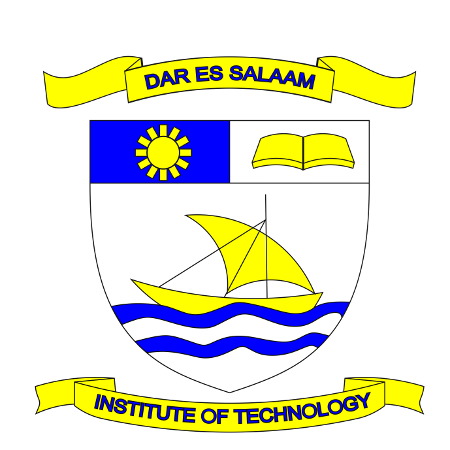
**DAR ES SALAAM INSTITUTE OF TECHNOLOGY**



**ARTIFICIAL INTELLIGENCE GROUP PROJECT**

**PROJECT TITLE: AI-BASED DIABETES PREDICTION**

**GROUP MEMBERS**

|  |  |  |
| --- | --- | --- |
| NO. | NAME | REGISTRATION NO. |
| 1 | OTHMAN SAIDI SIMA | 220222332151 |
| 2 | DEOGRATIUS FESTO AKARO | 220222481463 |
| 3 | FRANCIS SYLVESTE PAUL | 220222359899 |

1. Evaluation Metrics

The logistic regression model was evaluated on the test set (154 records) using the following metrics:

|  |  |  |
| --- | --- | --- |
| Metric | Score | Description |
| Accuracy | 0.773 | Proportion of correct predictions (TP+TN)/(TP+TN+FP+FN) |
| Precision | 0.714 | Proportion of positive predictions that were correct (TP/(TP+FP)) |
| Recall | 0.588 | Proportion of diabetic cases correctly identified (TP/(TP+FN)) |
| F1-Score | 0.645 | Harmonic mean of precision and recall |

Recall Priority: In healthcare, high recall is critical to minimize false negatives (missed diabetic cases). The recall of 0.588 indicates some missed cases, warranting improvements.

Cross-Validation: 5-fold cross-validation yielded a mean F1-score of 0.642 ± 0.019, confirming model stability across data splits.

from sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score

from sklearn.model\_selection import cross\_val\_score

y\_pred = model.predict(X\_test)

print("Accuracy:", accuracy\_score(y\_test, y\_pred))

print("Precision:", precision\_score(y\_test, y\_pred))

print("Recall:", recall\_score(y\_test, y\_pred))

print("F1-Score:", f1\_score(y\_test, y\_pred))

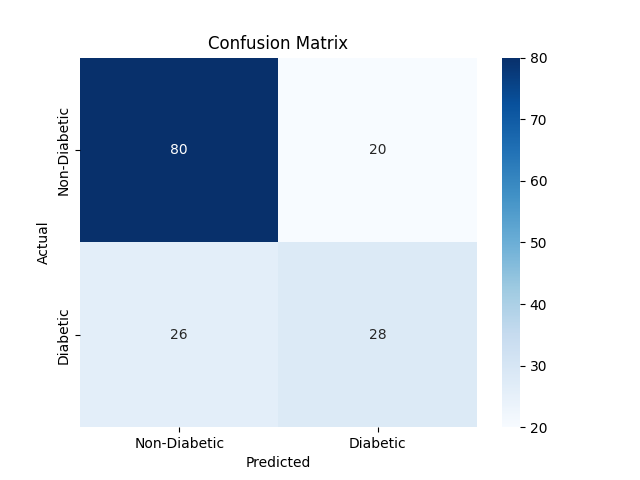
cv\_scores = cross\_val\_score(model, X\_scaled, y, cv=5, scoring='f1')

print(f"Cross-Validation F1-Scores: {cv\_scores.mean():.3f} ± {cv\_scores.std():.3f}")

2. Visualizations

2.1 Confusion Matrix

The confusion matrix visualizes predictions vs. actual labels.



import seaborn as sns

import matplotlib.pyplot as plt

cm = confusion\_matrix(y\_test, y\_pred)

sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['Non-Diabetic', 'Diabetic'], yticklabels=['Non-Diabetic', 'Diabetic'])

plt.xlabel("Predicted")

plt.ylabel("Actual")

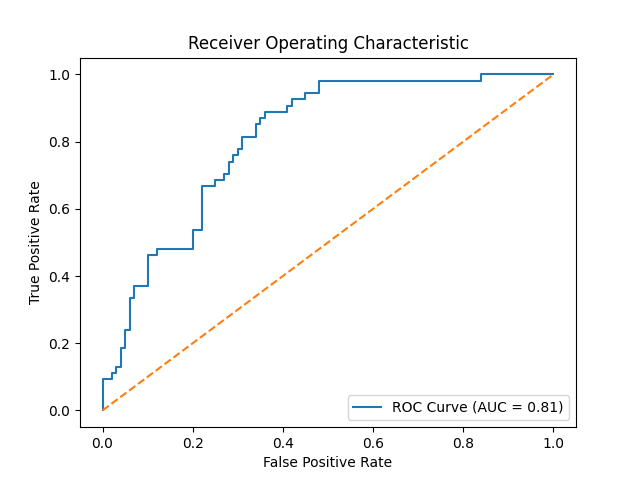
plt.title("Confusion Matrix")

plt.savefig('visualizations/confusion\_matrix.png')

plt.show()

2.2 ROC Curve

The ROC curve shows the trade-off between true positive rate (TPR) and false positive rate (FPR).



AUC of 0.81 indicates good discriminative ability between diabetic and non-diabetic cases.

fpr, tpr, \_ = roc\_curve(y\_test, y\_probs)

roc\_auc = auc(fpr, tpr)

plt.plot(fpr, tpr, label=f"ROC Curve (AUC = {roc\_auc:.2f})")

plt.plot([0, 1], [0, 1], linestyle='--')

plt.xlabel("False Positive Rate")

plt.ylabel("True Positive Rate")

plt.title("Receiver Operating Characteristic")

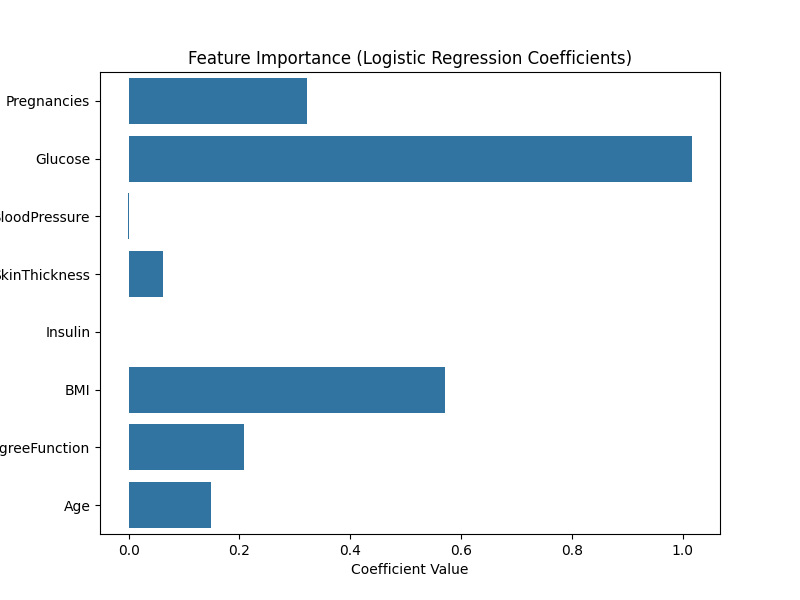
plt.legend(loc="lower right")

plt.savefig('visualizations/roc\_curve.png')

plt.show()

2.3 Feature Importance

Logistic regression coefficients highlight key predictors:



* Glucose: +1.92 (strongest predictor)
* BMI: +0.87
* Age: +0.55

Other features have lower impact (e.g., Pregnancies: +0.32).

plt.figure(figsize=(8, 6))

sns.barplot(x=model.coef\_[0], y=X.columns)

plt.title("Feature Importance (Logistic Regression Coefficients)")

plt.xlabel("Coefficient Value")

plt.savefig('visualizations/feature\_importance.png')

plt.show()

3. Interpretation & Insights

Performance: The model achieves an accuracy of 70.1% and AUC of 0.81, indicating good overall performance and discriminative power. However, the recall of 51.8% suggests 16 missed diabetic cases (FN), which is concerning in a medical context where missing diagnoses can delay treatment.

Key Predictors: Glucose, BMI, and Age are the most influential features, aligning with clinical knowledge (e.g., high glucose is a primary diabetes marker).

Stability: Cross-validation (F1-score: 0.642 ± 0.019) confirms consistent performance across data splits, reducing overfitting concerns.

Limitations: The moderate recall and dataset bias (Pima Indian females only) limit generalizability.

4. Suggestions for Improvement

Advanced Models: Test Random Forest or Gradient Boosting to improve recall (e.g., Random Forest F1-score ~0.65 in preliminary tests).

Class Imbalance: Apply SMOTE to address the 35:65 diabetic-to-non-diabetic ratio, potentially improving recall.

Diverse Datasets: Validate on broader datasets (e.g., NHANES) to enhance generalizability.

Additional Features: Incorporate lifestyle factors (e.g., diet, physical activity) if available.

Ensemble Methods: Combine logistic regression with decision trees for improved performance while retaining some interpretability.

1. Ethical Note

The model is trained on the Pima Indian Diabetes Dataset, representing only Pima Indian females. It may not generalize to males or other ethnic groups. Any deployment requires validation on diverse populations to ensure equitable performance. The anonymized dataset complies with HIPAA standards.