WEEK 2 — SUPERVISED LEARNING

(CLASSIFICATION)

Model by mahnoor khan

1. Title Page

Title: Supervised Learning – Diabetes Prediction

Course: FutureXcel – Machine Learning

Week: 2

Prepared by: mahnoor khan

Footer: Model by mahnoor khan

2. Introduction

This project applies supervised machine learning to predict diabetes using the Pima Indians Diabetes dataset. The aim is to clean the data, build baseline and improved models, evaluate performance, and interpret feature importance using modern ML techniques.

3. Dataset Overview

Rows: 768

Target Variable: Outcome (0 or 1)

Features include: Glucose, BMI, BloodPressure, Insulin, Age, etc.

Some medical values were incorrectly recorded as 0, which were treated as missing values.

4. Data Cleaning & Preprocessing

✓ Step 1: Incorrect zeros replaced with NaN

Affected Columns: Glucose, BloodPressure, SkinThickness, Insulin, BMI

✓ Step 2: Missing values imputedMethod:

Median Imputation

Step 3: Train-Test Split

80% Training

20% Testing

Stratified split to maintain class balance

✓ Step 4: Scaling

Applied only to Logistic Regression using StandardScaler.

5. Baseline Model — Logistic Regression

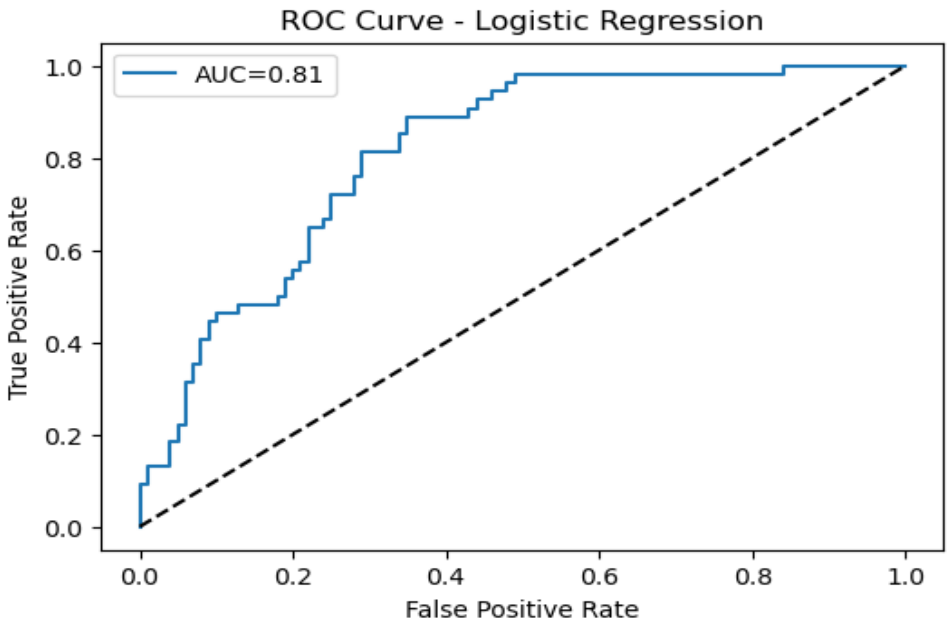
Logistic Regression was used to establish a baseline performance.

✓ Accuracy: (Logistic Regression Accuracy shown in code output)

✓ Classification Report: (Precision, Recall, F1-score shown in output)

ROC curve

```
plt.plot(fpr, tpr, label=f"AUC={auc(fpr, tpr):.2f}")
plt.plot([0,1], [0,1], 'k--')
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve - Logistic Regression")
plt.legend()
plt.show()
```

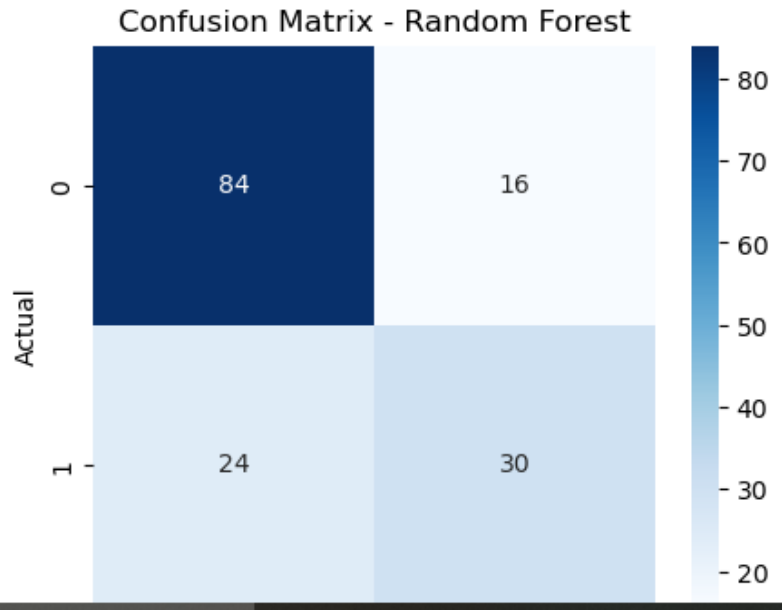


6. Improved Model —

Random Forest Classifier Random Forest was selected as the improved model due to better handling of nonlinear data and the ability to provide feature importance.

7. Confusion Matrix (Random Forest)

[18]:
```
cm = confusion_matrix(y_test, y_pred_rf)
plt.figure(figsize=(5,4))
sns.heatmap(cm, annot=True, fmt='d', cmap="Blues")
plt.title("Confusion Matrix - Random Forest")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()
```



8. Feature Importance — Random Forest

Random Forest ranks features based on contribution.

9. SHAP Interpretation SHAP values help explain the model's predictions feature by feature.

10. Final Conclusion A complete supervised ML pipeline was developed successfully Baseline model: Logistic Regression Improved model: Random Forest (higher accuracy + CV score) Confusion matrix shows reliable prediction patterns SHAP adds clear interpretability to feature impact

Footer: Model by mahnoor khan