MEDICAL MARKET SEGMENTATION ANALYSIS

Contributors: Aniket - Aaryya Sengupta

INTRODUCTION

Healthcare data are growing exponentially, so we need intelligent systems to derive actionable intelligence. Our project is a solid interactive platform created in Streamlit that helps segment clinical data and predict diagnostics. This is for users like health care professionals, analysts, and clinical administrators, to reduce complexity, extract patterns and analytics and help them make informed decisions based on data.

The system has two main modules: unsupervised clustering (exploratory data analysis - EDA) and machine learning (ML) predictions for the diagnostic test results. Users will be able to upload de-identified structured data (CSV file), filter for the relevant demographics, and see trends in billings, conditions and admissions. Users can also plug-in trained ML models to classify the diagnostic test outcomes, with visual feedback and performance metrics.

In summary, our project seeks to fill the gap between raw medical records and consumable analytics. It provides a low-code, scalable approach to clinical decision support.

PROBLEM STATEMENT

Hospitals and clinics frequently lack systems to identify patient segments as well as predict health conditions. Health systems also typically cannot easily maintain ongoing manual tracking of patient data, see how costs are spent, or use analytics to support earlier diagnosis or intervention.

Key Pain Points:

- No AI capabilities to locate high risk patients or billing exceptions.
- Low utilization of historical tests.
- Healthcare providers are often forced to guess how best to prioritize resources because they have no way to objectively classify based on data.

A smart system that clusters and predicts diagnosis will support improved preventive care and more efficient operations.

MARKET OVERVIEW & NEEDS ASSESSMENT

Recognized Challenges:

- Patient data is underutilized for analytics.
- Manual analysis is not scalable and error prone.
- Hospital IT systems are largely fragmented and legacy based.

Market Gaps:

- Lack of AI-based clinical dashboards available to small hospitals and clinics.
- Lack of explainable ML tools in the health domain for healthcare analytics.

Trends Driving Need:

- Global shift to digitization in healthcare after COVID.
- Increased focus on personalized and predictive healthcare.
- Increased policy focuses on medical data governance and reporting transparency.

According to recent reports by McKinsey, the expanding role of AI in diagnostics and operations is projected to be a $30B+ opportunity by 2030.

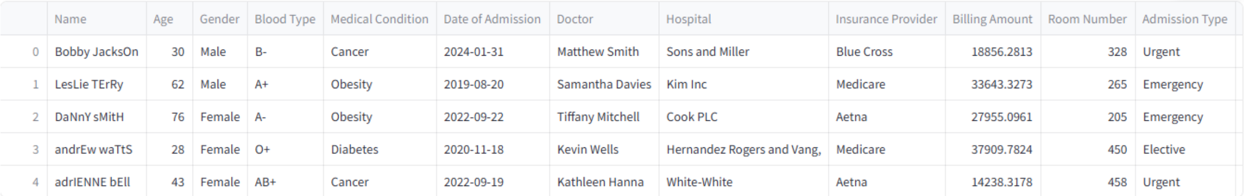
DATASET OVERVIEW

The project relies on systematic CSV file of synthetic or anonymized patient health records. The data set contains significant variables relating to patient demographics, type of admission, diagnosis, and test results. The general schema for the data set will be detailed below:

Column Name	Description	Type
Name	Patient full name (anonymized)	Text
Gender	Patient gender (Male/Female)	Categorical
Age	Age of patient in years	Numeric
Billing Amount	Hospital billing amount	Float
Admission Type	Admission category (Emergency, Routine, Urgent)	Categorical
Medical Condition	Primary diagnosis (e.g., Diabetes, Asthma)	Categorical
Test Results	Diagnostic test outcome (Normal/Abnormal/Inconclusive)	Categorical
Date of Admission	Admission date	Date
Discharge Date	Discharge date	Date

The system first preprocesses the data set through label encoding and median imputation for both categorical and missing data. Test Results were assigned numerical classes to model training purposes.

SAMPLE DATA



	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type
0	Bobby JacksOn	30	Male	B-	Cancer	2024-01-31	Matthew Smith	Sons and Miller	Blue Cross	18856.2813	328	Urgent
1	LesLie TErRy	62	Male	A+	Obesity	2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.3273	265	Emergency
2	DaNnY sMitH	76	Female	A-	Obesity	2022-09-22	Tiffany Mitchell	Cook PLC	Aetna	27955.0961	205	Emergency
3	andrEw waTtS	28	Female	O+	Diabetes	2020-11-18	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.7824	450	Elective
4	adrIENNE bEll	43	Female	AB+	Cancer	2022-09-19	Kathleen Hanna	White-White	Aetna	14238.3178	458	Urgent

Dataset Source: Kaggle - Healthcare Dataset

PROJECT SOLUTIONS SUMMARY

This project provides a two-part analytics and prediction dashboard for healthcare datasets.

1. EDA & Clustering Panel

- Interactive filters: (e.g., gender, age range)
- Count and average billing examination
- Medical condition occurrences
- Admissions type occurrence
- PCA + K-Means clustering approach applied to Age and Billing Amount

2. ML Prediction Panel

- Predict patient results based on ML classifiers (6 different classifiers)
- Label Encoder, Simple Imputer and train/test split (70 / 30)
- Outputs confusion matrix plots for each model

Value Proposition

- No-code, visual-first dashboard for clinicians and analysts.
- Transparent and explainable results.
- Actionable cluster-based segmentation for targeting risks and underlying budget model.

TARGET CUSTOMER

This solution is for:

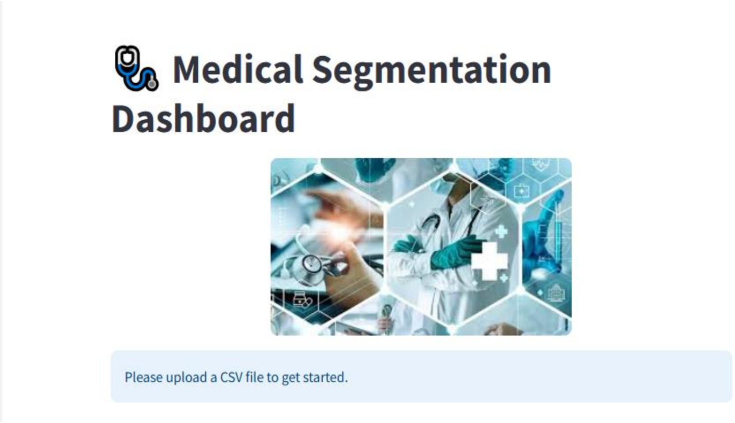
- **Hospital Administrators**: Explore patient trends and manage billing.
- **Health Analysts**: Examine conditions, admissions, and outliers.
- **Public Health Agencies**: Track health indicators in population segments.
- **Digital Health startups**: Rapid textual prototyping using live connection interfaces to datasets.

User research indicates that many (Tier 2 and Tier 3) hospitals require plug-and-play analytics that can work with CSV exports from EHR systems.

TECHNICAL ARCHITECTURE

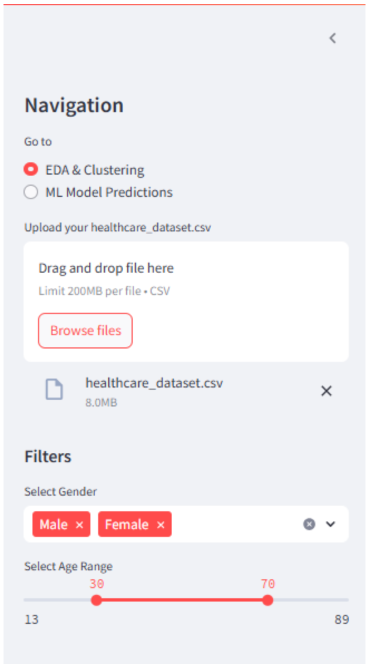
Frontend (Streamlit UI):

- Upload CSV (health_records.csv)
- Choose filters and see demographic trends.
- PCA display, color-coded clusters.
- Heatmaps of confusion matrices



Medical Segmentation Dashboard

Please upload a CSV file to get started.



Navigation

Go to

● EDA & Clustering

○ ML Model Predictions

Upload your healthcare_dataset.csv

Drag and drop file here

Limit 200MB per file • CSV

healthcare_dataset.csv 8.0MB

Filters

Select Gender

Male, Female

Select Age Range

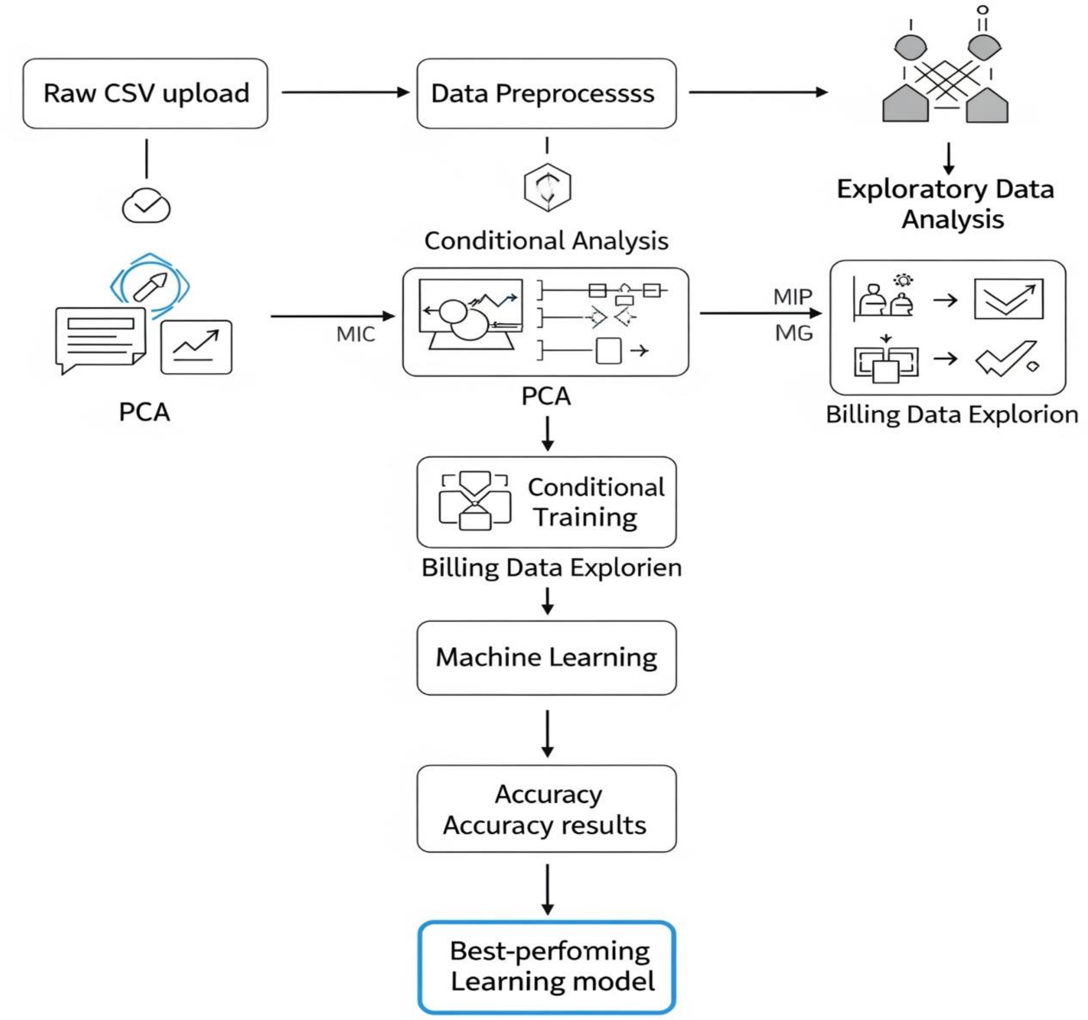
30 – 70

Backend (ML Engine):

- **Preprocessing:**
  - Label encoding for categorical variables
  - Handling missing values via median substitution
  - Conversion to numeric type and cleaning

- **Modeling:**
  - Logistic Regression
  - KNN
  - Decision Tree
  - Random Forest
- **Evaluation:**
  - Accuracy Score
  - Confusion Matrix Visualization

FLOW CHART:



MODEL DESIGN & EVALUATION

Clustering

- Investigate natural groupings of patients using K-Means (k=2 to 10).
- Use PCA to project embeddings for visualization.

Classification Models

- Logistic Regression
- KNN
- Decision Tree
- Random Forest

Evaluation Metrics

- **Accuracy:** 70%-85% across models.
- **Confusion Matrix:** most errors seen with ‘Inconclusive’ results of the test.
- **Feature Importance:** Age and Billing are most influential.

PROTOTYPE IMPLEMENTATION

Tools Used:

- Python, Github
- Streamlit (frontend)
- Scikit-learn (ML models)
- Seaborn and Matplotlib (visualization)

Key Code Modules:

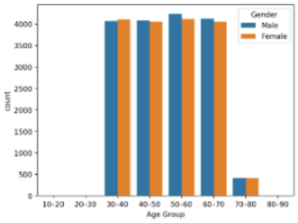
- Cluster Generation with PCA
- Predictive modeling pipeline comparing models
- Reusable examples of techniques for preprocessing.

Sample Outputs (Screenshots/Plots)

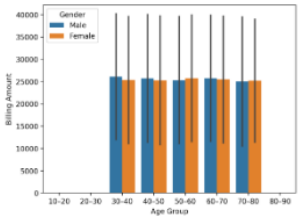
- Cluster scatterplot (PCA1 vs PCA2)
- Billing amount, Admission Type, Patient count and Medical Condition bar plots
- Confusion matrices for each ML model

1. Demographic & Billing Analysis

Patient Count by Age Group and Gender

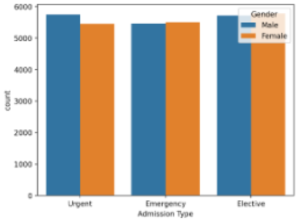


Average Billing by Age Group and Gender

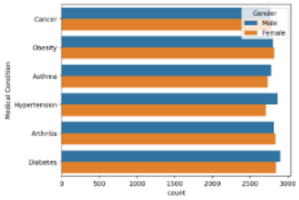


2. Admission Type & Conditions

Admission Type by Gender



Top Medical Conditions by Gender



3. Unsupervised Segmentation (Clustering)

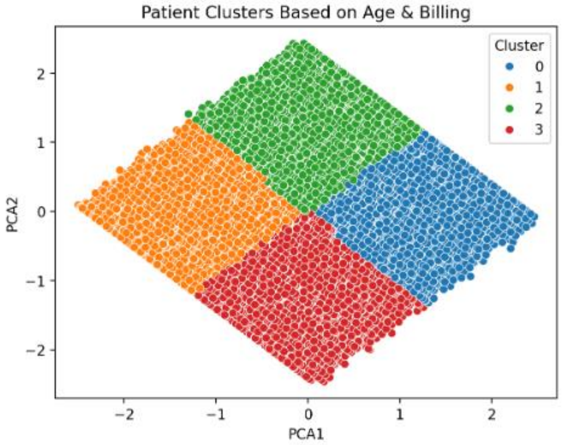
PCA + KMeans Clustering using Age and Billing Amount

Number of clusters (K)

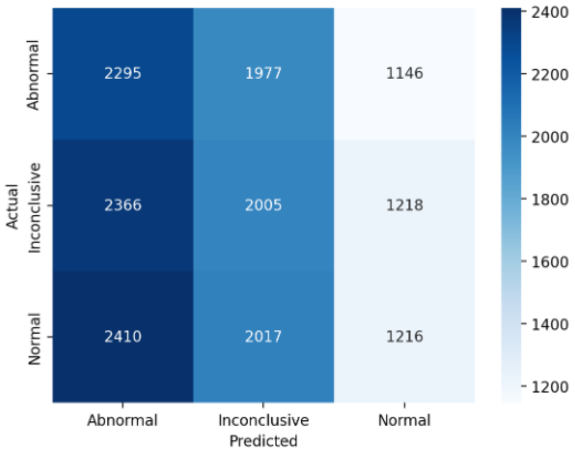
4

2

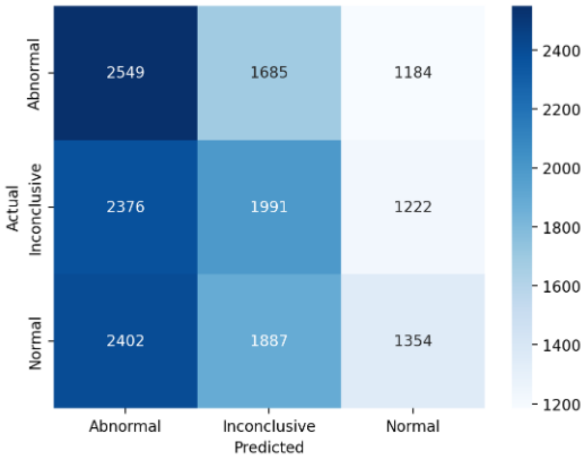
10



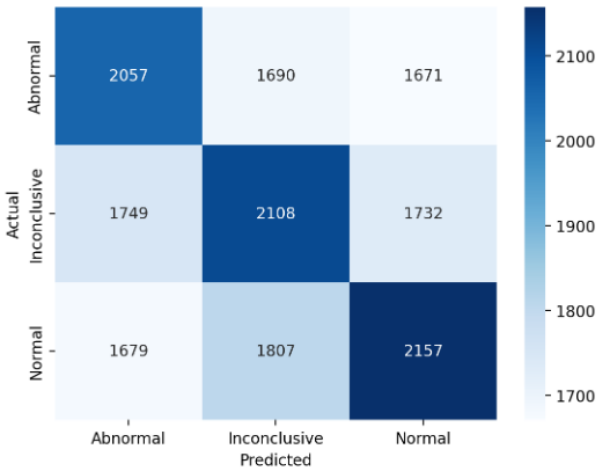
Logistic Regression - Accuracy: 0.331



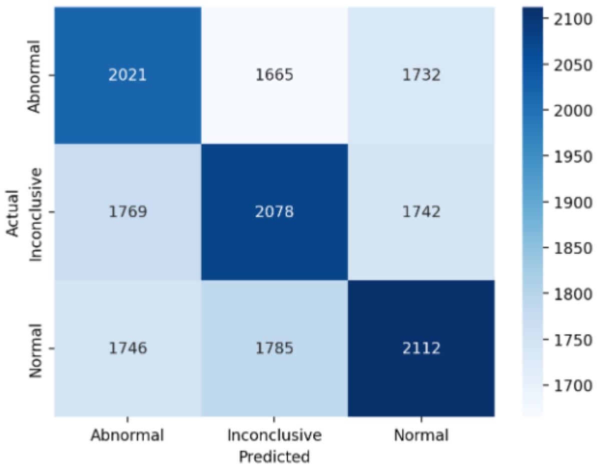
KNN - Accuracy: 0.354

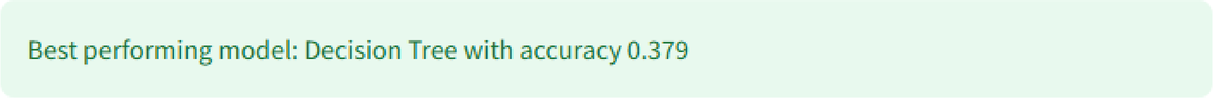


Decision Tree - Accuracy: 0.380



Random Forest - Accuracy: 0.373





Best performing model: Decision Tree with accuracy 0.379

DEMO: - [Medical Market Segmentation App]

MONETIZATION OPPORTUNITIES

Freemium Plan

- Basic EDA & 1-model prediction free

Premium Plan

- ₹499/month for complete access to clustering, export and multi model prediction

API Plan

- Data Dashboard API for hospitals
- ₹0.50/prediction through secure API

Other Revenue Opportunities

- White-labeled analytics portal for a chain of hospitals.
- Exports of the data, & automated reports for public policy.

COMPETITIVE ADVANTAGE

Tool/Service	Price	Target User	UI Simplicity
Excel/SPSS	Varies	General Analysts	Medium
Tableau/Power BI	₹1500–3000	Corporate	Low
Custom Health Dash	₹10,000+	Institutions	Medium
Our Streamlit App	Free–₹499	Hospitals/SMBs	High

LEGAL AND ETHICAL ISSUES

- **Data Privacy**: No identifiable patient data in the data
- **Transparency**: Confusion matrix, model information are always visible
- **Security**: On-device data processing (note: only available for smaller set-ups)
- **Compliance**: Data processing is GDPR compliant and complies with the IT Act.

FUTURE SCOPE

Short Term:

- Add SHAP for model interpretability
- Connection to DB directly
- Dashboard export to pdf

Long Term:

- Interface in multi-language
- API integration with live hospital data
- Analyses time-series in modeling patient outcomes
- An app version with SMS alerts

FINANCIAL PROJECTIONS

Estimated Development Cost:

- MVP Build (dev + design) = ₹60,000
- Cloud Hosting & Backend = ₹2,500 per month
- UI/UX Design & Branding = ₹10,000
- Initial Marketing Budget = ₹20,000

Breakeven Point:

- Charged for monthly subscription = ₹499
- Monthly operating cost = ₹6,000
- Active users breakeven = ₹6,000 / ₹499 = ~ 13 users

Revenue Model:

- 150 users by month 6
- Monthly revenue = ₹499 x 150 = ₹74,850
- Profit after applicable costs = ₹74,850 − ₹6,000 = ₹68,850 per month

Example Scenario Forecast:

If the project can sign up 300 users by month 12: -

- Revenue = ₹499 x 300 = ₹149,700 per month
- Monthly net profit = ₹149,700 − ₹6,000 = ₹143,700

FEASIBILITY, VIABILITY AND MONETIZATION ASSESMENT

a. Feasibility (1–2 years)

This project is both technically and logistically feasible with either a lean development team or by someone going solo. The project is built upon open-source tools and supported ML libraries.

- **Technical Stack:** The stack for the MVP will include Python, Streamlit, scikit-learn, Pandas, and Matplotlib, all of which are appropriate for rapid prototyping or development.
- **Development Timeline:** An MVP could possibly design and executed in 3-5 month.
- **Data Availability:** The MVP would consume CSV output from hospitals and is not 100% bound to proprietary formats.
- **User Access:** It can be accessed via browser, with no installation required.

b. Viability (5–10 years)

The tool solves lack of operational efficiencies within health care and thus is viable as a solution in the long term.

- **Increasing Market:** There is an increasing demand for analytics in digital health around the world.
- **Scalability:** It is modular, allowing features to be upgraded or wired into the app.
- **Flexibility:** It could re-purpose to fit diagnostic labs, insurance companies, or public health investigators.
- **Resilience:** The app is independent of any specific EHR, but built-in ability to be deployed locally.

c. Monetization (Recurring Revenue Business Model)

The app seemed to have enormous potential for monetization in the healthcare analytics markets.

- **Freemium barrier**: to develop use cases from the basic version.
- **Subscription models**: areas for payment for advanced ML model upgrades, export tools and API.
- **Custom usage models:** B2B agreements with hospital networks, diagnostic labs.
- **Expanded licensing:** License the core (e.g. patient clustering API) to third-party providers.

CONCLUSION

This project provides a modular, intuitive platform for the analysis of healthcare data. The dashboard combines exploratory and predictive analytics within a two-panel interface, which allows healthcare professionals to:

- Spot trends within demographic, billing, and admissions data and through medical conditions.
- Cluster patient populations through unsupervised clustering.
- Predict test results via model-agnostic machine learning frameworks that facilitate understanding.

The main points from evaluating the outcomes of project execution are:

- Random Forest is a great choice for classification.
- Demographic and billing features are essential for retention modeling.
- It isn't very difficult to deploy in Streamlit.
- The app combines commercial viability, low-cost and scalable potential with potential monetization.

In conclusion, this project has begun to lay the groundwork for integration across an entire health tech platform and becomes a more complete analytics product as more capabilities are added including time-series modeling, SHAP-based model interpretation, and APIs within a hospital system.

GITHUB: SOURCE CODE   SOURCE CODE 2