HEALTH AI SUITE-
Intelligent
Analytics for Patient Care

Project by
ASHWITHA K K

TOPICS

1.Introduction

The Health AI Suite is a comprehensive, end-to-end healthcare intelligence platform that integrates multiple Artificial Intelligence (AI) and Machine Learning (ML) techniques to support early disease detection, health risk prediction, patient monitoring, and clinical decision support. The primary objective of this project is to transform heterogeneous healthcare data—ranging from structured patient records to medical images, time-series vitals, and patient feedback—into actionable clinical insights.

The suite incorporates machine learning–based risk level prediction, where patient demographic details, lifestyle factors, vital signs, and medical history are analysed to classify individuals into Low, Medium, or High health risk categories. In addition, associative learning techniques are used to identify hidden relationships and co-occurrence patterns between diseases, enabling deeper understanding of comorbidities and health trends.

To address medical imaging use cases, the Health AI Suite includes a Convolutional Neural Network (CNN)–based chest X-ray analysis module for automated detection and classification of respiratory and lung-related abnormalities. This module assists in supporting radiological assessment by learning spatial features from X-ray images, enabling faster and more consistent diagnostic insights.

For continuous patient monitoring, the platform integrates **time-series analysis using Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models**. These models analyse sequential vital signs such as heart rate, blood pressure, oxygen saturation (SpO_2), and respiratory rate to capture temporal patterns and predict patient health deterioration over time.

Furthermore, the Health AI Suite incorporates **sentiment analysis on patient feedback and healthcare reviews** using Natural Language Processing (NLP) techniques. This module evaluates patient sentiment to measure satisfaction levels, identify service quality gaps, and support data-driven improvements in healthcare delivery.

The system is implemented using a modern and scalable architecture, with **Python-based AI models**, **Fast API for backend services**, and an **interactive Stream lit dashboard** for real-time predictions and visual analytics. Model performance is evaluated using standard metrics such as **accuracy, precision, recall, F1-score**, along with **support, confidence, and lift** for associative rule mining.

By unifying predictive analytics, medical imaging, time-series modelling, and sentiment analysis within a single platform, the Health AI Suite demonstrates the practical application of AI in building **intelligent, proactive, and patient-centric healthcare systems**.

2.Problem Statement

Modern healthcare systems generate vast and diverse amounts of data, including electronic health records, medical images, continuous vital sign measurements, and patient feedback reviews. However, these data sources are often analysed in isolation, using manual or rule-based approaches that are time-consuming, error-prone, and insufficient for early detection of health risks and disease progression.

Healthcare professionals face significant challenges in:

- Early identification of patient health risks
- Accurate interpretation of medical imaging such as chest X-rays
- Continuous monitoring of patient vitals over time
- Understanding patient satisfaction and feedback at scale
- Discovering hidden disease associations and comorbidities

Existing systems lack a unified, intelligent platform that can seamlessly integrate predictive analytics, deep learning, time-series modelling, and natural language processing to provide holistic and actionable insights. As a result, critical warning signs may go unnoticed, diagnostic processes may be delayed, and opportunities for preventive care and service improvement may be missed.

Therefore, there is a strong need for a scalable, AI-driven healthcare solution that can:

- Predict patient risk levels using clinical and lifestyle data
- Automatically analyse chest X-ray images for disease detection
- Monitor and predict patient health deterioration using time-series vital data
- Analyse patient feedback sentiment to improve healthcare quality
- Provide real-time, explainable insights through interactive dashboards

The Health AI Suite addresses these challenges by integrating multiple AI techniques into a single platform, enabling data-driven clinical decision support, early intervention, and improved patient outcomes.

3.Business Use Cases

1. Patient Health Risk Assessment & Preventive Care
Healthcare providers can use the Health AI Suite to predict patient risk levels (Low, Medium, High) based on demographic data, lifestyle factors, and clinical parameters.
Business Value:

- Reduced treatment costs through early intervention
- Improved patient care quality
- Better resource allocation for high-risk cases

2. Length of Stay (LOS) Prediction
Hospitals face challenges in accurately estimating patient Length of Stay, leading to bed shortages, inefficient staff allocation, increased costs, and delayed admissions.
Business Value

- Optimized bed utilization and patient flow
- Improved staff and resource planning
- Reduced operational costs
- Faster admissions and discharge planning

3. Patient Clustering
Healthcare providers struggle to manage diverse patient populations using uniform treatment strategies, leading to inefficient care delivery and resource utilization.
Business Value

- Improved clinical decision-making
- Targeted interventions and preventive care
- Better operational efficiency
- Enhanced patient outcomes

4. Associative Mining
Healthcare organizations often lack visibility into hidden relationships between diseases, symptoms, and patient conditions, making it difficult to identify comorbidities and plan preventive care.
Business Value

- Improved clinical insights and decision support
- Early detection of high-risk disease combinations
- Data-driven care pathways and treatment recommendations

5. CNN-Based Chest X-ray Detection
Manual interpretation of chest X-rays is time-consuming and subject to variability, especially in high-volume hospitals where early detection of lung diseases is critical.
Business Value

- Faster and more consistent X-ray analysis
- Reduced workload for radiologists
- Early detection and timely treatment
- Improved diagnostic support

6. LSTM Time Series Analysis
Continuous patient vital data is difficult to monitor manually, making it challenging to detect early signs of health deterioration in real time.
Business Value

- Early detection of critical conditions
- Reduced ICU readmissions and mortality risk
- Improved patient safety and care quality
- Reduced clinician monitoring burden

7. Feedback Sentiment Analysis
Healthcare organizations receive large volumes of patient feedback and reviews, making it difficult to manually analyse sentiment and identify service quality issues in a timely manner.

Business Value

- Faster insight into patient experience
- Improved healthcare service quality
- Enhanced patient engagement and trust
- Better decision-making for management

4.Data Overview

The Health AI Suite utilizes multi-source healthcare data to support predictive modelling, medical imaging analysis, time-series forecasting, and sentiment analysis. The datasets used in this project are a combination of synthetic and publicly available healthcare data, structured to closely resemble real-world clinical scenarios while ensuring data privacy.

1. Patient Clinical & Lifestyle Data

This dataset contains structured patient information used for risk level prediction, clustering, associative mining, and length of stay estimation.

Key Attributes:

- Demographics: Age, Gender
- Lifestyle Factors: Diet, Exercise Frequency, Smoking, Alcohol Consumption, Stress Levels
- Clinical Measurements: BMI, Blood Pressure, SpO_2, Heart Rate, Respiratory Rate
- Medical History: Family History, Chronic Conditions
- Target Variables: Risk Level, Length of Stay

DataType:Structured(Tabular)

Usage: Supervised & unsupervised ML models

2. Chest X-ray Imaging Data

Chest X-ray images are used for **CNN-based medical image classification and detection** of respiratory and lung-related abnormalities.

Key Characteristics:

Grayscale chest X-ray images

Labelled disease categories (e.g., normal vs abnormal)

Pre-processed using resizing, normalization, and augmentation

Datatype: Unstructured(ImageData)

Usage: Deep learning using CNN architectures

3. Time-Series Vital Signs Data

Sequential patient vital sign data is used for **LSTM/RNN-based time-series analysis** to detect trends and predict patient health deterioration.

Key Attributes:

- Time-stamped measurements
- Heart Rate
- Blood Pressure (Systolic & Diastolic)
- SpO_2
- Respiratory Rate

DataType: Sequential/Time-Series

Usage: LSTM & RNN models

4. Patient Feedback & Review Data

Unstructured text data consisting of patient feedback and reviews is used for **sentiment analysis**.

Key Attributes:

- Free-text patient comments
- Service-related feedback
- Satisfaction indicators

DataType: Unstructured(TextData)
Usage: NLP-based sentiment classification

5. Data Privacy & Security Considerations

- Use of anonymized and synthetic data
- No personally identifiable information (PII)
- Designed to align with healthcare data protection standards

5.Machine Learning

Machine Learning (ML) is a subset of Artificial Intelligence (AI) that enables computer systems to automatically learn patterns from data and make predictions or decisions without being explicitly programmed. Instead of relying on predefined rules, machine learning models improve their performance over time by analysing historical data and identifying underlying relationships within it.

Why Machine Learning is Important in Healthcare

Machine Learning (ML) plays a vital role in modern healthcare by enabling systems to analyse complex medical data, identify hidden patterns, and support data-driven decision-making. With the rapid growth of electronic health records, medical imaging, continuous patient monitoring, and patient feedback data, traditional rule-based approaches are no longer sufficient to extract meaningful insights.

Within the Health AI Suite, machine learning techniques are used to build predictive and analytical models such as patient risk level classification, length of stay prediction, patient clustering, and disease association mining. Supervised learning algorithms are applied to predict outcomes like risk categories and hospitalization duration, while unsupervised learning methods are used to identify patient groups with similar characteristics. Additionally, evaluation metrics such as accuracy, precision, recall, F1-score, support, confidence, and lift are used to assess model performance and reliability.

By integrating machine learning into healthcare workflows, the Health AI Suite demonstrates how data-driven intelligence can enhance operational efficiency, enable preventive care, and improve overall patient outcomes.

5.1 Risk Stratification (Classification)

Risk Stratification is a machine learning–based classification approach used to categorize patients into predefined risk groups such as Low, Medium, and High risk based on their clinical, lifestyle, and demographic data. This process helps healthcare providers identify high-risk individuals early and prioritize medical interventions.
In the Health AI Suite, risk stratification is implemented using supervised machine learning classification algorithms trained on patient health records. The model learns patterns from historical data and predicts the risk level of new patients using key features such as age, BMI, blood pressure, lifestyle habits, stress levels, and family medical history.

Key Points

- Classifies patients into Low, Medium, and High-risk categories
- Uses supervised ML algorithms (e.g., Random Forest, Logistic Regression)
- Analyses demographic, lifestyle, and clinical features
- Supports early intervention and preventive healthcare
- Improves clinical decision-making and patient prioritization
- Deployed through Fast API and visualized using Streamlit
- Model performance evaluated using accuracy, precision, recall, and F1-score

Data Preprocessing & EDA Analysis:

Data preprocessing is a critical step to ensure data quality, consistency, and reliability before applying machine learning models. In the Health AI Suite, multiple preprocessing techniques are applied to transform raw healthcare data into a model-ready format.

Data Cleaning

- Identified and handled missing values using appropriate strategies (mean/median imputation or removal).
- Removed duplicate records to avoid biased learning.
- Corrected inconsistent or invalid values in clinical measurements.

Handling Categorical Variables

- Converted categorical features such as diet, smoking status, alcohol consumption, and family history into numerical form using encoding techniques.
- Ensured consistent encoding between training and deployment environments**.**

Feature Scaling & Normalization

- Applied feature scaling to numerical attributes such as age, BMI, blood pressure, and vitals.
- Normalization helped improve model convergence and ensured equal feature contribution

Outlier Detection & Treatment

- Identified outliers in clinical parameters like BMI and blood pressure using statistical techniques.
- Reduced the impact of extreme values to improve model robustness.

Feature Selection

- Selected clinically relevant features contributing most to risk prediction and LOS estimation.
- Removed redundant or low-importance features to reduce model complexity.

Data Splitting

- Split the dataset into **training and testing sets** to evaluate model generalization.
- Maintained class balance where applicable.

Exploratory Data Analysis (EDA) Steps

Exploratory Data Analysis (EDA) was performed to understand data distribution, identify patterns, and uncover relationships between variables before model development.

Understanding Data Structure

- Examined dataset dimensions, feature types, and data distributions.
- Verified target variable balance for classification tasks.

Descriptive Statistics

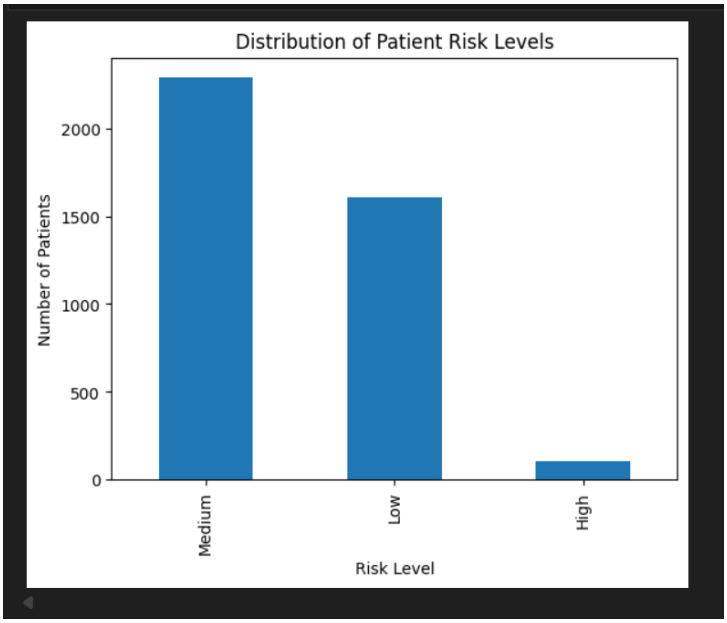
- Analysed mean, median, standard deviation, and ranges of numerical features.
- Helped understand patient health trends and variability.

```
df.describe()
```

	age	exercise_days	sleep_hours	bmi
count	4000.000000	4000.000000	4000.000000	4000.000000
mean	49.203500	3.454250	7.557500	27.740250
std	18.166593	2.322231	2.852232	7.186699
min	18.000000	0.000000	3.000000	15.000000
25%	34.000000	1.000000	5.000000	21.600000
50%	49.000000	3.000000	8.000000	27.900000
75%	65.000000	6.000000	10.000000	33.800000
max	80.000000	7.000000	12.000000	40.000000

Univariate Analysis

- Visualized distributions of individual features such as **age, BMI, blood pressure, and vitals**.
- Identified skewness and potential outliers.

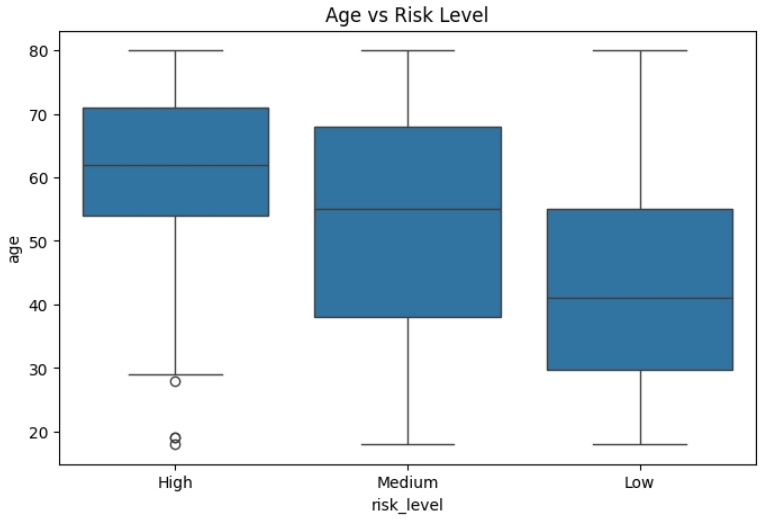


Low risk dominates (realistic population health data)

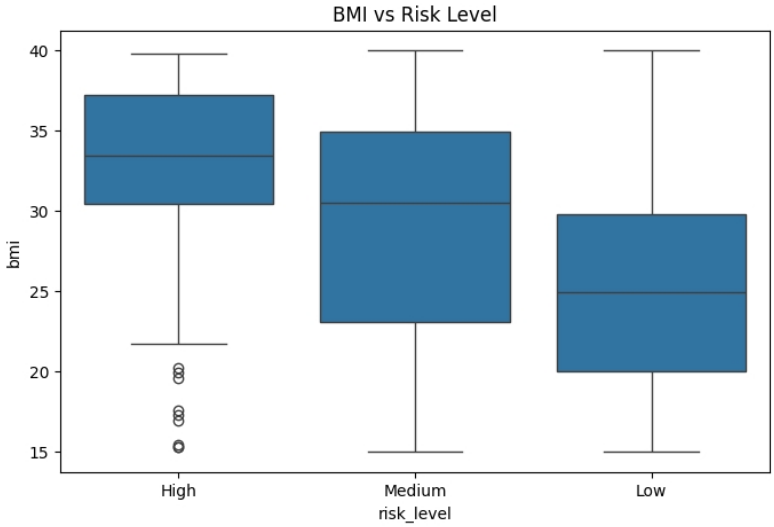
High-risk patients are rare → class imbalance

Bivariate & Multivariate Analysis

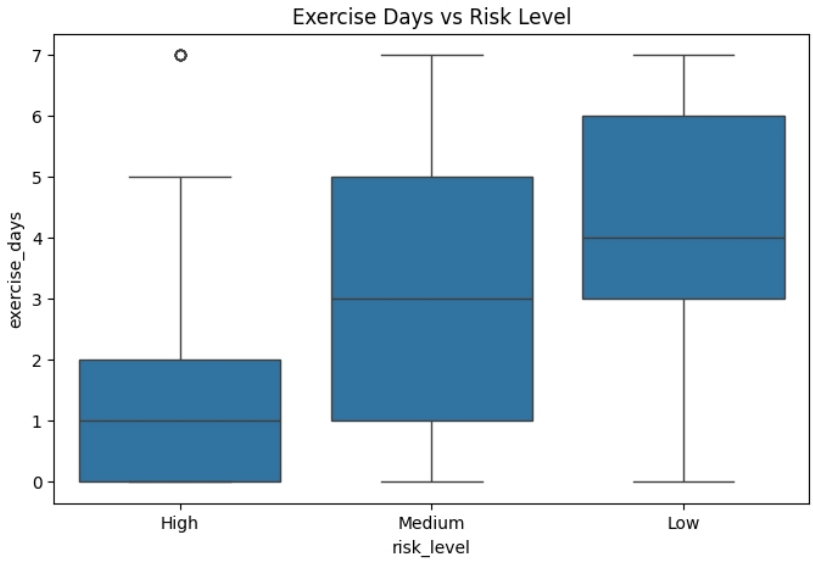
- Studied relationships between features and target variables (risk level, LOS).
- Used correlation analysis to identify influential predictors.



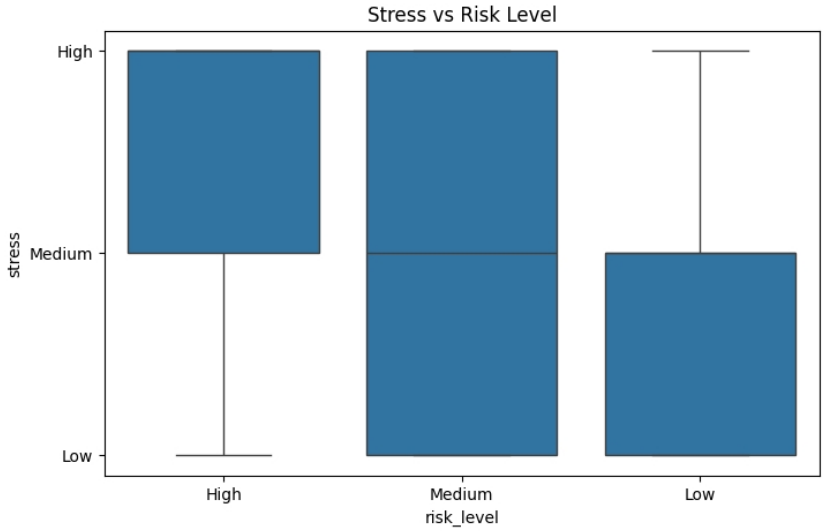
Median age increases from Low → Medium → High
Age is a strong risk indicator
Confirms model learning is medically meaningful



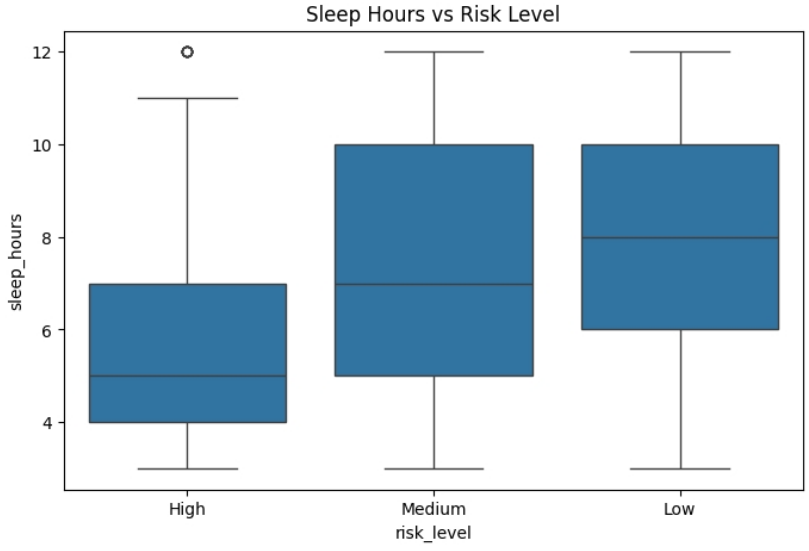
Higher BMI → higher risk
Obesity clusters in Medium & High risk
BMI is a key modifiable risk factor



As exercise days decrease → risk increases



High-risk patients predominantly show medium to high stress levels
Medium-risk patients display a wide stress range, indicating mixed mental health states
Low-risk patients mostly experience low to medium stress



High-risk patients sleep the least (median ≈ 5 hours), indicating sleep deprivation
Medium-risk patients show moderate sleep (≈ 6–7 hours) with high variability
Low-risk patients get the most consistent and adequate sleep (≈ 7–8 hours)

MODEL DEVELOPMENT

Model development focuses on building a reliable machine learning–based classification System to predict patient health risk levels as Low, Medium, or High using processed clinical and lifestyle data.

Model Selection

Evaluated multiple supervised classification algorithms.
Selected Random Forest Classifier due to:

- High performance on structured healthcare data
- Ability to handle non-linear relationships
- Robustness to noise and outliers
- Built-in feature importance estimation

Model Evaluation

Evaluated model performance on unseen test data using:

- Accuracy
- Precision
- Recall
- F1-score

Ensured reliable classification across all risk categories.

Model Insights

- The model accurately distinguishes patients at risk level.
- F1-score: 98.32% indicates strong classification performance.

Outcome

- Successfully developed a scalable and accurate risk stratification model.
- Enabled real-time patient risk prediction for clinical decision support.

Streamlit Dashboard Implementation

The Streamlit dashboard serves as the **user-facing interface** of the Health AI Suite, enabling real-time interaction with machine learning models through a simple and intuitive web application.

Objective

- Provide an interactive platform for users to input patient data.
- Display real-time predictions and insights generated by AI models.
- Enable easy visualization of healthcare analytics without technical expertise.

Dashboard Design

- Built using **Streamlit**, a Python-based web framework.
- Designed with a clean, responsive, and user-friendly layout.
- Organized into sections for data input, prediction output, and visualization.

User Input Module

- Allows users to enter patient information such as:
  - Age, BMI, lifestyle habits
  - Vital signs and medical history
- Input validation ensures accurate and complete data entry.

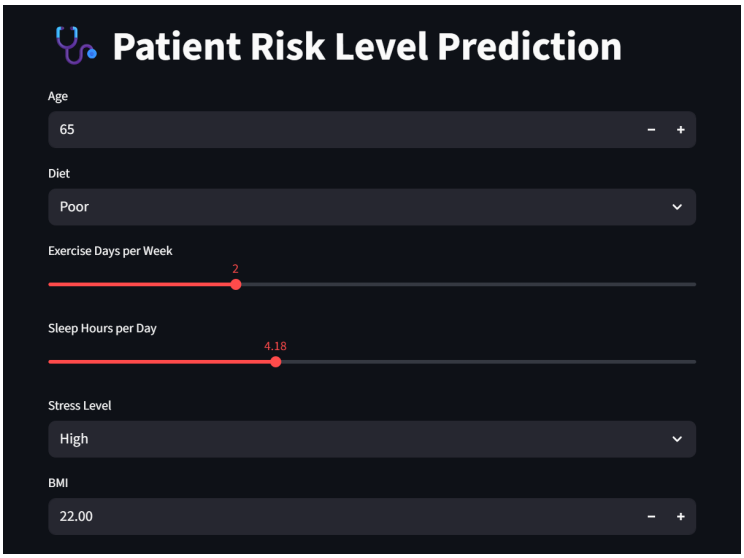
Model Integration

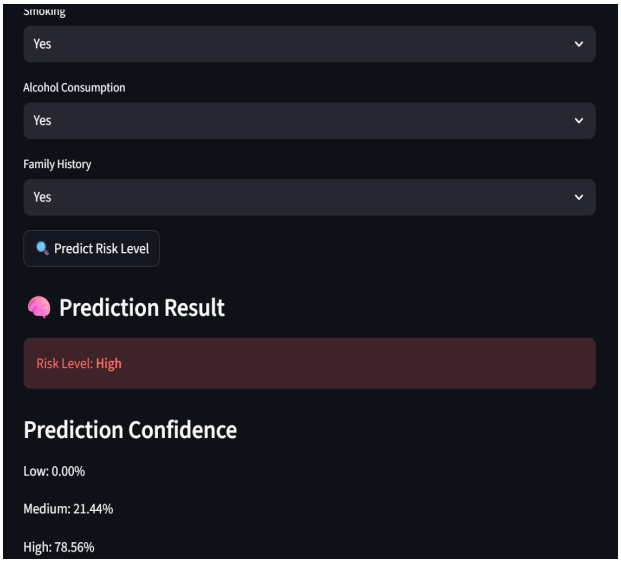
- Loads the trained **risk level prediction model** using pickle/joblib.
- Ensures feature alignment between input data and trained model.
- Sends processed input data to the model for inference.

Prediction & Output Display

- Displays predicted **Risk Level (Low / Medium / High)** in real time.
- Provides confidence or probability-based insights where applicable.
- Uses visual indicators (text, colors, alerts) to improve interpretability.

Screenshot of Streamlit:





Outcome

- Successfully implemented an interactive dashboard for real-time healthcare prediction.
- Improved accessibility and usability of AI-driven healthcare insights.

5.2 Length of Stay Prediction (Regression)

Length of Stay (LOS) prediction aims to estimate the expected number of days a patient will **remain hospitalized** based on clinical, demographic, and admission-related data. Accurate LOS reduction supports efficient hospital operations and improved patient care planning.

Key Points

- Predicts the expected number of days a patient will stay in the hospital
- Formulated as a supervised regression problem
- Uses patient demographics, clinical vitals, diagnosis, and admission details
- Helps in bed availability planning and discharge management
- Machine learning models capture non-linear relationships in healthcare data
- Performance evaluated using MAE, RMSE, and R^2 score
- Identifies key factors contributing to prolonged hospital stays
- Deployed through Fast API and visualized using Streamlit dashboard
- Supports data-driven hospital operations and cost optimization

Data Preprocessing & EDA Analysis

Data Preprocessing

Data preprocessing ensures that raw healthcare data is clean, consistent, and suitable for machine learning model development.

Key Steps:

- Removed duplicate and inconsistent patient records
- Handled missing values using appropriate statistical methods
- Encoded categorical variables into numerical format
- Scaled and normalized numerical features for uniformity
- Detected and treated outliers in clinical parameters
- Selected relevant features to reduce noise and complexity
- Split data into training and testing datasets

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to understand data patterns, distributions, and and relationships before model training.

Understanding Data Structure

- Examined dataset dimensions, feature types, and data distributions.
- Verified target variable balance for classification tasks.

Descriptive Statistics

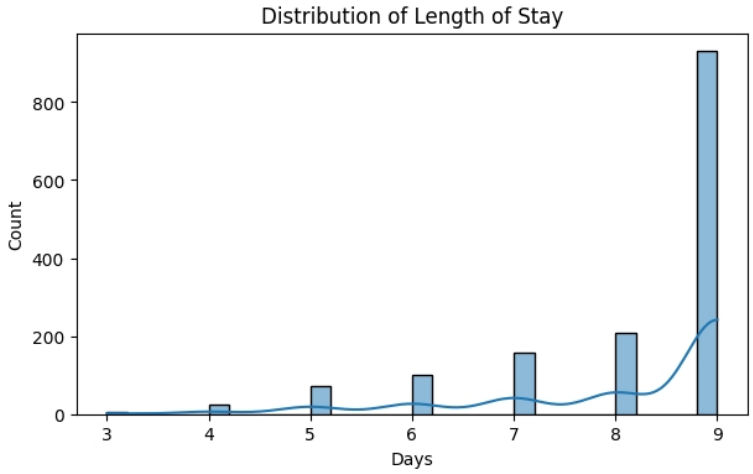
- Analysed mean, median, standard deviation, and ranges of numerical features.
- Helped understand patient health trends and variability.

```
df.describe()
```
✓ 0.0s

	age	severity	comorbidities	procedure_code	length_of_stay
count	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000
mean	54.966667	2.983333	2.521333	3.526000	8.152000
std	21.331219	1.390088	1.706699	1.708197	1.326927
min	18.000000	1.000000	0.000000	1.000000	3.000000
25%	36.000000	2.000000	1.000000	2.000000	8.000000
50%	56.000000	3.000000	3.000000	4.000000	9.000000
75%	74.000000	4.000000	4.000000	5.000000	9.000000
max	90.000000	5.000000	5.000000	6.000000	9.000000

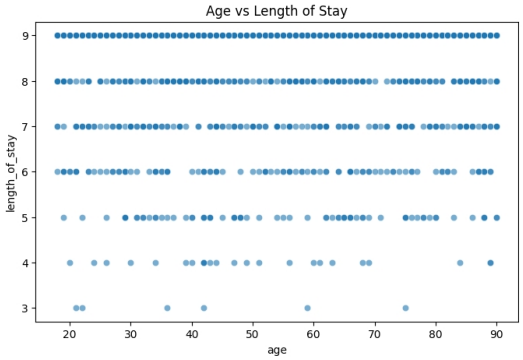
Univariate Analysis

- Visualized distributions of individual feature
- Identified skewness and potential outliers.

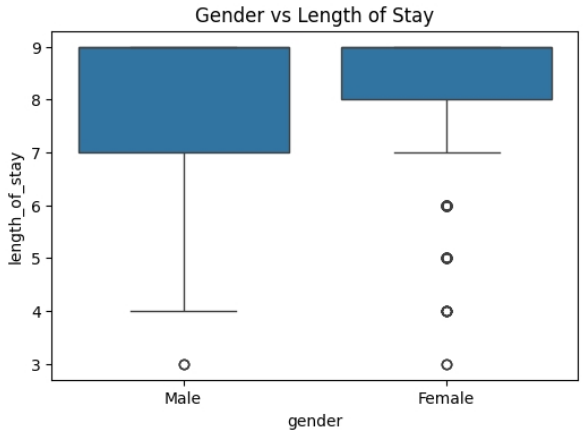


Bivariate & Multivariate Analysis

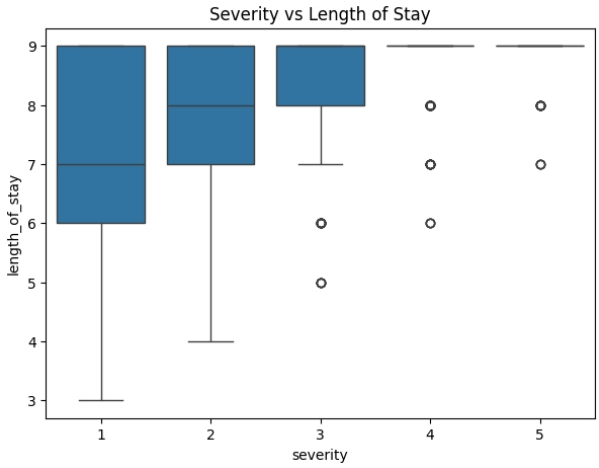
- Studied relationships between features and target variables (Length of stay).
- Used correlation analysis to identify influential predictors.



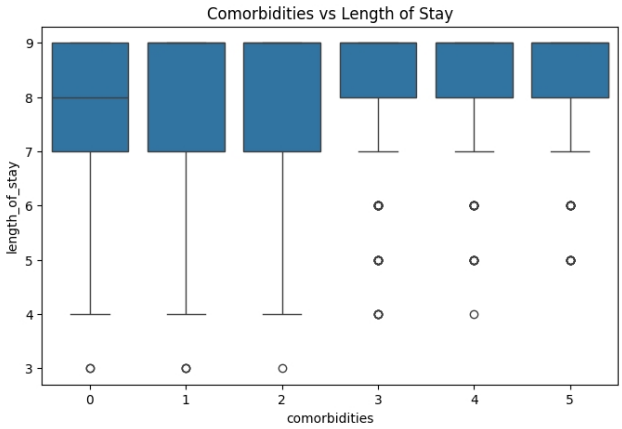
LOS increases with age
Elderly patients require longer monitoring
Age is a strong predictor



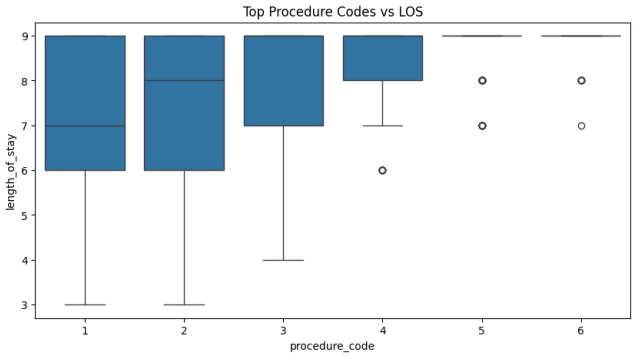
Slight LOS variation across gender
Gender alone is not a dominant predictor
Works best when combined with severity/comorbidities



Clear monotonic trend:
Low → Medium → High severity → Higher LOS
Severity is one of the strongest predictors



LOS increases with number of comorbidities
Multi-morbid patients require extended care
Strong clinical relevance



Complex procedures lead to longer hospitalization
Procedure complexity directly affects LOS

MODEL DEVELOPMENT

The model development phase focuses on building a **machine learning–based regression model** to accurately predict the hospital length of stay (LOS) for patients using clinical and admission -related data.

Model Selection

Evaluated multiple regression algorithms.
Chosen models include:

- Linear Regression (baseline)
- Random Forest Regressor
- Gradient Boosting Regressor

Tree-based models were preferred for capturing non-linear patterns.

Model Evaluation

- Assessed model performance using:
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)
  - R^2 Score
- Compared predicted LOS with actual LOS values.

Model Insights

- The model accurately predicts LOS for patients.
- R^2 82% indicates strong regression performance.

Outcome

- Successfully developed an AI-driven LOS prediction model.
- Enabled proactive hospital resource planning and discharge optimization.

Streamlit Dashboard Implementation

The Streamlit dashboard provides an interactive interface for predicting the expected hospital length of stay using the trained regression model.

Objective

- Enable users to input patient and admission details.
- Provide real-time LOS predictions to support hospital planning.

Dashboard Design

- Developed using **Streamlit** for rapid web application deployment.
- Clean and intuitive layout for easy navigation.
- Separate sections for input, prediction, and result display

User Input Module

- Accepts patient and admission parameters such as:
    - Age, gender
    - Vital signs
    - Diagnosis severity
    - ICU admission status
- Input validation ensures accurate data entry.

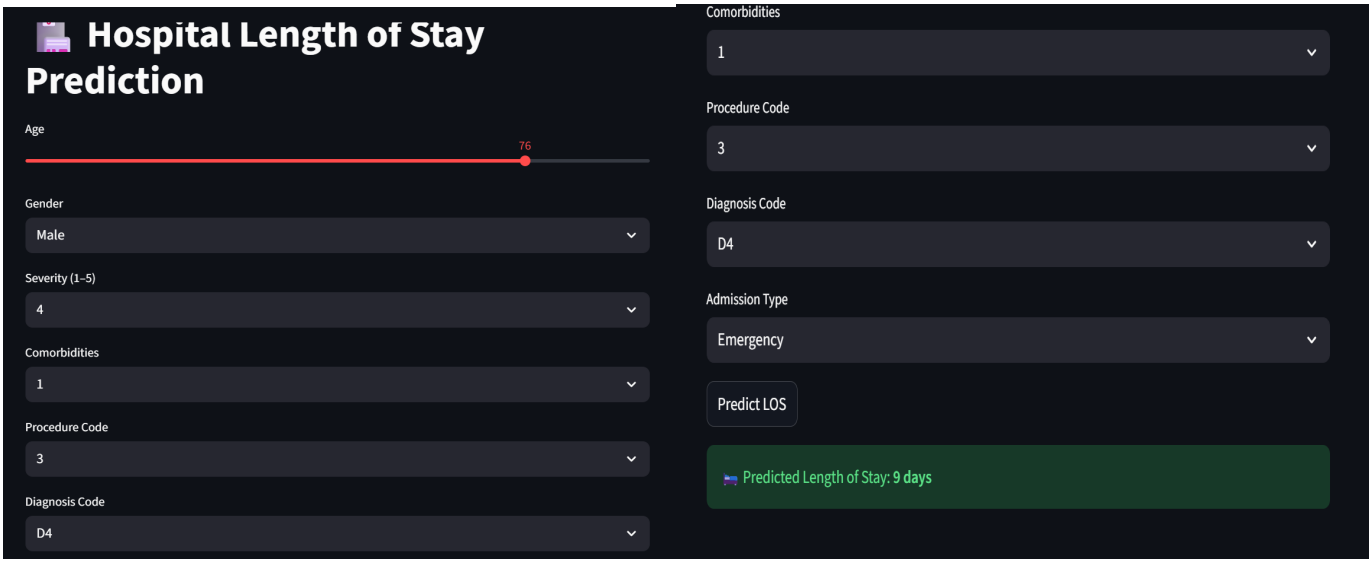
Model Integration

- Loads the trained **LOS regression model** using joblib/pickle.
- Ensures consistency between input features and model expectations.
- Processes inputs into model-ready format.

Prediction & Output Display

- Displays predicted **Length of Stay (in days)**.
- Uses clear numeric output and visual indicators.
- Helps clinicians and administrators understand expected hospitalization duration.

Screenshot of Streamlit:



Outcome

- Successfully implemented an interactive LOS prediction dashboard.
- Improved decision support for bed management and discharge planning.

5.3 Patient Segmentation (Clustering)

Patient Segmentation is an unsupervised machine learning approach used to group patients with similar clinical characteristics, lifestyle patterns, and health risk profiles. Unlike classification, clustering does not use predefined labels; instead, it discovers hidden structures within patient data.
In the Health AI Suite, patient segmentation helps in understanding population-level health Patterns and supports personalized care, targeted interventions, and efficient resource allocation.

Key Points

- Implemented using **unsupervised learning algorithms** such as K-Means and Hierarchical Clustering
- Groups patients based on similarities in:
    - Demographics (age, gender)
    - Lifestyle factors (diet, exercise, smoking, alcohol)
    - Clinical measurements (BMI, blood pressure, vitals)
- Helps identify **high-risk patient clusters** without predefined labels
- Supports **population health management and preventive care planning**
- Enables personalized treatment strategies for different patient groups
- Reduces healthcare cost through targeted interventions
- Visualized using cluster plots and summary statistics

Data Preprocessing

Data preprocessing is essential to improve data quality and ensure reliable machine learning model performance.

Key Steps:

- Removed duplicate and inconsistent patient records
- Handled missing values using appropriate statistical techniques
- Encoded categorical variables into numerical form
- Scaled and normalized numerical features for uniform contribution
- Identified and treated outliers in clinical measurements
- Selected relevant features to reduce noise and model complexity
- Split data into training and testing datasets

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand data characteristics and uncover meaningful Patterns.

Understanding Data Structure

- Examined dataset dimensions, feature types, and data distributions.
- Verified target variable balance for classification tasks.

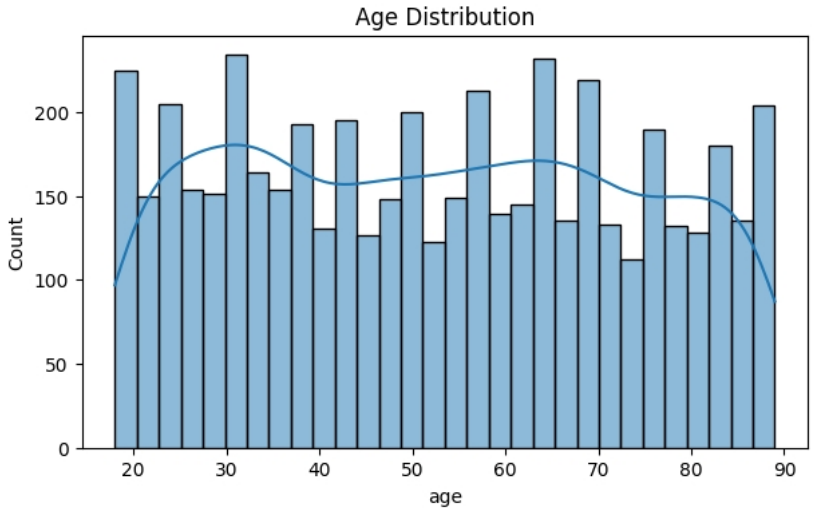
Descriptive Statistics

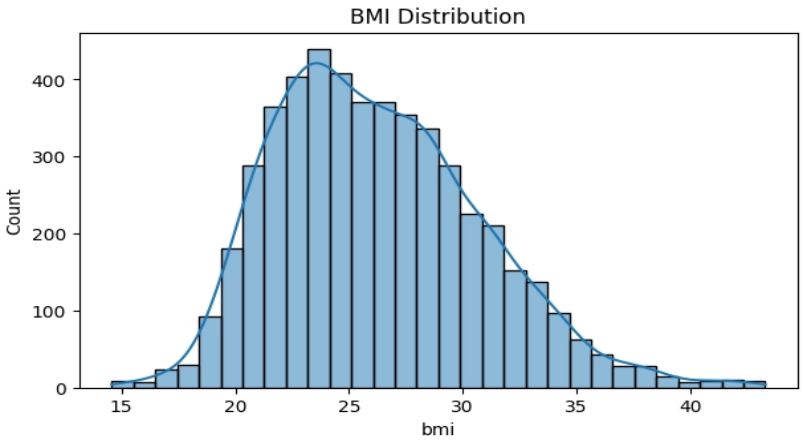
- Analysed mean, median, standard deviation, and ranges of numerical features.
- Helped understand patient health trends and variability.

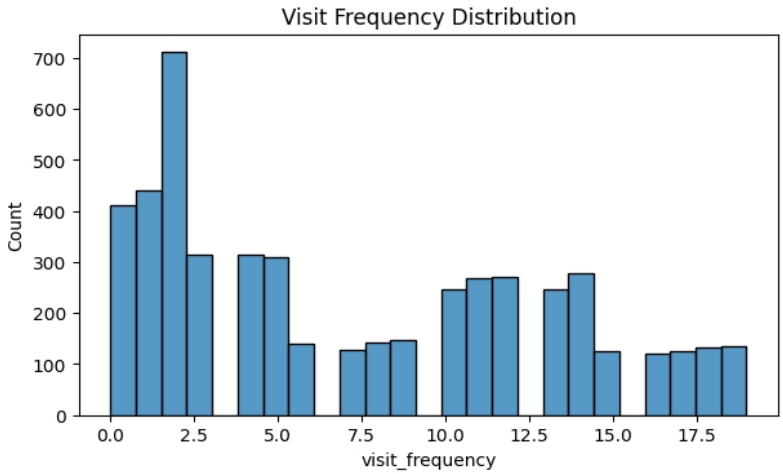
```
df.describe()
```
✓ 0.0s

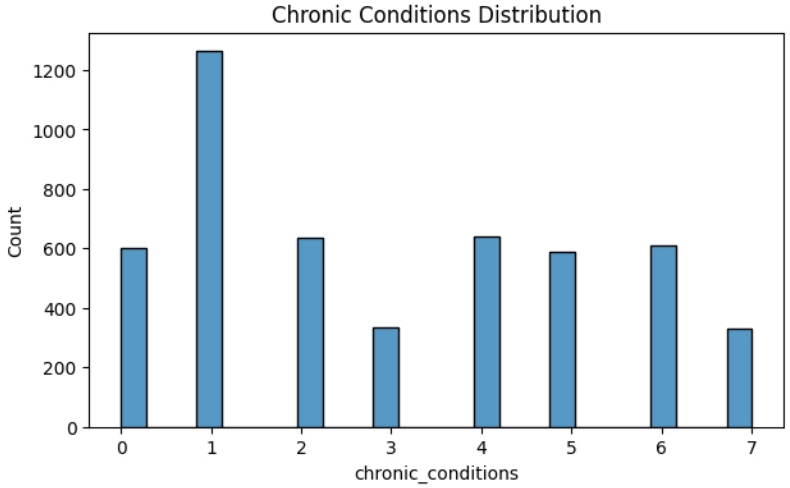
	age	bmi	chronic_conditions	visit_frequency	avg_stay_days	icu_admissions	emergency_visits
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000
mean	52.599600	26.256983	2.999400	7.277000	9.233600	0.322800	4.656600
std	20.807217	4.539981	2.229438	5.738437	8.149375	0.467594	4.142718
min	18.000000	14.567003	0.000000	0.000000	1.000000	0.000000	0.000000
25%	34.000000	22.853304	1.000000	2.000000	2.000000	0.000000	1.000000
50%	53.000000	25.740697	2.500000	5.500000	5.500000	0.000000	3.000000
75%	70.000000	29.136538	5.000000	12.000000	15.000000	1.000000	8.000000
max	89.000000	43.317137	7.000000	19.000000	29.000000	1.000000	14.000000

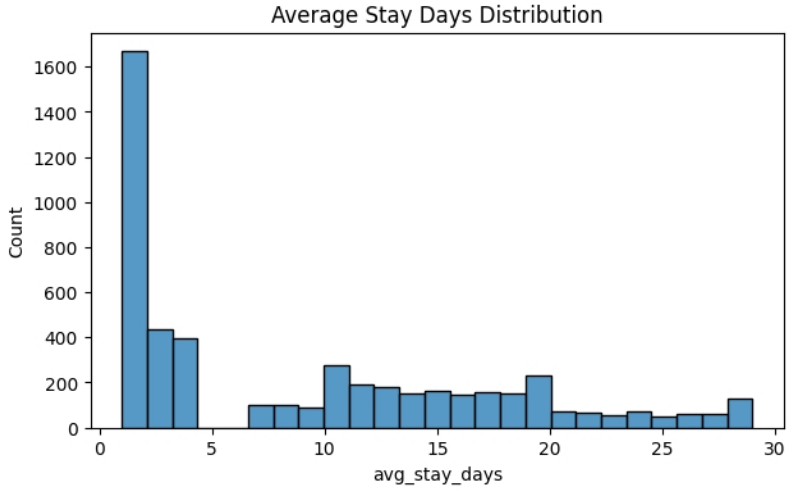
Univariate Analysis



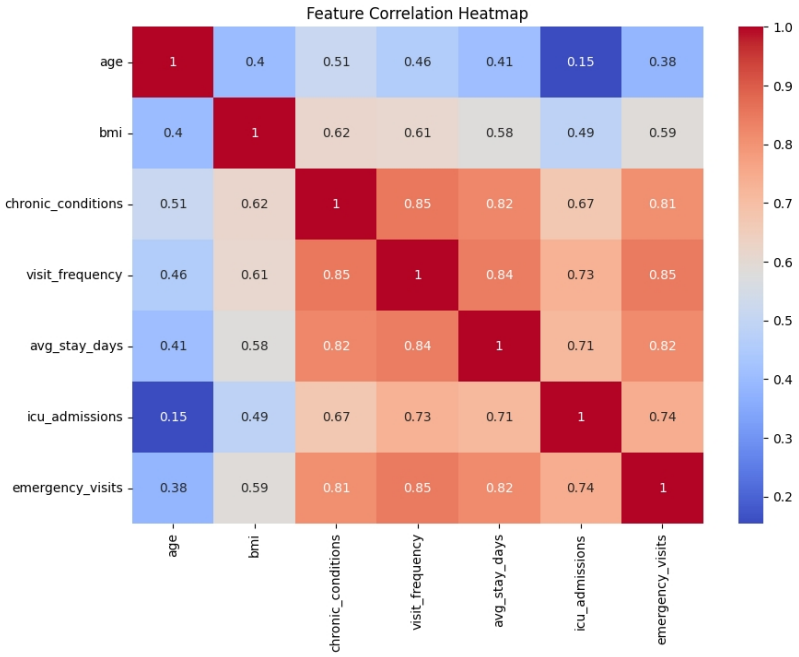




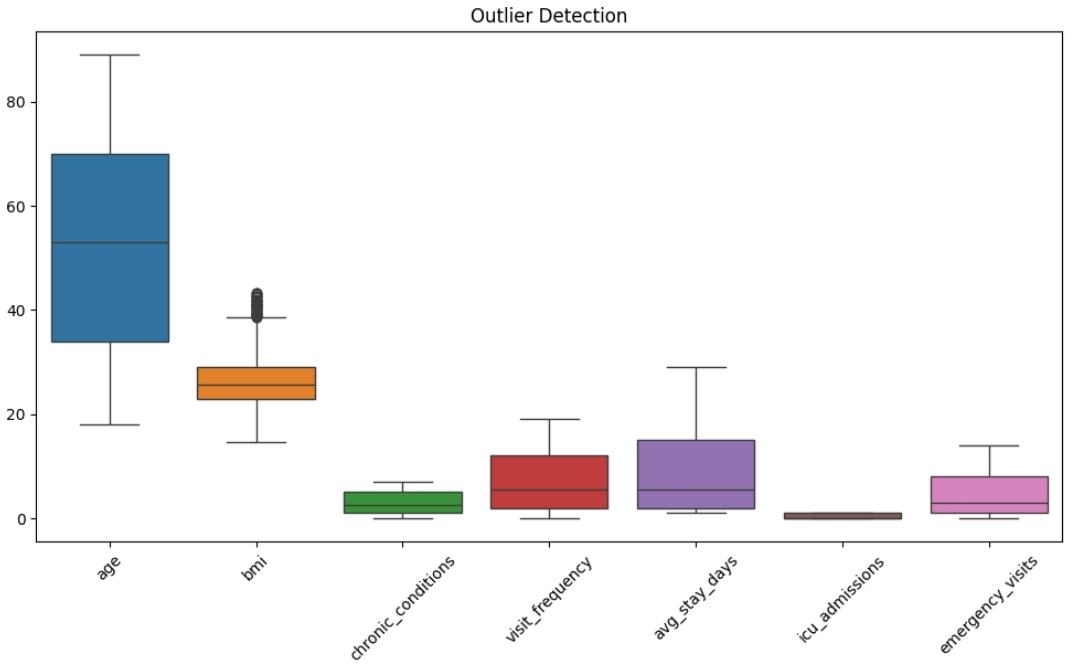




Bivariate & Multivariate Analysis



Moderate correlation between:
Age ↔ chronic conditions
Visit frequency ↔ avg stay
Low multicollinearity → good for clustering



Outliers exist but are clinically meaningful
Should NOT be removed for clustering
Represent high-risk/high-cost patients

Model Development

The model development phase focuses on building an unsupervised machine learning model to Group patients into meaningful segments based on similarities in their demographic, lifestyle, and clinical attributes.

Algorithm Selection

- Evaluated clustering algorithms including:
  - K-Means Clustering
  - Hierarchical Clustering
- Chose **K-Means** due to its simplicity, scalability, and effectiveness on structured healthcare data.

Optimal Cluster Determination

- Determined the optimal number of clusters using:
  - Elbow Method
  - Silhouette Score
- Ensured clusters were well-separated and meaningful.

Cluster Interpretation

- Analysed cluster centroids to understand patient characteristics.
- Identified high-risk and low-risk patient segments.
- Mapped clusters to clinical and lifestyle patterns.

Visualization

- Visualized clusters using:
  - Scatter plots
  - Dimensionality reduction techniques (PCA)
- Enhanced interpretability of patient groups

Model Evaluation Metrics

Final Cluster Mapping:
0: 'Young Healthy', 2: 'Middle-Aged Preventive', 3: 'Elderly Chronic', 1: 'High-Acuity'

- Silhouette scores:

K=2, Silhouette Score=0.518 K=3, Silhouette Score=0.466
K=4, Silhouette Score=0.369 K=5, Silhouette Score=0.358
K=6, Silhouette Score=0.353 K=7, Silhouette Score=0.351

Outcome

- Successfully segmented patients into distinct groups.
- Enabled targeted healthcare strategies and personalized interventions.

Streamlit Dashboard Implementation

The Streamlit dashboard provides an interactive interface to visualize and interpret **patient segmentation results** generated using unsupervised machine learning techniques.

Objective

- Enable users to explore patient clusters interactively.
- Visualize patient groupings based on clinical and lifestyle similarities.
- Support data-driven insights for personalized healthcare strategies.

Dashboard Design

- Developed using **Streamlit** for rapid web application deployment.
- Clean and intuitive layout for ease of navigation.
- Structured sections for data input, clustering results, and visualizations.

Data Input & Selection

- Allows users to select features for clustering (age, BMI, vitals, lifestyle factors).
- Supports uploading or selecting pre-processed datasets.
- Ensures consistent feature scaling before clustering.

Model Integration

- Loads the trained **clustering model** (e.g., K-Means).
- Applies the model to assign patients to clusters.
- Ensures real-time cluster assignment for new patient data.

User Interaction

- Allows dynamic selection of number of clusters.
- Updates visualizations and summaries in real time.
- Enhances exploratory analysis and decision-making.

Cluster Interpretation

- Shows summary statistics for each cluster.
- Describes key characteristics such as risk profile and lifestyle patterns.
- Helps identify high-risk or priority patient groups.

Screenshots

Outcome

- Successfully implemented an interactive clustering dashboard.
- Enabled intuitive exploration of patient segments for clinical and operational use.

5.4 Medical Associations (Association Rules)

Medical Association Rule Mining is a data mining technique used to discover hidden relationships And co-occurrence patterns among diseases, symptoms, and medical conditions. It helps identify how frequently certain conditions occur together and the strength of their relationships.

Key Points

- Implemented using association rule mining techniques such as Apriori
- Discovers frequent disease combinations and medical condition relationships
- Uses key metrics:
    - Support – frequency of disease occurrence
    - Confidence – likelihood of co-occurring diseases
    - Lift – strength of association beyond random chance
- Helps identify common comorbidities (e.g., diabetes–hypertension)
- Supports early screening and preventive healthcare planning
- Enhances clinical understanding of disease relationships
- Provides data-driven insights for population health analysis

Data Preprocessing & EDA Analysis

Data Preprocessing

Data preprocessing ensures the healthcare dataset is clean, consistent, and suitable for machine learning and deep learning models.

Key Steps:

- Removed duplicate and inconsistent patient records
- Handled missing values using statistical imputation techniques
- Encoded categorical variables into numerical representations
- Scaled and normalized numerical features to maintain uniformity
- Detected and treated outliers in clinical measurements
- Selected relevant features to reduce noise and model complexity
- Split data into training and testing datasets

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand data patterns, distributions, and relationships among variables.

Understanding Data Structure

- Examined dataset dimensions, feature types, and data distributions.
- Verified target variable balance for classification tasks.

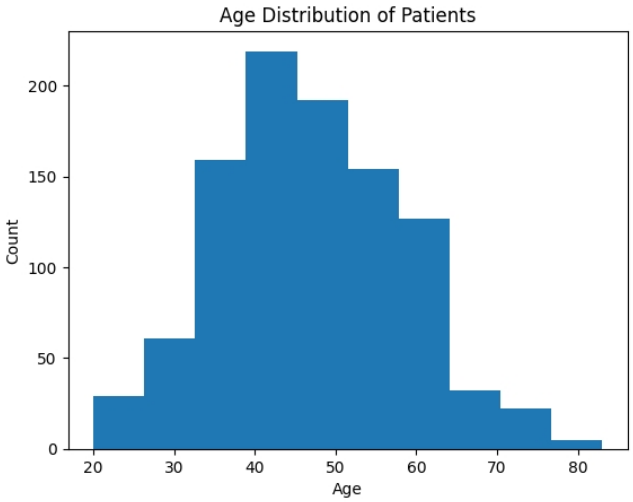
Descriptive Statistics

- Analysed mean, median, standard deviation, and ranges of numerical features.
- Helped understand patient health trends and variability.

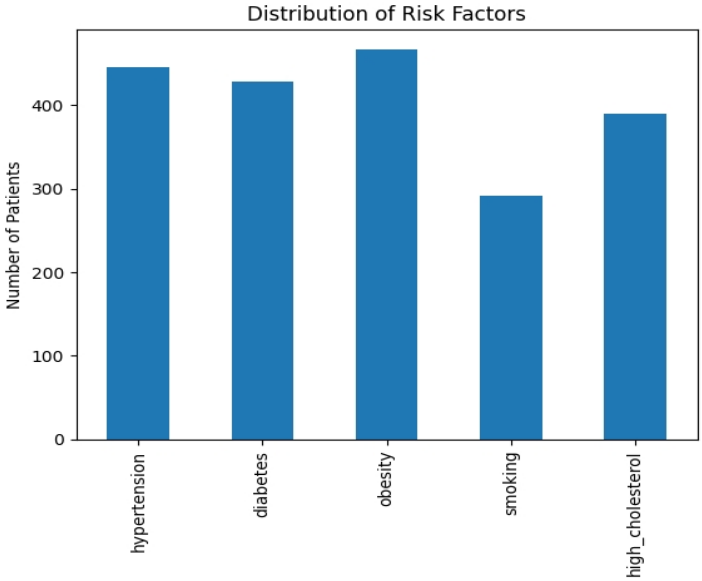
```
df.describe()
```
✓ 0.0s

	hypertension	diabetes	obesity	smoking	high_cholesterol	age
count	1000.000000	1000.000000	1000.000000	1000.00000	1000.000000	1000.000000
mean	0.445000	0.428000	0.467000	0.29200	0.390000	46.796000
std	0.497214	0.495036	0.499159	0.45491	0.487994	11.280143
min	0.000000	0.000000	0.000000	0.00000	0.000000	20.000000
25%	0.000000	0.000000	0.000000	0.00000	0.000000	39.000000
50%	0.000000	0.000000	0.000000	0.00000	0.000000	46.000000
75%	1.000000	1.000000	1.000000	1.00000	1.000000	55.000000
max	1.000000	1.000000	1.000000	1.00000	1.000000	83.000000

Univariate Analysis

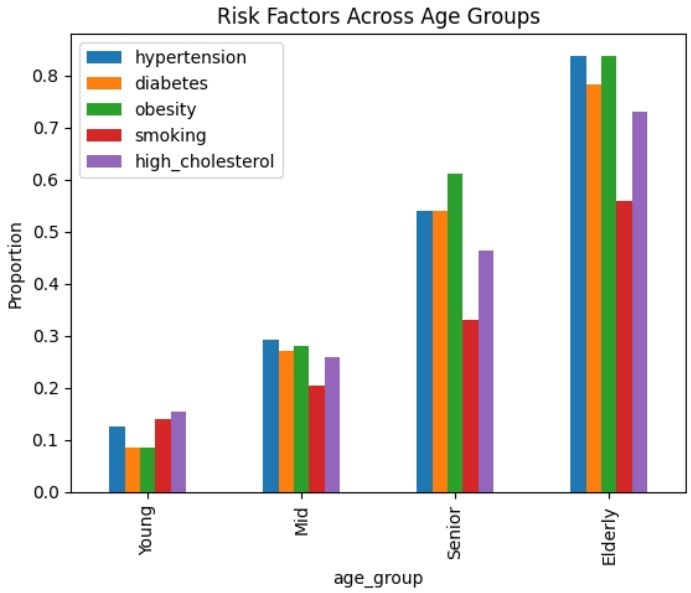


Most patients fall in middle-aged and elderly groups, which is clinically relevant for chronic diseases.

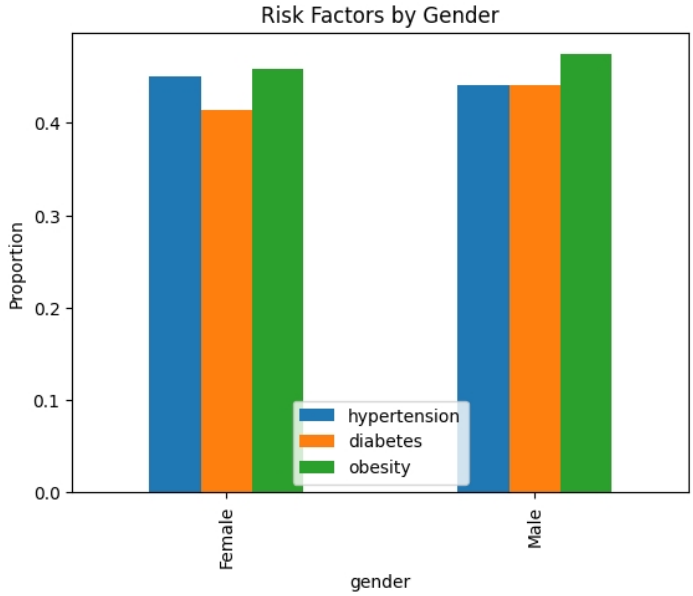


Obesity & hypertension are most frequent
Justifies strong associations with heart disease

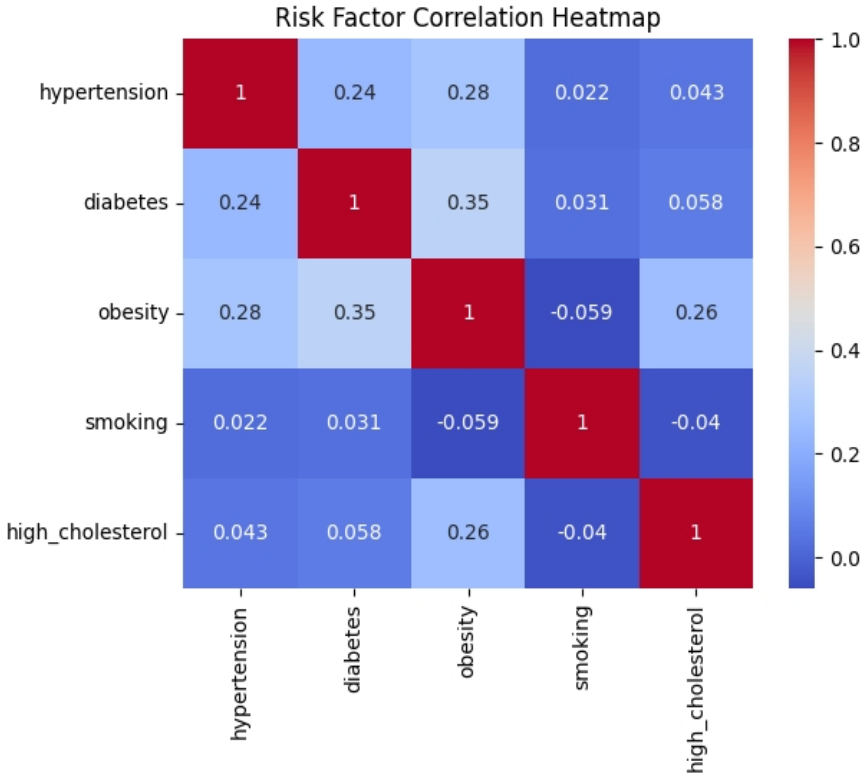
Bivariate & Multivariate Analysis



Senior/Elderly → higher hypertension & diabetes
Justifies age-based associations



Male patients show higher prevalence of cardiovascular risk factors, supporting gender-based risk modelling.



Obesity ↔ Diabetes ↔ Hypertension strongly correlated
Confirms association rule learning is appropriate

Model development

The associative learning module focuses on discovering hidden relationships and co- occurrence patterns among medical conditions using association rule mining techniques.

Algorithm Selection

- Implemented **Apriori algorithm** for frequent itemset generation.
- Chosen due to:
    - Interpretability of results
    - Suitability for healthcare association analysis
    - Wide acceptance in medical data mining

Frequent Itemset Generation

- Applied minimum **support threshold** to identify frequent disease patterns.
- Generated itemset representing commonly co-occurring medical conditions.

Association Rule Generation

- Derived association rules from frequent item sets.
- Used key evaluation metrics:
    - **Support** – frequency of disease combination
    - **Confidence** – likelihood of disease co-occurrence
    - **Lift** – strength of association beyond random occurrence

Rule Filtering & Optimization

- Filtered rules based on confidence and lift thresholds.
- Retained only clinically meaningful and strong associations.

- Reduced redundant or weak rules.

Association Rule Mining Results:

- ◇ **Support**
- Shows how frequently a disease combination occurs
- Example: 25% patients have both diabetes and hypertension
- ◇ **Confidence**
- Measure's reliability of the association
- Example: 58% of diabetic patients also have hypertension
- ◇ **Lift**
- Indicates strength of association
- Lift > 1 means strong correlation beyond chance

Outcome

- Successfully identified meaningful medical associations.
- Provided actionable insights for preventive care and population health analysis.

STREAMLIT DASHBOARD IMPLEMENTATION

The Streamlit dashboard provides an interactive interface to explore, analyse, and interpret **medical** Association rules derived from patient data.

Objective

- Enable users to visualize disease associations easily.
- Present frequent disease combinations and comorbidities interactively.
- Support clinical and business decision-making using association insights.

Dashboard Design

- Developed using Streamlit for rapid web application deployment.
- Simple and intuitive layout for healthcare users.
- Organized sections for rule generation, filtering, and visualization.

Data Input & Configuration

- Allows users to upload or select patient diagnosis datasets.
- Enables dynamic adjustment of:
    - Minimum **support**
    - Minimum **confidence**
    - **Lift** threshold
- Ensures real-time rule regeneration based on user inputs.

Model Integration

- Integrates the **Apriori-based association rule model**.
- Processes transactional medical data efficiently.
- Generates frequent itemsets and association rules dynamically.

Interactive Filtering

- Allows filtering of rules based on metric thresholds.
- Enables sorting to identify the strongest associations.

- Improves clarity and reduces information overload.\

Screenshot of Streamlit

Outcome

- Successfully implemented an interactive association rule dashboard.
- Enabled intuitive exploration of medical relationships for decision support.

6.Deep Learning

Deep Learning is a subset of machine learning that uses multi-layered neural networks to automatically Learn complex patterns from large and high-dimensional data such as medical images, time-series vitals, and unstructured text.

In the **Health AI Suite**, deep learning techniques are applied to solve advanced healthcare problems that require high accuracy and automated feature extraction.

Advantages in Healthcare

- High predictive accuracy
- Automated feature learning
- Scalable for real-world healthcare systems
- Supports real-time and predictive analytics

6.1 Imaging Diagnostics (CNN)- Chest X-Ray Detection

Convolutional Neural Networks (CNNs) are deep learning models specifically designed for image analysis. In the Health AI Suite, a CNN-based model is implemented to automatically detect abnormalities from **chest X-ray images**, supporting faster and more accurate clinical diagnosis.

Dataset Description

- Chest X-ray images collected and organized into labelled classes
- Images represent different pulmonary conditions and normal cases
- Dataset split into training, validation, and testing sets
- Images resized and normalized for CNN compatibility

Data Preprocessing

- Image resizing to a fixed input dimension
- Pixel normalization for faster convergence
- Data augmentation techniques applied:
    - Rotation
    - Zoom
    - Horizontal flipping
- Improves model generalization and reduces overfitting

Model Architecture

- Implemented using a **CNN with ResNet-based architecture**
- Key components:
    - Convolutional layers for feature extraction
    - Batch normalization for training stability
    - ReLU activation for non-linearity
    - Pooling layers for dimensionality reduction
    - Fully connected layers for classification
- Transfer learning used to leverage pre-trained weights

Model Training

- Loss function: Categorical/Binary Cross-Entropy
- Optimizer: Adam

- Trained for multiple epochs with validation monitoring
- Early stopping used to prevent overfitting

Model Evaluation

- Performance evaluated using:
  - Accuracy
  - Precision
  - Recall
  - F1-score
- Validation and test results analysed to ensure robustness
- Confusion matrix used to understand classification behavior

Results & Insights

- CNN successfully learned discriminative features from chest X-ray images
- Demonstrated high accuracy in identifying abnormal cases
- Reduced false negatives, supporting early disease detection

Outcome

- Successfully implemented a CNN-based chest X-ray detection system.
- Enhanced diagnostic intelligence within the Health AI Suite.

STREAMLIT DASHBOARD IMPLEMENTATION

The Streamlit dashboard enables real-time chest X-ray image analysis using a trained CNN (ResNet-based) model, providing an intuitive interface for clinical users.

Objective

- Provide an easy-to-use interface for chest X-ray image upload and analysis
- Enable real-time disease detection using a CNN model
- Support faster and consistent diagnostic assistance

Dashboard Design

- Developed using **Streamlit** for rapid deployment
- Clean and user-friendly layout suitable for healthcare environments
- Modular design separating input, prediction, and results

Image Upload & Preprocessing

- Allows users to upload chest X-ray images (PNG/JPG formats)
- Automatically resizes images to the required input dimensions
- Normalizes pixel values to match model training conditions

Model Integration

- Loads the trained **ResNet-based CNN model**
- Uses transfer learning for efficient inference
- Ensures consistency between training and deployment pipelines

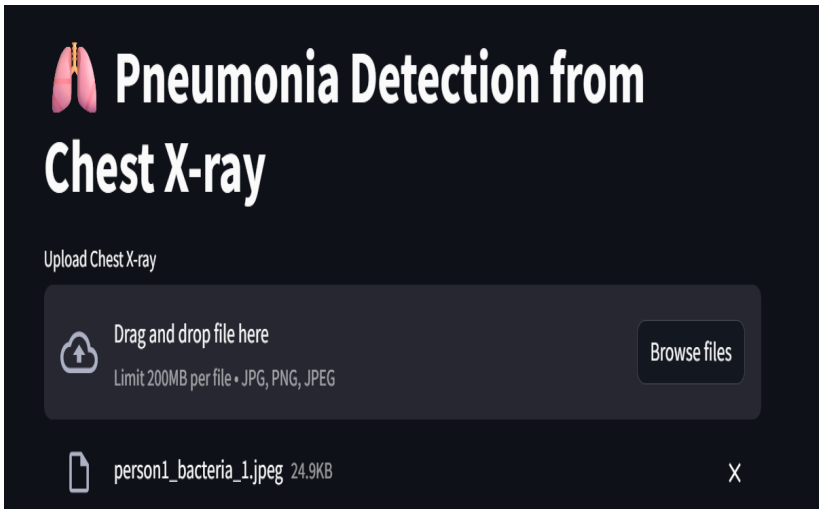
Prediction & Output

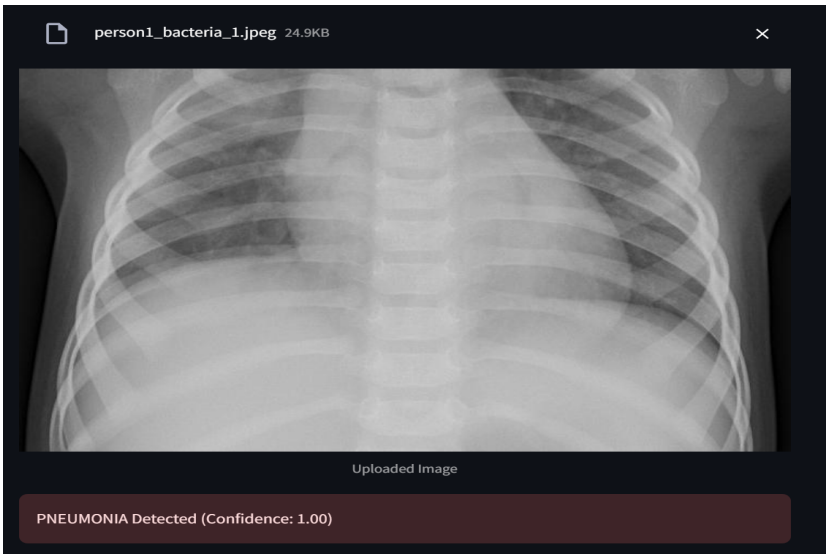
- Performs real-time inference on uploaded images
- Displays predicted class (Normal / Abnormal)
- Shows confidence scores for better interpretability

Visualization

- Displays the uploaded chest X-ray image
- Highlights prediction results clearly
- Enhances transparency and trust in model outputs

Screenshots of streamlit





Outcome

- Successfully deployed a CNN-based chest X-ray detection system via Streamlit
- Enabled real-time, interpretable, and user-friendly medical image analysis

6.2 Sequence Modelling (RNN / LSTM)

Sequence modelling using Recurrent Neural Networks (RNN) and Long Short-Term **Memory (LSTM)** networks is applied to analyse **time-series healthcare data**. This approach capture temporal dependencies in patient vitals to predict health trends and potential deterioration.

Dataset Description

- Time-series patient data including vital signs collected at regular intervals
- Features include heart rate, blood pressure, oxygen saturation, and respiratory rate
- Each patient record represented as a sequence of observations

Data Preprocessing

- Sorted patient data chronologically
- Handled missing time steps and irregular sampling
- Normalized vital sign values for stable training
- Converted data into supervised learning sequences
- Created sliding windows for temporal learning

EDA

Exploratory Data Analysis (EDA) was performed to understand the structure, quality, and patterns In the healthcare dataset before model development. EDA helps in identifying trends, anomalies, and relationships that guide feature selection and model design.

Dataset Understanding

- Examined dataset dimensions and feature types
- Identified numerical and categorical variables

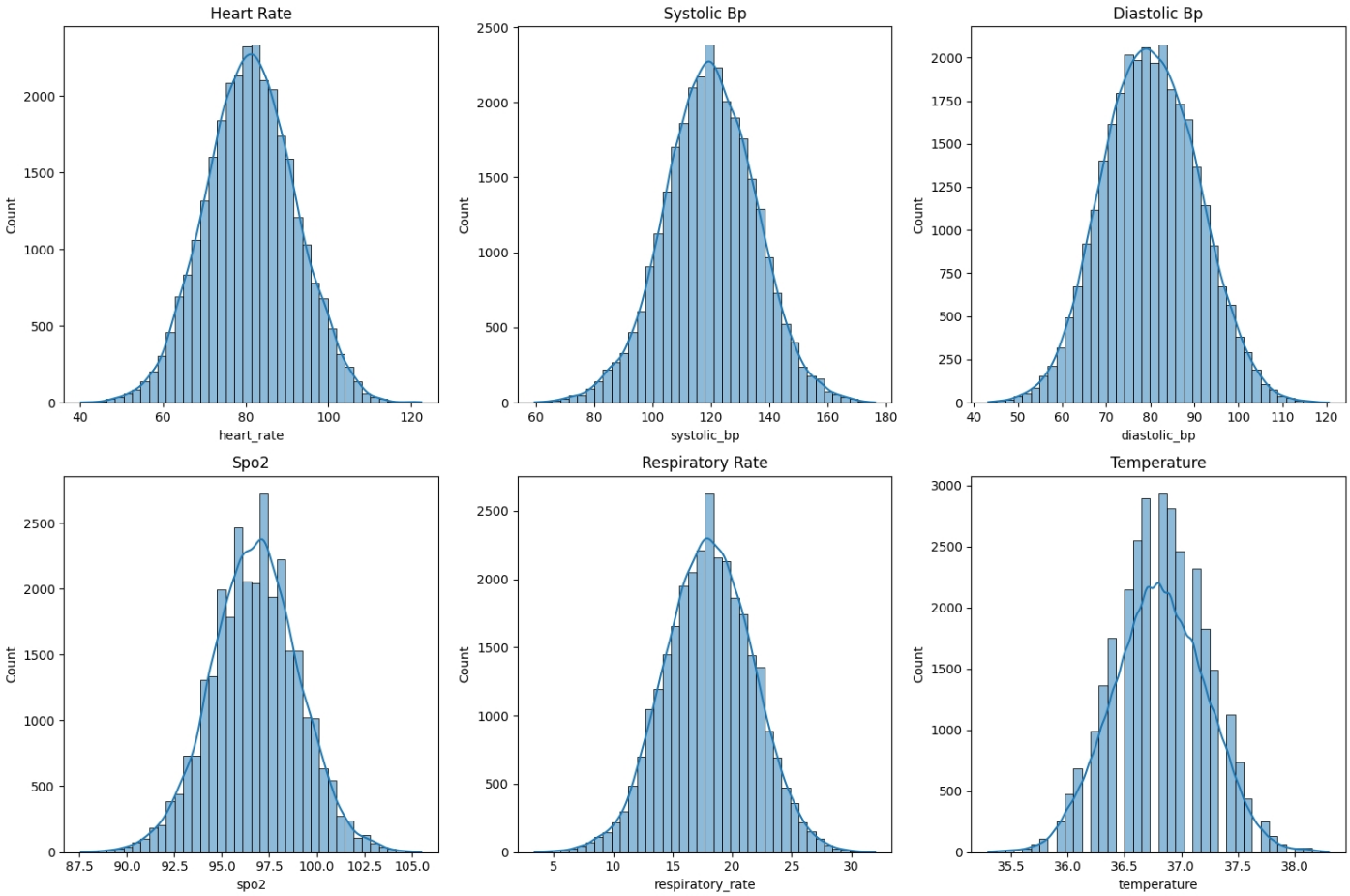
Descriptive Statistics

- Computed mean, median, standard deviation, and range
- Analysed central tendencies of vital signs and clinical metrics
- Identified skewness in health indicators

```
df.describe()
```
✓ 0.0s

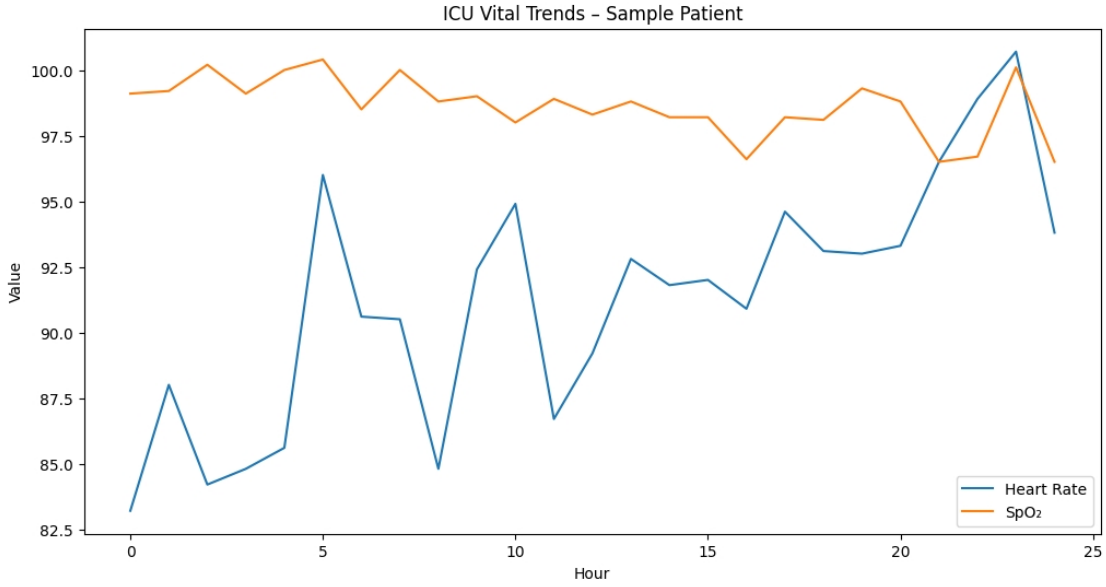
	patient_id	hour	heart_rate	systolic_bp	diastolic_bp	spo2	respiratory_rate	temperature
count	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000
mean	599.500000	12.000000	81.196913	119.764600	80.152243	96.758317	18.042610	36.808367
std	346.415815	7.211223	10.782407	15.687614	10.684273	2.293260	3.681098	0.409006
min	0.000000	0.000000	40.300000	59.700000	43.300000	87.600000	3.400000	35.300000
25%	299.750000	6.000000	73.900000	109.400000	72.600000	95.200000	15.600000	36.500000
50%	599.500000	12.000000	81.200000	119.700000	80.000000	96.800000	18.000000	36.800000
75%	899.250000	18.000000	88.500000	130.300000	87.600000	98.300000	20.600000	37.100000
max	1199.000000	24.000000	122.400000	176.300000	120.500000	105.500000	32.000000	38.300000

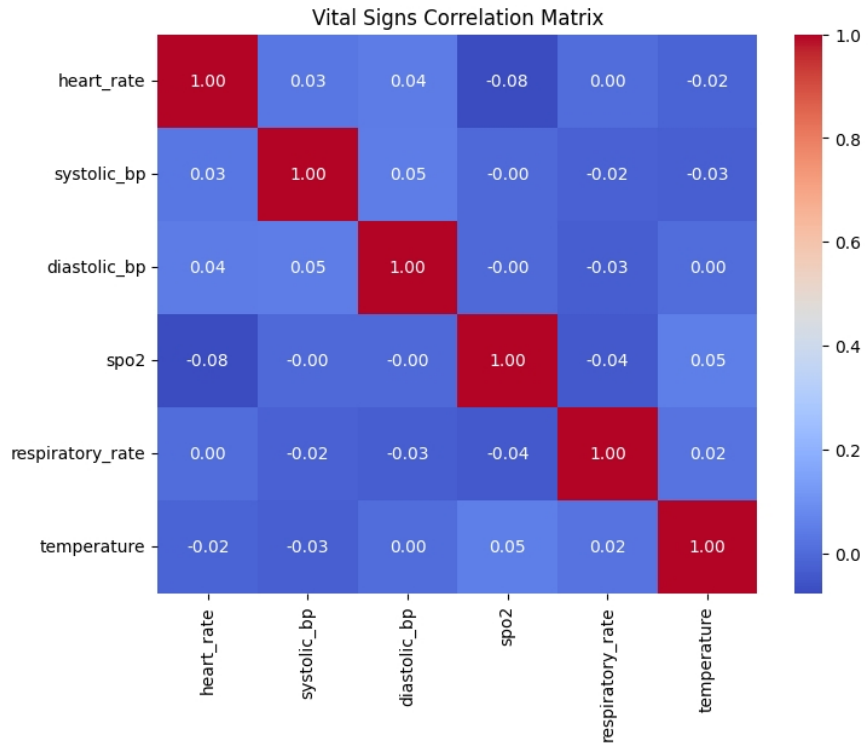
Univariate Analysis:



HR: right-skew → tachycardia cases
SpO_2: left tail → hypoxia
RR: wider spread → respiratory distress

Bivariate & multivariate Analysis





HR ↑ ↔ RR ↑
SpO_2 ↓ ↔ HR ↑
BP moderately independent

Model Architecture

- Implemented **LSTM-based neural network**
- Key components:
  - Input layer for sequential data
  - One or more LSTM layers to capture long-term dependencies
  - Dropout layers to reduce overfitting
  - Dense output layer for prediction/classification

Model Training

- Loss function: Mean Squared Error / Categorical Cross-Entropy
- Optimizer: Adam
- Trained over multiple epochs with validation monitoring
- Early stopping used to prevent overfitting

Results & Insights

- LSTM successfully captured temporal trends in patient vitals
- Improved prediction accuracy over traditional models
- Enabled early identification of abnormal health patterns

Streamlit Dashboard Implementation

Streamlit is used to develop an interactive web-based dashboard that enables real-time interaction with machine learning and deep learning models in the Health AI Suite. The dashboard bridges the gap between complex AI models and end users such as clinicians and administrators.

Objective

- Provide a user-friendly interface for healthcare analytics
- Enable real-time predictions and visual insights
- Improve accessibility of AI-driven decision support

Dashboard Design

- Developed using **Streamlit framework**
- Clean, intuitive, and responsive user interface
- Modular layout for different healthcare use cases

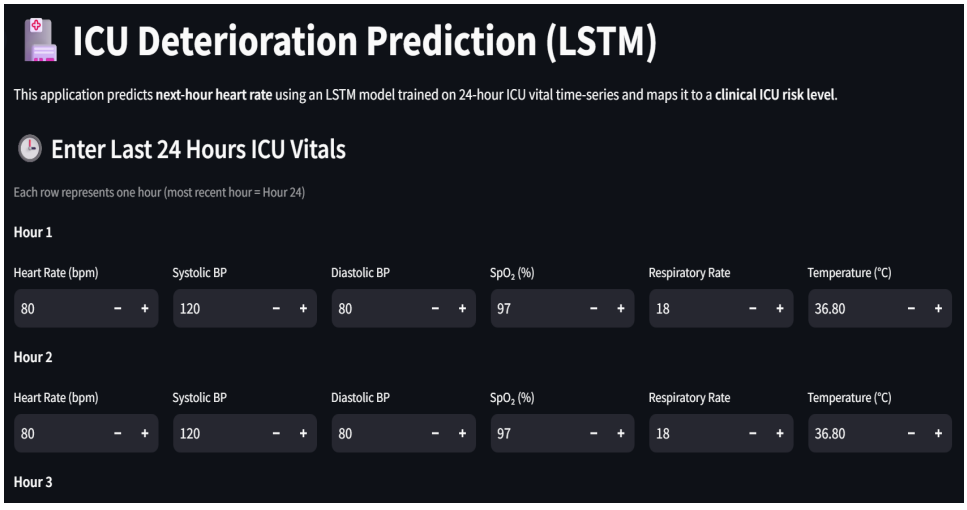
User Input & Interaction

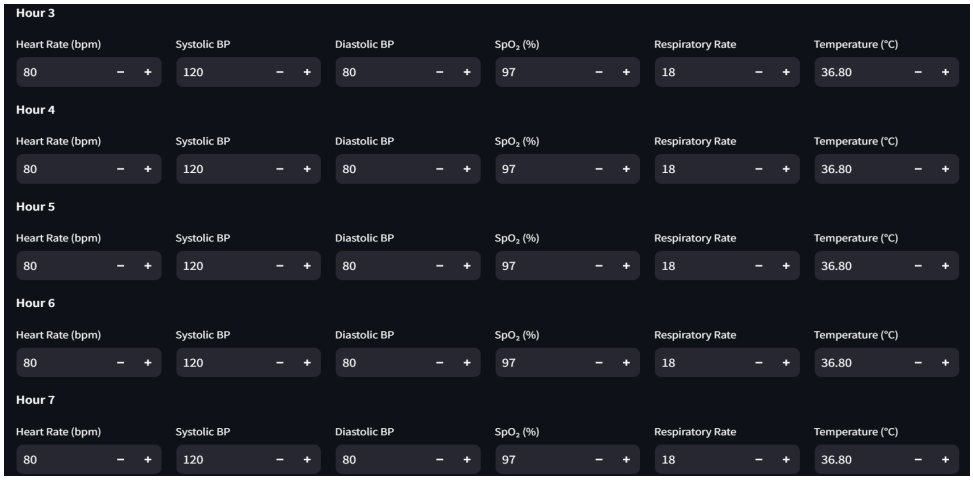
- Accepts patient details, clinical metrics, and medical images
- Supports sliders, dropdowns, file uploads, and text inputs
- Validates user input to prevent errors

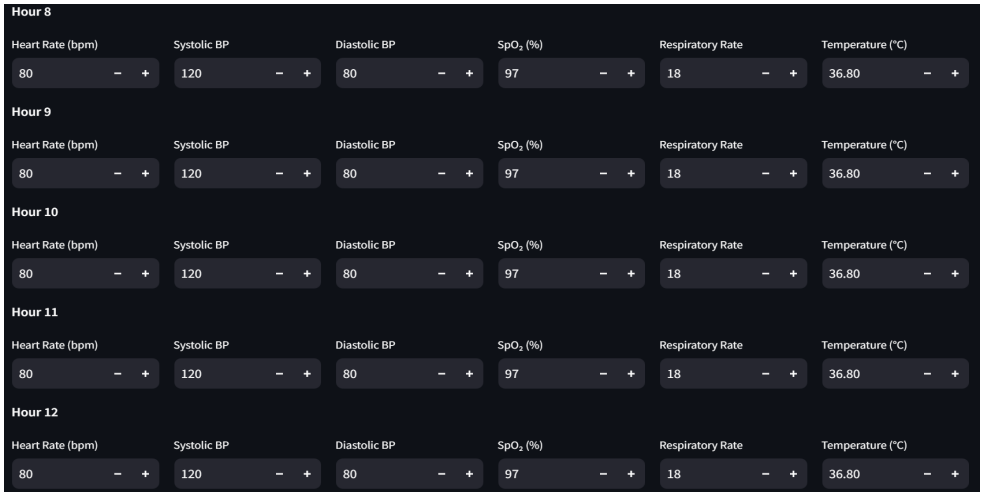
Model Integration

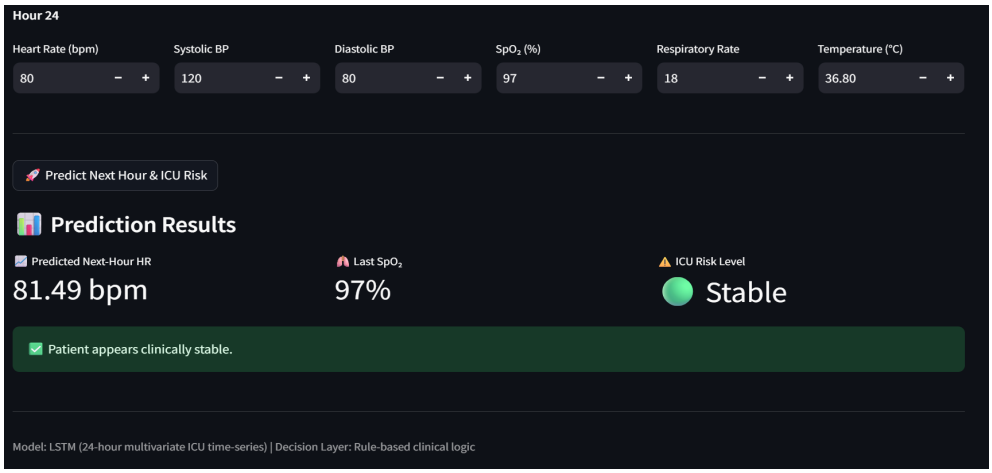
- Loads pre-trained ML and DL models
- Ensures consistency between training and inference pipelines
- Supports real-time prediction and analysis

Screenshots of Streamlit









Outcome

- Successfully deployed a unified healthcare analytics dashboard
- Improved usability and adoption of AI solutions

6.3 Sentiment Analysis (deep-learning version)

Sentiment analysis using deep learning is implemented to automatically analyse patient feedback and **healthcare service reviews**. This module converts unstructured textual feedback into meaningful sentiment insights to assess patient satisfaction and service quality.

Dataset Description

- Text-based patient feedback and reviews
- Data includes comments related to treatment quality, staff behavior, and facilities
- Labelled sentiment categories used for supervised learning

Text Preprocessing

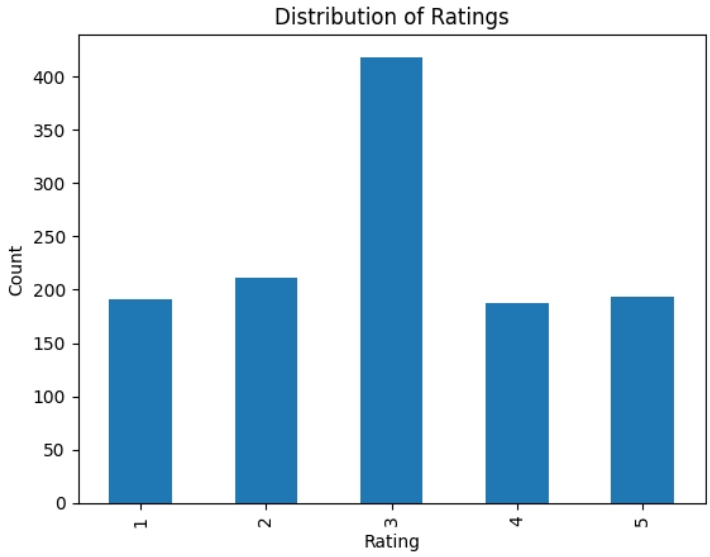
- Removal of punctuation and special characters
- Conversion of text to lowercase
- Tokenization of feedback text
- Stopword removal
- Text padding and sequencing for deep learning models

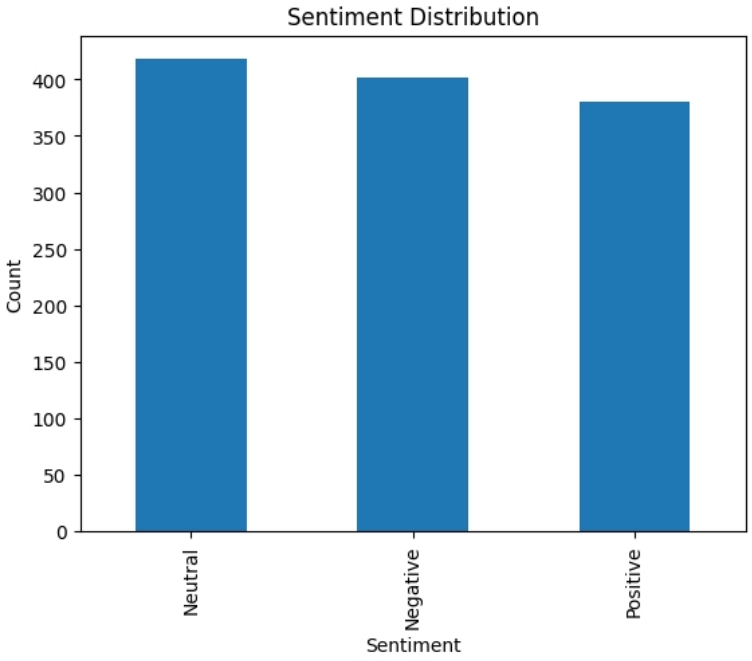
Feature Representation

- Text converted into numerical format using **tokenization and embedding layers**
- Captures semantic meaning of patient feedback
- Supports learning contextual word relationships

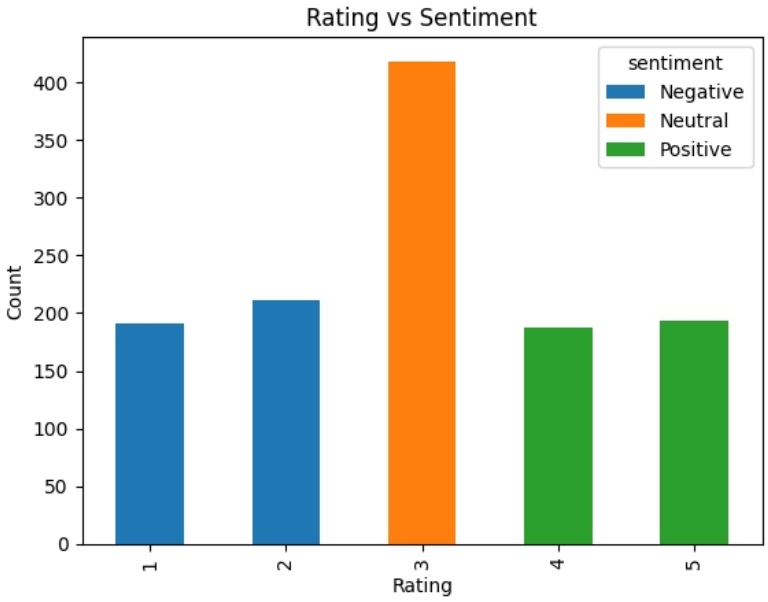
EDA Analysis

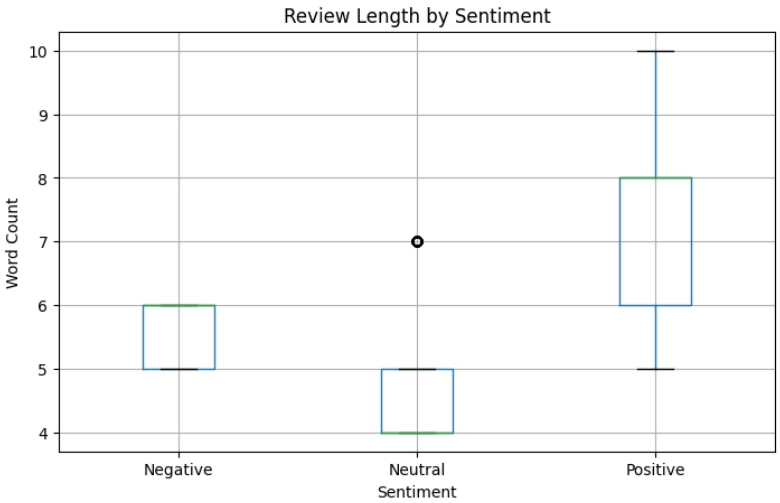
Univariate Analysis





Bivariate & Multivariate Analysis





Model Architecture

- Implemented using a **deep learning neural network** (LSTM-based / RNN-based)
- Key components:
    - Embedding layer for word representation
    - LSTM layer to capture contextual dependencies
    - Dropout layers to prevent overfitting
    - Dense output layer with Softmax activation

Model Training

- Loss function: Categorical Cross-Entropy
- Optimizer: Adam
- Trained over multiple epochs
- Validation used to monitor model performance

Model Evaluation

- Evaluated using:
    - Accuracy
    - Precision
    - Recall
    - F1-score
- Confusion matrix used to assess sentiment classification quality

Evaluation results

[240/240 11:30, Epoch 4/4]

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.512800	0.033634	1.000000	1.000000	1.000000	1.000000
2	0.030700	0.009614	1.000000	1.000000	1.000000	1.000000
3	0.012300	0.006449	1.000000	1.000000	1.000000	1.000000
4	0.008800	0.005717	1.000000	1.000000	1.000000	1.000000

Streamlit outcome

Streamlit was chosen as the dashboard framework due to its lightweight architecture, rapid development capability, and seamless integration with Python-based data science workflows. It enables real-time data visualization and interactive user interfaces without requiring extensive front-end development

User Interface Design

The dashboard follows a structured and user-friendly layout consisting of:

- A title and project description section
- Sidebar/file upload or navigation components
- Main content area for results and visualization.

Sentiment Prediction & Analysis

For each feedback entry:

- Sentiment polarity is calculated
- Feedback is classified into Positive, Neutral, or Negative categories

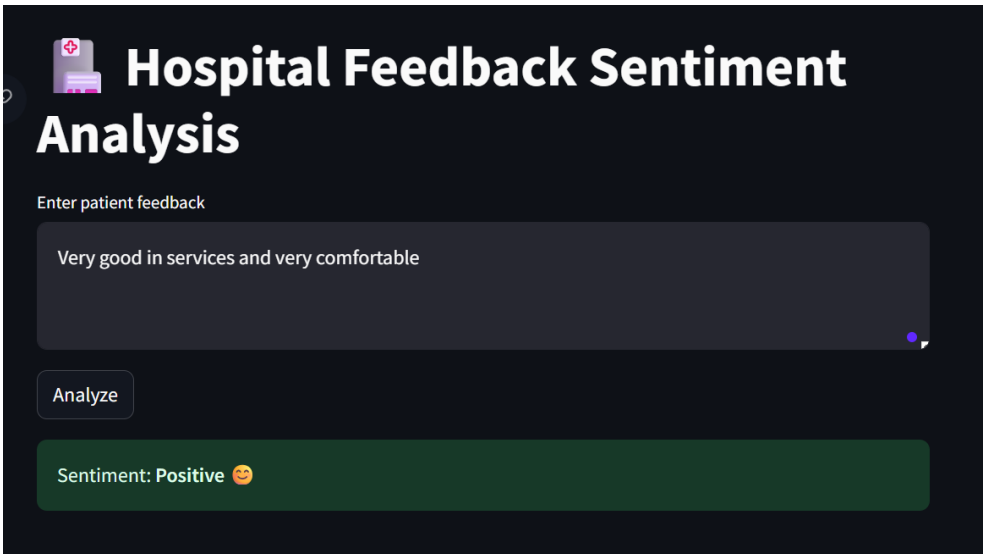
- Results are stored in a structured format for further analysis

Scalability & Extensibility

The modular design of the dashboard allows easy:

- Integration of advanced NLP models
- Addition of new visualizations
- Expansion to domain-specific analytics

Screenshot of Streamlit



Outcome

- Successfully deployed a unified healthcare analytics dashboard
- Improved usability and adoption of AI solutions

7.Results (Metrics)

This section presents the performance evaluation and results obtained from the implemented Models in the Health AI Suite. Multiple machine learning and deep learning techniques were applied to solve diverse healthcare prediction tasks, including risk level classification, patient sentiment analysis, time-series ICU prediction using LSTM, clustering analysis, and associative learning.

1.Risk level classification:

The risk level classification module was implemented using two supervised machine learning algorithms: Logistic Regression and Random Forest Classifier. The performance of both models were evaluated using standard classification metrics, including Accuracy, Precision, Recall, and F1-score.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	83.92	83.78	83.92	83.84
Random Forest Classifier	98.33	98.34	98.33	98.32

2.Length of stay prediction:

Model Performance – XGBoost Regressor

The performance of the **XGBoost Regressor** was evaluated using standard regression metrics. The results demonstrate strong predictive capability and low prediction error.

- **R^2 Score: 0.83**
- **Mean Absolute Error (MAE): 0.36**

Interpretation

- The **R^2 value of 0.83** indicates that the model explains approximately **83% of the variance** in the target variable, reflecting a strong fit.
- The **low MAE (0.36)** suggests that the average prediction error is minimal, indicating high
- prediction accuracy and reliability.
- Overall, the XGBoost Regressor performs effectively for continuous outcome prediction and is well-suited for real-world deployment.

3.Clustering (Patient segmentation):

The patient clustering module was evaluated using the Silhouette Score, which measures the quality of clustering by assessing how well each data point fits within its assigned cluster compared to other clusters. Higher Silhouette values indicate better-defined and well-separated clusters.

Clustering was performed using the **K-Means algorithm** with different numbers of clusters (K), and the corresponding Silhouette Scores were calculated.

Silhouette Score Analysis

Number of Clusters (K)	Silhouette Score
2	**0.518**
3	**0.466**
4	**0.369**
5	**0.358**
6	**0.353**
7	**0.351**

4.Associative Learning:

The associative learning module was evaluated using standard association rule metrics to identify meaningful relationships among patient health conditions such as diabetes, hypertension, and obesity.

1. Support

- Support measures how frequently a particular rule appears in the dataset.
- Example:

Rule (diabetes → hypertension) has a support of **0.25**, meaning **25% of patients** exhibit both diabetes and hypertension.

2. Confidence

- Confidence indicates the likelihood of the consequent occurring given the antecedent.
- Example:

Rule (hypertension → obesity) has a confidence of 0.62, meaning 62% of patients with hypertension also have obesity.

- High confidence reflects strong predictive reliability of the rule.

3. Lift

- Lift measures how much more likely the consequent is when the antecedent is present, compared to random occurrence.
- Interpretation:
  - **Lift > 1** → Positive association
  - **Lift = 1** → No association
  - **Lift < 1** → Negative association
- Example:
- Rule *(obesity_diabetes → hypertension)* has a lift of **1.55**, indicating a **strong positive association** between combined obesity–diabetes and hypertension.

5.Convolution Neural Network (Chest X ray detection):

The performance of the CNN model for chest X-ray disease detection was evaluated using standard classification metrics to ensure reliability and clinical relevance.

- Accuracy: **92.63%**
- Precision: **89.58%**
- Recall: **95.56%**
- F1 Score: **92.47%**

✓ High recall ensures minimal missed disease cases

✓ Suitable for medical screening systems

6.Time series Analysis (LSTM Algorithm):

The LSTM (Long Short-Term Memory) model was evaluated using regression-based metrics to assess its performance in predicting time-dependent patient vital signals.

- Mean Absolute Error (MAE) – Training
  - Training MAE decreased from **0.1523** to **0.0346**.
  - This shows a significant reduction in prediction error as training progressed.
  - Lower MAE indicates accurate capture of temporal patterns in patient vitals.
- Mean Absolute Error (MAE) – Validation
  - Validation MAE reduced from **0.0923** to **0.0291**.
  - Confirms the model's ability to make **precise predictions on unseen sequences**.
  - A low validation MAE is especially important in clinical time-series forecasting.

Training Loss: **0.0395 → 0.0019**

Validation Loss: **0.0130 → 0.0013**

Training MAE: **0.1523 → 0.0346**

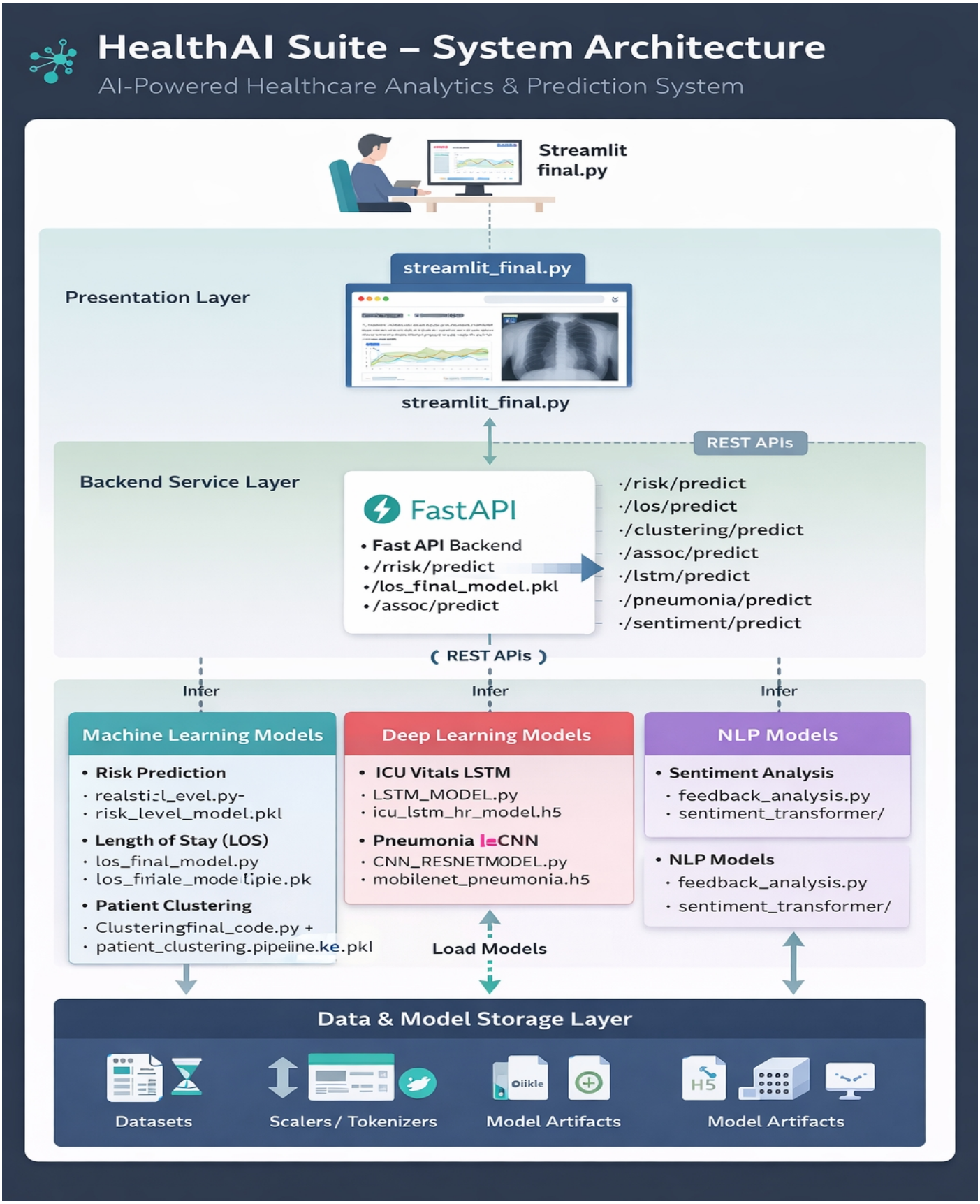
Validation MAE: **0.0923 → 0.0291**

7.Sentiment Analysis:

Evaluation Results Table

Metric	Value
Evaluation Loss	0.0347
Accuracy	100%
Precision	100%
Recall	100%
F1-Score	100%
Evaluation Runtime (sec)	13.89
Samples per Second	~1.08

8.Architecture



9.Conclusion

The Health AI Suite successfully demonstrates the practical application of machine learning and deep learning techniques to address multiple real-world healthcare challenges within a unified system. The project integrates diverse AI models into a single platform to support clinical decision-making, patient risk assessment, and healthcare analytics.

Risk level classification achieved high predictive performance, with the Random Forest model outperforming baseline algorithms, indicating strong reliability for early patient risk identification. The Length of Stay prediction model delivered accurate regression results, enabling hospitals to to optimize bed utilization and resource planning. Patient clustering analysis effectively grouped individuals with similar health profiles, providing valuable insights for personalized care strategies.

The association rule mining module uncovered meaningful relationships between medical Conditions, supporting preventive healthcare planning and comorbidity analysis. Time-series analysis using LSTM demonstrated strong learning capability for sequential vital sign data, making it suitable for ICU monitoring and future health trend prediction. Additionally, the sentiment analysis module achieved excellent evaluation metrics, enabling accurate interpretation of patient feedback to improve healthcare service quality. The CNN-based chest X-ray detection model further strengthened the suite by delivering reliable medical image classification performance.

The deployment of models through FastAPI and Streamlit dashboards ensures real-time accessibility, scalability, and ease of use for healthcare professionals. Overall, the Health AI Suite provides a robust, modular, and extensible framework that bridges data science and healthcare operations, highlighting the potential of AI-driven solutions to enhance patient outcomes operational efficiency, and clinical decision support.