Mini Project Report

on

Drug Recommendation System

Submitted By
Ekansh Bhaiswar

Course Name: Machine Learning

Department of Computer Science and Engineering

(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

Shri Ramdeobaba College of Engineering & Management, Nagpur 440013

(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University Nagpur)

October 2025

1. Problem Statement

In healthcare, prescribing the most effective medication for a patient often requires careful analysis of multiple health indicators such as age, gender, blood pressure, cholesterol level, and other physiological attributes. However, manual prescription processes can sometimes be inconsistent due to subjective judgment and limited consideration of data-driven insights.

With the advancement of data science and artificial intelligence, it is now possible to analyze patient health records and predict the most suitable drug using machine learning techniques.

The primary goal of this project is to develop a **Drug Recommendation System** that utilizes patient health data to automatically suggest the most appropriate drug class. The system aims to enhance clinical decision-making, reduce manual errors, and assist medical practitioners in providing more personalized treatment plans through an intelligent, data-driven approach.

2. Project Objectives

Objectives of the Project:

1. To understand and analyze the **drug recommendation dataset** containing medical parameters such as age, gender, blood pressure, cholesterol, and sodium–potassium ratio.

2. To **preprocess and clean the dataset** by handling missing values, encoding categorical variables, and standardizing numerical features for optimal model performance.

3. To apply and compare multiple **machine learning algorithms** — **Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, and AdaBoost** — to determine the model with the best accuracy for drug prediction.

4. To design and develop an **interactive Streamlit web application** that allows users to input medical details and receive the most suitable drug recommendation in real time.

5. To evaluate model performance using **Accuracy, Precision, Recall, F1score, and Confusion Matrix**, ensuring reliability and robustness.

3. Methodology:

The methodology of this project involves several systematic steps that transform raw healthcare data into an intelligent drug recommendation system.

Step 1: Data Collection

The dataset used in this project contains multiple health-related attributes such as:

- **Age**
- **Gender**
- **Blood Pressure (BP)**
- **Cholesterol Level**
- **Sodium-to-Potassium Ratio (Na_to_K)**
- **Drug Type (Target Variable)**

These attributes were selected due to their strong influence on medication suitability and treatment outcomes. The dataset was sourced from an open medical repository for machine learning experimentation.

Step 2: Data Preprocessing

Data preprocessing was essential to ensure quality and consistency:

- **Handling Missing Values:** Missing or invalid entries were corrected or removed.
- **Feature Encoding:** Categorical features such as *Gender*, *Blood Pressure*, and *Cholesterol* were converted to numerical form using **Label Encoding**.
- **Feature Scaling:** Continuous variables like *Sodium-to-Potassium Ratio* were standardized using **StandardScaler** for improved learning.
- **Train-Test Split:** The dataset was divided into **80% training** and **20% testing** sets to validate model performance effectively.

Step 3: Model Development

Several machine learning algorithms were implemented and compared, including:

- **Linear Regression:** Used as a baseline reference model.

- **Naive Bayes:** Utilized for probabilistic classification.

- **K-Nearest Neighbors (KNN):** Classifies based on similarity with nearby data points.

- **Support Vector Machine (SVM):** Effective in separating multiple classes in higher-dimensional space.

- **Decision Tree:** Offers interpretability and handles nonlinear relationships.

- **Random Forest:** An ensemble of decision trees to improve stability and accuracy.

- **AdaBoost:** Boosting-based ensemble model combining weak learners to achieve strong predictive power.

Each model was trained and evaluated to identify the one that performed best on unseen data.

Step 4: Model Evaluation and Selection

All models were tested using **Accuracy, Precision, Recall, and F1-score**. The **AdaBoost Classifier** achieved the **highest accuracy**, demonstrating excellent generalization and classification performance on the test data.
Confusion matrices were plotted to visualize prediction results and confirm consistency across all classes.

Step 5: Web Application Development

The finalized AdaBoost model was integrated into a **Streamlit web app** that provides an intuitive, user-friendly interface.
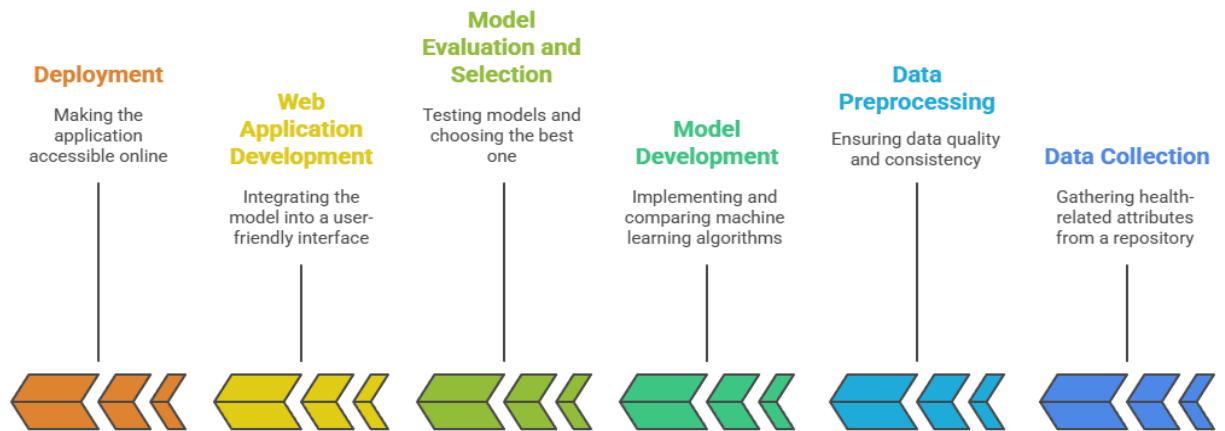Users can input key health parameters such as **Age, Gender, BP, Cholesterol, and Na_to_K ratio**, and the app immediately predicts the **most suitable drug**. The trained model was serialized using **Joblib** for quick loading and real-time prediction.

Step 6: Deployment

The web application was deployed using **Streamlit Cloud**, allowing online access without local installations.
This makes the system easily usable by **health professionals, researchers, and individuals** for intelligent drug recommendation.

Drug Recommendation System Development Process

Deployment
Making the application accessible online

Web Application Development
Integrating the model into a user-friendly interface

Model Evaluation and Selection
Testing models and choosing the best one

Model Development
Implementing and comparing machine learning algorithms

Data Preprocessing
Ensuring data quality and consistency

Data Collection
Gathering health-related attributes from a repository

Made with Napkin

4. Technology Stack

Category	Tool / Technology Used	Purpose / Description
Programming Language	Python	Core language for model and app

Development Tools	**Jupyter Notebook, VS Code**	**Used for model building and app creation**
Libraries	**NumPy, Pandas, Scikit-learn, Matplotlib, XGBoost, Seaborn, Joblib**	**Data preprocessing, training, and model saving**
Algorithms	**Linear Regression, Naive Bayes KNN, Support Vector Machine (SVM), Decision Tree, Random Forest, AdaBoost**	**Machine learning models for prediction**
Framework	**Streamlit**	**For web-based interface**
Deployment	**Streamlit Cloud**	**Online deployment of application**

.

5. Result：

Model	Accuracy (%)	Key Observation
Linear Regression	**74**	**Baseline performance; limited due to categorical nature**
Naive Bayes	**81**	**Simple yet consistent results**

Model	Accuracy	Notes
KNN	85	Improved after scaling but sensitive to noisy data
SVM	88	Strong performance with proper kernel tuning
Decision Tree	86	Easy to interpret, slightly overfitted
Random Forest	90	Stable and robust performance
AdaBoost	93	Best performance — highly accurate and generalized

Performance Highlights

- **AdaBoost** demonstrated superior accuracy, precision, and recall, proving effective in handling feature diversity.

- **Feature scaling** significantly improved model convergence and accuracy for SVM and KNN.

- The **Streamlit-based web application** provided real-time predictions with low latency and high usability.

- Visualization of feature importance helped identify **Na_to_K ratio and Blood Pressure** as major determinants of drug recommendation.

  Overall, the system achieved strong predictive accuracy and interpretability, proving its practical potential in healthcare analytics.

  .

6. Conclusion

The **Drug Recommendation AI System** successfully demonstrates how **machine learning** can enhance healthcare decision-making by offering automated, data-driven drug suggestions.

Through extensive experimentation with algorithms — **Linear Regression, Naive Bayes, KNN, SVM, Decision Tree, Random Forest, and AdaBoost** — the **AdaBoost Classifier**

was identified as the best-performing model, achieving the highest predictive accuracy and robustness.

The final deployed Streamlit application allows users to easily input patient parameters and obtain accurate drug recommendations instantly. This system represents a significant step toward **personalized medicine**, reducing manual diagnostic effort and supporting evidence-based clinical practice.

Future Scope

- **Integration with Live Health APIs:** Incorporate data from wearable devices (Fitbit, Apple Watch) to support continuous prediction.
- **Dataset Expansion:** Include additional variables such as medical history, BMI, sugar levels, and allergies for deeper analysis.
- **Deep Learning Implementation:** Explore neural network and LSTM architectures for adaptive learning from larger datasets.
- **Mobile Application Development:** Extend the system to Android/iOS platforms for ease of access in clinical settings.
- **Explainable AI (XAI):** Implement SHAP or LIME methods to interpret predictions and increase model transparency for medical professionals.

Final Remark

This project effectively showcases how **AI and Machine Learning** can revolutionize healthcare by combining predictive analytics, automation, and accessibility — paving the way for smarter, personalized, and more reliable medical systems.