Executive Summary

Project Title:

Predicting Hospital Performance and Resource Optimization Using Machine Learning

Team Members:

- Devika Rani Sanaboyina – NB79056
- Abhinav Varma Vathadi – CS77898

1. The problem that we aimed to solve

Throughout the health system, hospitals are perpetually functioning in a state of tension—facing resource shortages, congested emergency departments, and variable performance. Although there are efficiently functioning hospitals, others lack the ability for bed, staff, and service management.

Our project sought out the answer to the following simple question

Yes, machine learning can be utilized to predict hospital performance and resource requirements even before the problems actually occur.

We sought to

Identify the hospitals with a high risk of high demand

Classify hospitals according to operational efficiency

Provide insights to improve staffing, planning of capacity, and long-term planning

2. Data Underpinning Our Results

We employed a large dataset from the California Office of Statewide Health Planning and Development (OSHPD) that comprised over 578,000 hospital discharges.

Based on this set, we focused on

• Hospital capacity: Licensed beds, available beds, staffed beds

• Operational details: Facility name, location, and number of days in service

• Service consumption: Services offered and their volumes (Volume)

We also included the new target: a binary indicator of whether the hospital is anticipated to have high resource demands, calculated from the median Amount value.

Sometimes,

3. Our Approach to the Problem

- We used the CRISP-DM process, widely accepted for projects in data science:
- Defining the problem – converting real-life hospital demands into forecasting objectives
- Investigating the data - examining and ascertaining the dataset for trends, nulls, and outliers
- Data preparation – feature selection and scaling of representative features
- Model building – employing both regression, clustering, and classification techniques
- Outcome evaluation – comparison of model performance with appropriate metrics
- Making conclusions – linking our findings with the starting healthcare problem

4. What Did We Create – and What Did We Discover

Regression (Service Utilization Prediction)

We employed three models in order to predict the number of services (Amount):

Random Forest, XGBoost, and Linear Regression

- Random Forest Regressor worked best with R^2 Score: 92.27%
- RMSE: ~70 million (reasonable given the data is in billions)

Clustering (Grouping Hospitals by Efficiency)

- K-Means: Identified 3 separate clusters (low, medium, and high efficiency)
- DBSCAN: Did not perform well with density separation on shaped data
- Agglomerative Clustering: Worked well on sampled data

They found that certain hospitals process greater patient volumes with fewer resources, suggesting varying operational efficiency.

K-Means has best performance as it worked on entire dataset. Remaining 2 algorithms gave run time crashes when performed on entire dataset. Hence we have finalized K-Means for this problem.

Classification (High-Demand Hospital Forecast)

We trained models for predicting whether a hospital would have high demand

- Logistic Regression
- Random Forest Classifier
- Gradient Boosting

The Random Forest Classifier again gave the best performance

Accuracy: 84

• High recall for recognizing hospitals in distress

Surprisingly, we found that removing the outliers worsened performance, so we included them in the final model, recognizing that in real data, outlier values usually carry useful signals.

5. Why This Matters

They can help hospital administrators and policymakers by:

- Identifying Hospitals that may need additional support
- Showing how resource consumption differs between institutions
- Informed decision-making in staffing and infrastructure
- In brief, machine learning can help make healthcare proactive, rather than reactive.

6. Concluding Observations

This project proves that with a dirty, complicated dataset, careful modeling is beneficial. In terms of planning with forecast potential or resource maximizing, the methods we employed—regression, clustering, and classification—present a good foundation for evidence-based health improvement.

It is our hope that our research will contribute to better-informed public health planning and hospital administration.