🩺 Healthcare Data Analysis and Predictive Modeling – Project Report

# 1. Project Overview

This project presents an interactive web-based application built using Streamlit for performing predictive analytics on healthcare datasets. Users can upload a healthcare-related CSV file, select a prediction target, choose between a classification or regression model, and receive model performance metrics and visual insights.

# 2. Objectives

🎯 Provide a user-friendly interface to analyze healthcare data.

🧠 Allow users to choose between classification and regression based on the nature of the target variable.

⚙️ Preprocess data including categorical encoding and handling missing values.

🌲 Train a Random Forest model and evaluate its performance.

📊 Visualize prediction outcomes for interpretability.

# 3. Key Features

📂 File Upload: Accepts .csv healthcare datasets.

🧬 Dynamic Target Selection: Lists all columns for target selection.

🧪 Model Type Selection: Supports classification and regression.

🧹 Automatic Data Preprocessing: Encodes categorical data and handles missing values.

🤖 Model Training: Uses RandomForestClassifier or RandomForestRegressor.

📈 Performance Metrics: Accuracy for classification, MSE for regression.

🎨 Visualization: 2D scatter plots for predictions.

# 4. Technologies Used

🐍 Python Libraries:

- Streamlit – Web interface

- Pandas, NumPy – Data manipulation

- scikit-learn – Machine learning

- Matplotlib, Seaborn – Visualization

# 5. Workflow

1️⃣ Data Input – Upload dataset

2️⃣ Exploration – Preview the data

3️⃣ Model Configuration – Select target and model type

4️⃣ Preprocessing – Encode and clean data

5️⃣ Model Training – Fit the model

6️⃣ Evaluation – View Accuracy or MSE

7️⃣ Visualization – Graphical output

# 6. Example Use Cases

🩻 Predicting disease occurrence (classification)

💵 Estimating medical expenses or hospital stay duration (regression)

👨‍⚕️ Analyzing patient demographics and outcomes

# 7. Limitations and Future Improvements

⚠️ Current Limitations:

- No hyperparameter tuning

- Limited visualization options

- Basic missing value handling

- No model interpretability tools

🔧 Future Enhancements:

- Add SHAP values for explainability

- More model choices

- Advanced missing value treatment

- Downloadable predictions and models

# 8. Conclusion

✅ This interactive ML-powered tool empowers users to analyze healthcare data, train models, and visualize predictions with ease. It's a valuable prototype for healthcare analysts and data scientists.