Disease Prediction Using Patient Data (Week 1)

Dataset: Pima Indians Diabetes Database (Kaggle)
Goal: Predict diabetes presence (binary classification: diabetic vs. non-diabetic).

Preprocessing

- Missing values: Replaced zeros in Glucose, BloodPressure, SkinThickness, Insulin, BMI with NaN, then imputed using mean (low skew) or median (high skew).
- Scaling: All numeric features normalized between 0 and 1 (Min-Max).
- Encoding: Applied ordinal encoding to categorical columns (none in this dataset).
- Split: 80/20 train-test split, stratified by outcome label.

Exploratory Data Analysis (EDA)

- describe() summary confirmed reasonable distributions, but some missing values were present in medical fields (imputed later).
- Correlation heatmap: Showed that Glucose, BMI, and Age have moderate positive correlation with diabetes outcome, while BloodPressure and SkinThickness showed weaker relationships.

Models

1. Logistic Regression (baseline linear model, max_iter=1000)
2. Random Forest (ensemble model, n_estimators=300, random_state=42)

Results

Model	Accuracy
Logistic Regression	70.78%
Random Forest	74.03%

Best Model: Random Forest

Conclusion

The Random Forest outperformed Logistic Regression by ~3.2%, showing better handling of non-linear relationships in the dataset. This demonstrates that ensemble methods can provide stronger baselines for disease prediction tasks. Next steps could include hyperparameter tuning, cross-validation, and exploring other evaluation metrics (precision, recall, F1-score) to assess performance on imbalanced classes.