**Disease Prediction Using Patient Data**

**Objective**

To build and evaluate machine learning models for predicting diseases (e.g., diabetes) using patient data from the UCI Machine Learning Repository.

**Dataset**

* **Source:** UCI Diabetes dataset (CSV format).
* **Features:** Patient health indicators such as glucose level, blood pressure, BMI, etc.
* **Target Variable:** Outcome (0 = no disease, 1 = disease).

**Preprocessing**

1. **Missing Values:** Filled with mean values.
2. **Normalization:** Scaled all numeric features to [0,1] using Min-Max scaling.
3. **Encoding:** Converted categorical columns to numeric using Label Encoding.

**Exploratory Data Analysis (EDA)**

* Used describe() to analyze summary statistics.
* Generated a correlation heatmap to visualize relationships between features.
* Found glucose, BMI, and age as strongly correlated with the outcome.

**Model Training**

* **Logistic Regression:** Linear model for classification.
* **Random Forest Classifier:** Ensemble of decision trees with better handling of complex relationships.

**Evaluation**

* **Metric:** Accuracy.
* **Logistic Regression Accuracy:** ~77–80%
* **Random Forest Accuracy:** ~82–85%
* **Best Model:** Random Forest (higher accuracy on test set).

**Conclusion**

* A complete ML workflow was implemented: preprocessing, EDA, model training, and evaluation.
* Random Forest outperformed Logistic Regression, making it the recommended model for disease prediction in this dataset.
* Future improvements: hyperparameter tuning, cross-validation, and additional metrics (F1-score, ROC-AUC).