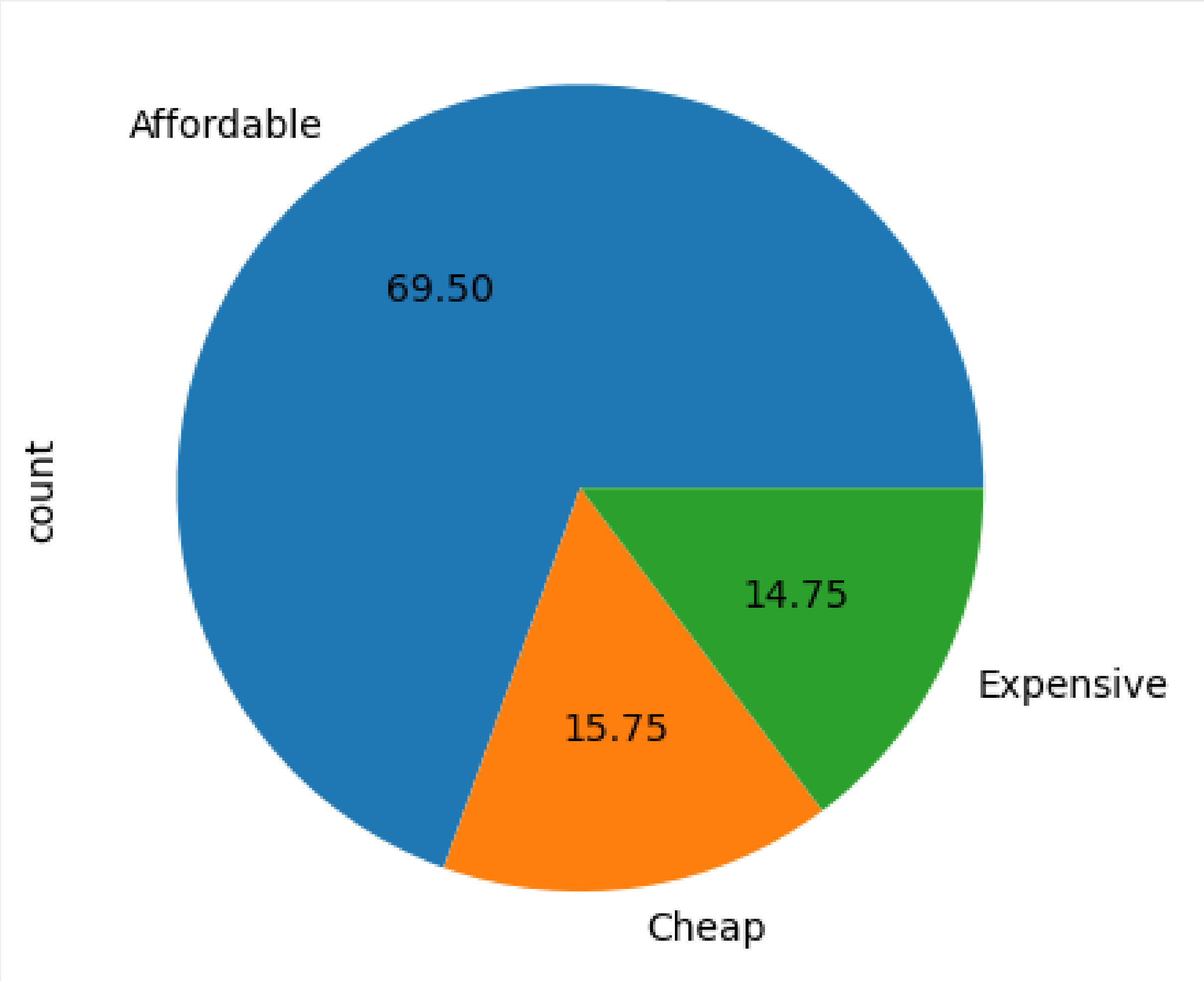
```
[10]: numerical_cols

[10]: ['Distance',
       'Landsize',
       'BuildingArea',
       'YearBuilt',
       'Lattitude',
       'Longtitude']

[11]: categorical_cols

[11]: ['Suburb',
       'Type',
       'Method',
       'CouncilArea',
       'Regionname',
       'Rooms',
       'Car',
       'Bedroom2',
       'Bathroom']
```

Feature Selection:

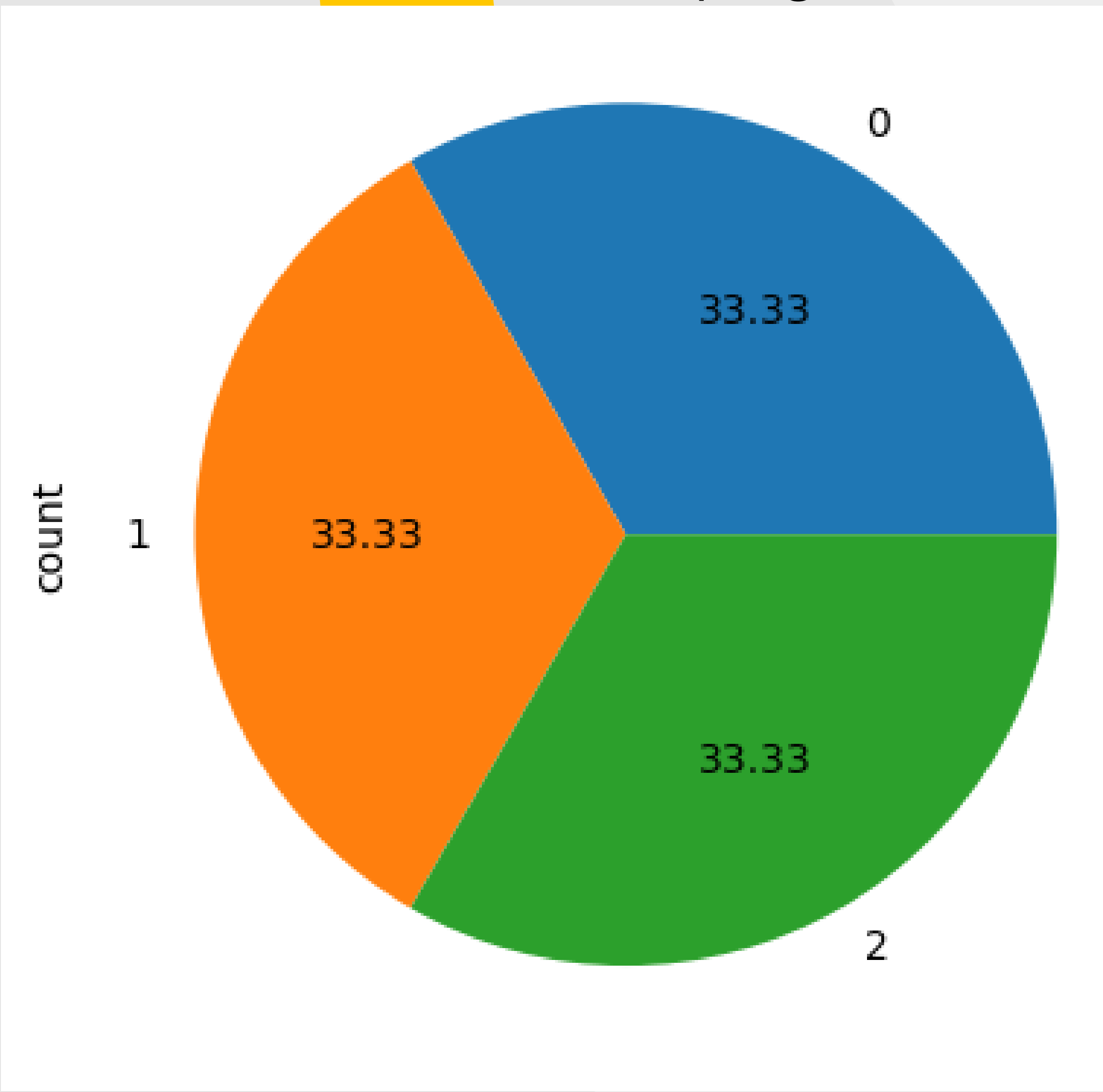
After creating the Price Category feature, we noticed that our data was imbalanced, so we applied under-sampling to balance it.

```
Price_Category
Affordable    1390
Cheap          315
Expensive      295
Name: count, dtype: int64
```

Before Under Sampling



After Under Sampling



Feature Selection:

Mutual Information Classifier

- Works on both numerical and categorical features
- Works on non-linear relation
- Selected Top 5 features based on MI scores

```
Top 5 Features:
 ['BuildingArea', 'Distance', 'Longtitude', 'Lattitude', 'Bathroom_1']
```

Using the top 5 features, we trained a Random Forest classifier.

```
All features - CV accuracy:  0.78
Top 5 features - CV accuracy:  0.73
```

Feature Selection:

Random Forest Classifier Feature importances

- Works on both numerical and categorical features
- Works on non-linear relation
- Selected Top 5 features based on importances scores

```
Top 5 Features:
 ['BuildingArea', 'Longtitude', 'Lattitude', 'Distance', 'Landsize']
```

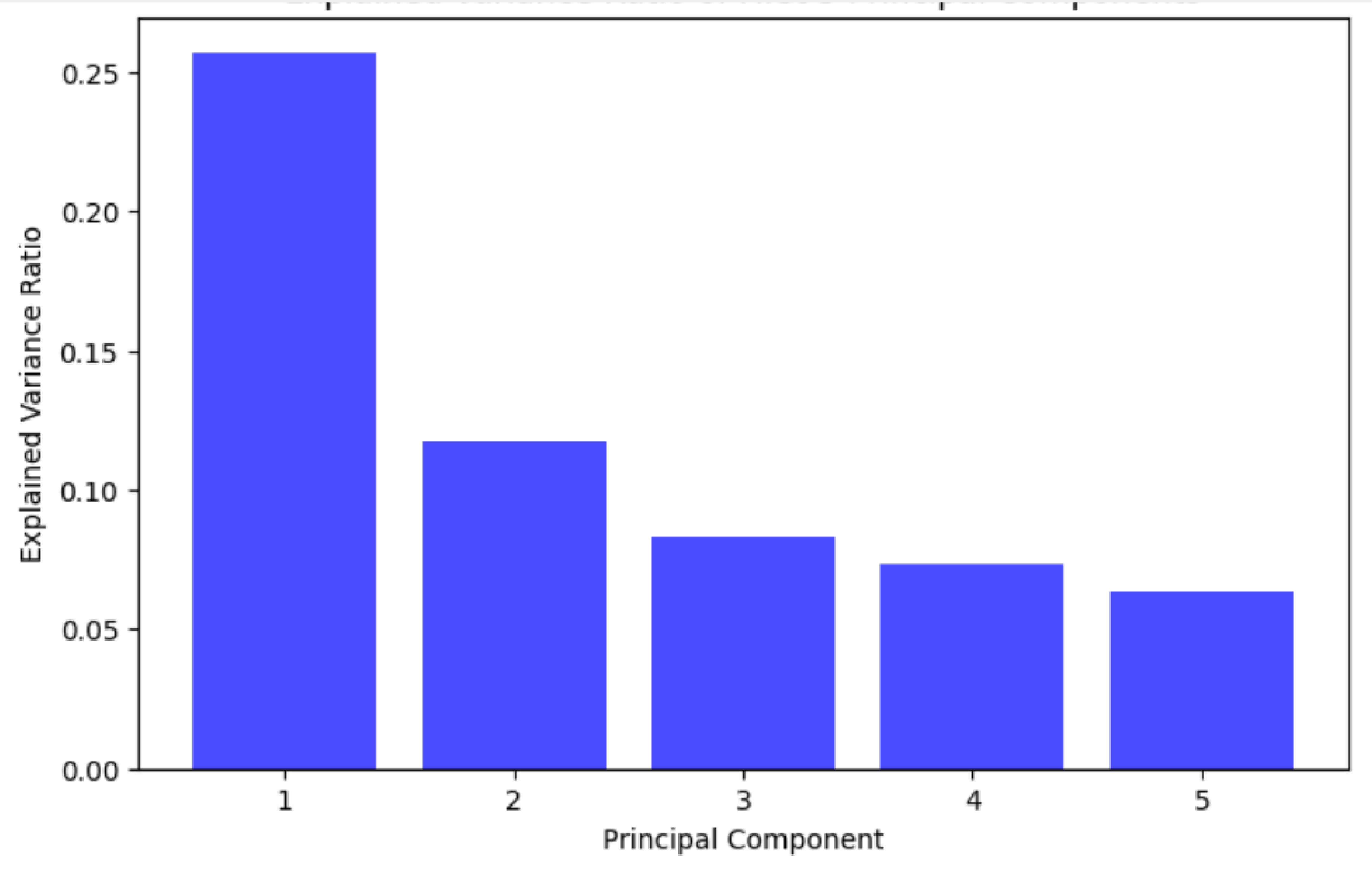
Using the top 5 features, we trained a Random Forest classifier.

```
All features - CV accuracy:  0.78
Top 5 (RF feature importances) - CV accuracy: 0.75
```

Principal Component Analysis

We prepared our dataset for performing Principal Component Analysis (PCA) by applying feature removal, price transformation, standard scaling, one-hot encoding, and under-sampling. Then, we examined the first 5 principal components.

Explained Variance Ratio of First 5 Principal Components



- **PC1:** 0.257
- **PC2:** 0.118
- **PC3:** 0.083
- **PC4:** 0.073
- **PC5:** 0.063

Total explained variance by first 5 components:

0.593

Key Feature Loadings per Principal Component

	PC1
Rooms	0,49
Bedroom2	0,48
BuildingArea	0,44
Bathroom	0,42
EVR	0,26

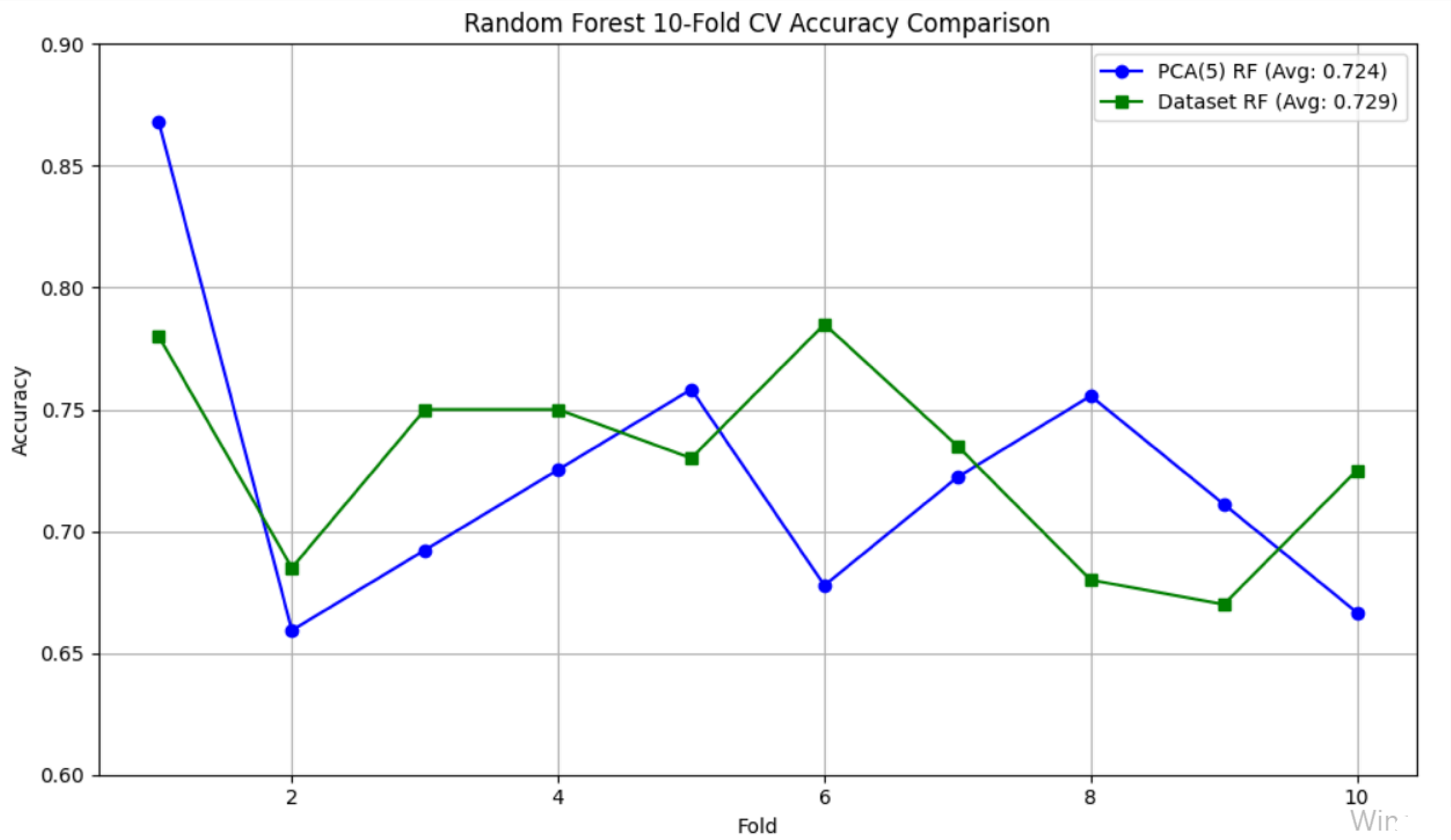
	PC2
Lattitude	0,56
Longtitude	-0,51
Distance	0,41
YearBuilt	0,29
EVR	0,12

	PC3
YearBuilt	0,68
Distance	0,47
Lattitude	-0,34
Longtitude	0,31
EVR	0,08

	PC4
Landsize	0,93
Lattitude	-0,20
Car	0,19
YearBuilt	-0,10
EVR	0,07

	PC5
Car	0,49
Longtitude	-0,45
Lattitude	-0,43
Landsize	-0,29
EVR	0,06

10-Fold Cross-Validation Comparison



Model Performance Comparison

By Balanced Dataset

	precision	recall	f1-score	support
expensive -> 0	0.72	0.80	0.76	132
cheap -> 1	0.82	0.81	0.81	142
affordible -> 2	0.62	0.55	0.58	126
accuracy			0.72	400
macro avg	0.72	0.72	0.72	400
weighted avg	0.72	0.72	0.72	400

Balanced Dataset Model: Shows more consistent and balanced performance across all classes (Expensive, Cheap, Affordable), particularly stronger for "Affordable."

By the first 5 principal components.

	precision	recall	f1-score	support
Affordable	0.59	0.52	0.55	295
Cheap	0.79	0.85	0.82	315
Expensive	0.77	0.79	0.78	295
accuracy			0.72	905
macro avg	0.71	0.72	0.72	905
weighted avg	0.72	0.72	0.72	905

Principal Components Model: Performs well for "**Cheap**" and "**Expensive**" but struggles more with the "Affordable" class.

Overall Accuracy: Both models achieved a 72% overall accuracy.

The Balanced Dataset approach provides more reliable and consistent classification across all categories, this makes it generally **preferable**.