Hospital Package Price Prediction and Patient Clustering

Machine Learning Assignment 3 - Mission Hospital Durgapur

By Kamya Sarda

Introduction

This project analyzes patient data from Mission Hospital Durgapur to solve two problems:

1. Predict hospital package prices using patient information
2. Group similar patients into clusters to understand different patient types

The dataset contains 248 patient records with information about demographics, medical conditions, treatment details, and costs.

I used Random Forest Regression for price prediction and Hierarchical Clustering to create patient groups.

Price Prediction Model

Methodology

I split the data into 80% training (198 patients) and 20% testing (50 patients) to evaluate model performance. I chose Random Forest Regressor because it handles both numerical and categorical data well and doesn't require assumptions about data distribution.

Model Configuration:

- Algorithm: Random Forest Regressor
- Number of trees: 200
- Random state: 42 (for reproducibility)

Results

The model performed well on the test set:

Metric	Value	What It Means
R^2 Score	0.728	Model explains 72.8% of cost variation
MAE	₹39,563	Average prediction error is ₹40K
RMSE	₹71,921	Larger errors are penalized more

The R^2 score of 0.728 means the model is quite accurate. Getting 72.8% is good because hospital costs naturally vary due to complications and individual patient responses.

The MAE of ₹39,563 tells us that on average, predictions are off by about ₹40,000. Since the average hospital cost is ₹217,845, this represents roughly 18% error, which is reasonable for cost estimation.

Patient Clustering

I used Hierarchical Clustering with Ward's linkage method to group patients based on their characteristics. Ward's method minimizes the variance within each cluster, creating groups of similar patients.

The clustering identified three distinct patient groups:

Cluster	Number of Patients	Percentage

Cluster 1	132 patients	53.2%
Cluster 3	83 patients	33.5%
Cluster 2	33 patients	13.3%

What the Dendrogram Shows:

The dendrogram displays how patients are grouped together based on similarity. The height of each branch shows how different groups are from each other. I chose to cut at 3 clusters because the dendrogram shows a clear separation at that level - cutting at 2 would be too broad, and cutting at 4 or more would create groups that are too similar.

PCA Visualization:

I used PCA (Principal Component Analysis) to reduce the data to 2 dimensions for plotting. The PCA plot shows the three clusters:

- Cluster 1 (Purple dots): Tightly grouped on the left side, indicating these patients are very similar
- Cluster 3 (Yellow dots): Scattered in the middle and right, showing more variation
- Cluster 2 (Teal dots): Separated at the top, representing distinct cases

The clusters have very less overlap, which means the grouping captured meaningful differences between patient types.

Understanding the Patient Groups

Cluster 1 has 132 patients and represents the majority group. These seem to be routine care patients with shorter hospital stays and lower costs. They don't need much ICU time and are probably elective procedures that went smoothly.

Cluster 3 has 83 patients in the middle group. These cases are more complicated than Cluster 1 but not critical. It's a mix of planned surgeries and emergencies. Some patients needed ICU and others didn't. Costs are in the middle range.

Cluster 2 is the smallest group with just 33 patients. These are the critical care patients with much longer stays and lots of ICU time. These are the expensive cases, usually emergencies with complications.

Key Findings

With 72.8% accuracy, the Random Forest model can estimate hospital costs with reasonable precision. The average error of ₹40,000 is acceptable given the complexity of healthcare pricing.
Hierarchical clustering revealed that patients naturally fall into three categories based on their characteristics and treatment complexity. The dendrogram clearly showed these separations.
The smallest cluster (33 patients) likely represents critical care cases that require significantly more resources, while the largest cluster (132 patients) represents routine procedures.
Although we didn't explicitly calculate feature importance, the clustering patterns and cost variations strongly suggest that how long patients stay in the hospital and ICU are the biggest factors affecting total cost.

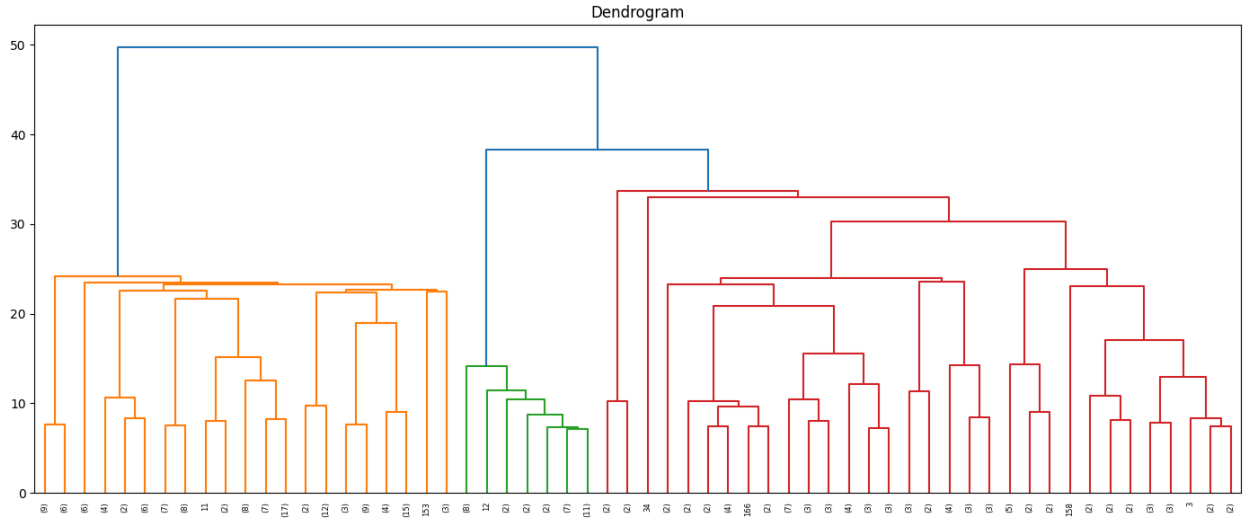
Summary

Random Forest regression got 72.8% accuracy predicting costs with ₹39,563 average error. Hierarchical clustering found three groups: routine care (53%), moderate cases (34%), and critical care (13%). These groups show clear differences in treatment intensity and resources needed.

VS Code Output & Visualizations Explanation/Justification

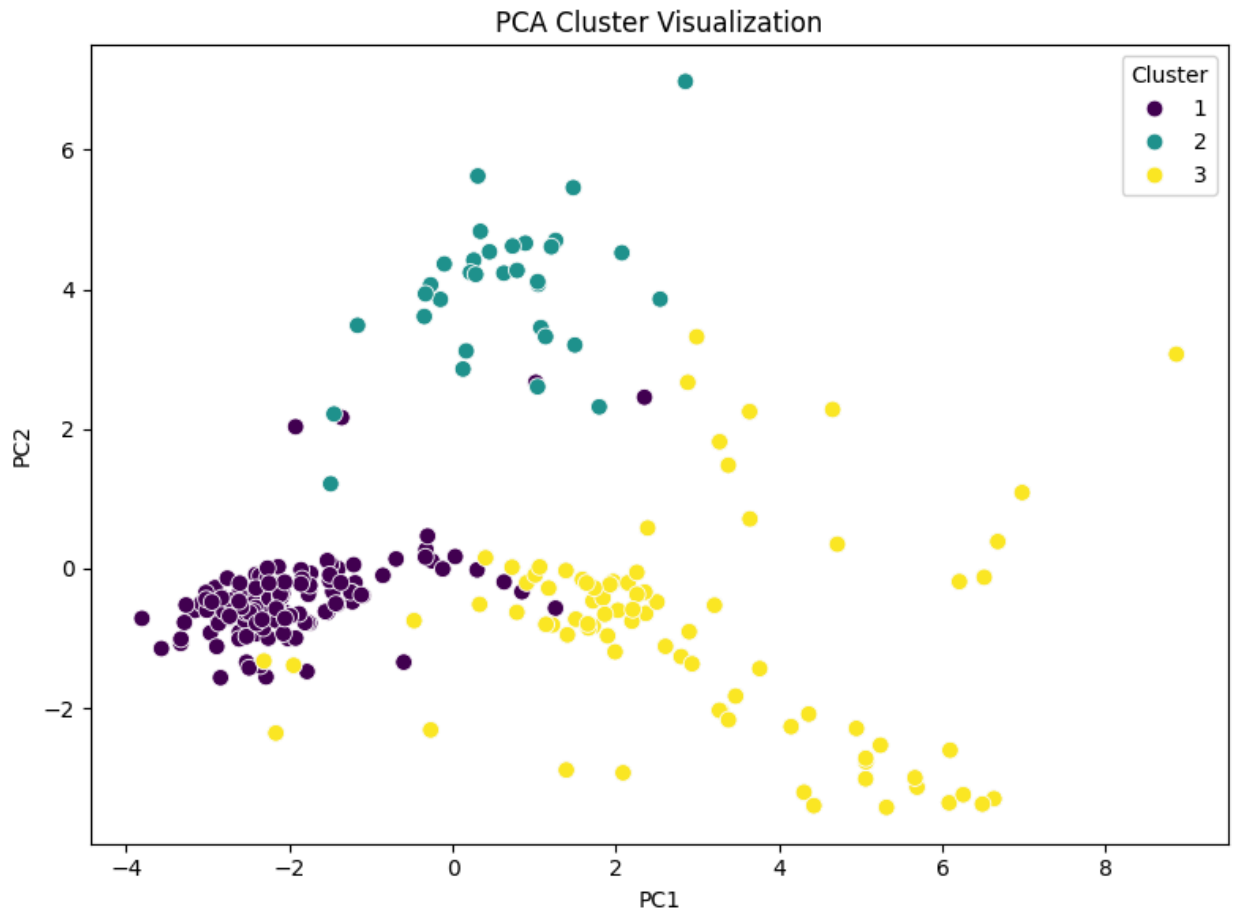
1. Dendrogram

The tree-like diagram shows how patients are grouped. Each vertical line represents joining two groups. The height indicates how different groups are - taller connections mean more different. The three main branches (colored differently) represent the three clusters.



2. PCA Cluster Plot

This scatter plot shows all 248 patients in 2D space, colored by cluster. Points close together are similar patients. The three colors (purple, yellow, teal) show the three clusters are well-separated, confirming they represent genuinely different patient types.



3. Model Performance

- R^2 of 0.728 means if we plot actual vs predicted costs, 72.8% of points would be close to a perfect diagonal line
- MAE of ₹39,563 is the average distance between predicted and actual costs
- These metrics confirm the model is reliable for cost estimation

```
[10]  ✓ 0.5s
```

```
...      SL.   AGE GENDER MARITAL STATUS KEY COMPLAINTS -CODE  ACHD  CAD-DVD  \
     0    1  58.0      M        MARRIED         other- heart     0        0
     1    2  59.0      M        MARRIED              CAD-DVD     0        1
     2    3  82.0      M        MARRIED              CAD-TVD     0        0
     3    4  46.0      M        MARRIED              CAD-DVD     0        1
     4    5  60.0      M        MARRIED              CAD-DVD     0        1

        CAD-SVD  CAD-TVD  CAD-VSD  ...  ALERT  TYPE OF ADMSN  ELECTIVE  \
     0        0        0        0  ...      1      EMERGENCY         0
     1        0        0        0  ...      1      EMERGENCY         0
     2        0        1        0  ...      1       ELECTIVE         1
     3        0        0        0  ...      1      EMERGENCY         0
     4        0        0        0  ...      1      EMERGENCY         0

        TOTAL COST TO HOSPITAL   TOTAL LENGTH OF STAY  LENGTH OF STAY - ICU  \
     0                660293.0                    25                    12
     1                809130.0                    41                    20
     2                362231.0                    18                     9
     3                629990.0                    14                    13
     4                444876.0                    24                    12

        LENGTH OF STAY- WARD  IMPLANT USED (Y/N)  IMPLANT   COST OF IMPLANT
     0                    13                   Y         1            38000
     1                    21                   Y         1            39690
     2                     9                   N         0                0
     ...
     R2: 0.7280078241008068
     MAE: 39562.873982000005
     RMSE: 71920.84639258002
     R2: 0.7280078241008068
```