VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A3a : Data Analysis - Logistic vs Tree

GAURI VINOD NAIR

V01110160

Date of Submission: 30-06-2024

TABLE OF CONTENTS

INTRODUCTION

This assignment delves into logistic regression and decision tree analysis using the "diabetes2" dataset, which includes key health indicators like pregnancies, glucose levels, BMI, and more, alongside an outcome variable indicating diabetes occurrence. The primary aim is to employ logistic regression to model the probability of diabetes based on these predictors, ensuring assumptions are met and evaluating model performance through tools like confusion matrices and ROC curves. Concurrently, decision tree analysis will uncover intricate, nonlinear relationships among these variables, providing a comparative perspective on its predictive power relative to logistic regression. By exploring these techniques, this study not only deepens insights into predictive modeling in healthcare but also offers practical applications for enhancing diabetes risk assessment and management strategies.

Through rigorous exploration of logistic regression and decision tree analysis using the "diabetes2" dataset, this assignment focuses on predicting diabetes based on essential health metrics. Logistic regression will model the likelihood of diabetes occurrence using variables such as pregnancies, glucose levels, and BMI, supported by validation of model assumptions and performance assessment using confusion matrices and ROC curves. Additionally, decision tree analysis will unveil complex, nonlinear interactions within the dataset, allowing for a comparative evaluation of its predictive efficacy alongside logistic regression. This dual approach not only enriches understanding of predictive analytics in healthcare contexts but also provides actionable insights for optimizing diabetes prevention and treatment strategies based on robust statistical methodologies.

OBJECTIVES

1. Conduct Logistic Regression Analysis
2. Evaluate Model Performance
   - Assess the predictive performance of the logistic regression model using a confusion matrix.
   - Plot and analyze the ROC curve to measure the model's sensitivity and specificity across various thresholds.
   - Interpret the results to understand how well the model predicts diabetes presence.
3. Perform Decision Tree Analysis
   - Implement a decision tree algorithm on the same dataset to explore nonlinear relationships between predictors and diabetes.
   - Compare the decision tree's predictive performance with that of logistic regression.
   - Evaluate the interpretability of the decision tree model relative to logistic regression.

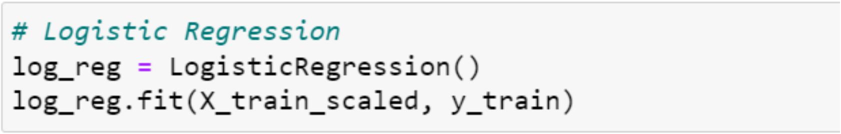
BUSINESS SIGNIFICANCE

Performing logistic regression and decision tree analysis on the "diabetes2" dataset holds significant business implications. Logistic regression provides a structured approach to quantifying diabetes risk, essential for healthcare decision-making. By accurately predicting the likelihood of diabetes based on variables like pregnancies, glucose levels, and BMI, healthcare providers can prioritize interventions and allocate resources effectively. This predictive capability supports early detection and intervention, potentially reducing long-term healthcare costs and improving patient outcomes. Furthermore, logistic regression insights can inform policy formulation by identifying high-risk populations and guiding public health initiatives aimed at diabetes prevention.

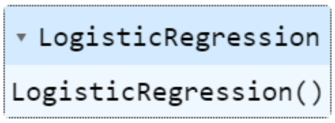
In contrast, decision tree analysis enhances predictive accuracy by uncovering complex, nonlinear relationships within the dataset. This method not only complements logistic regression but also offers a visual and intuitive representation of decision-making processes. The interpretability of decision trees makes them valuable for healthcare professionals and policymakers alike, facilitating informed decisions about patient care and resource allocation. By comparing the results of decision tree analysis with logistic regression, stakeholders gain a comprehensive understanding of diabetes risk factors, enabling tailored healthcare strategies that address individual patient needs effectively. Together, these analytical approaches provide robust tools for advancing diabetes management and enhancing healthcare outcomes.

RESULTS AND INTERPRETATIONS

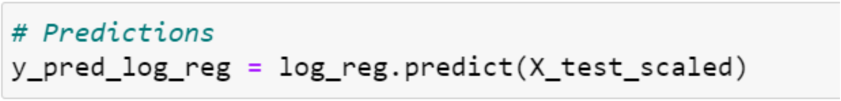
- ⇨ Python



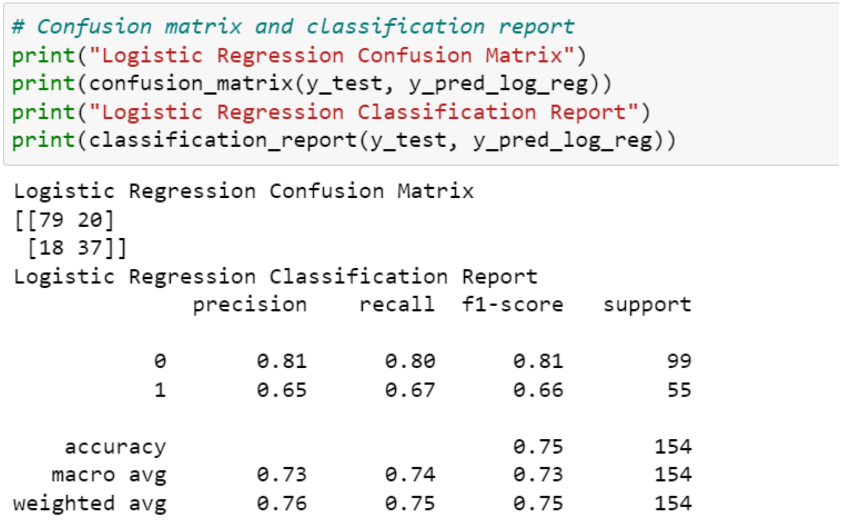
```python
# Logistic Regression
log_reg = LogisticRegression()
log_reg.fit(X_train_scaled, y_train)
```



```
▾ LogisticRegression
LogisticRegression()
```



```python
# Predictions
y_pred_log_reg = log_reg.predict(X_test_scaled)
```



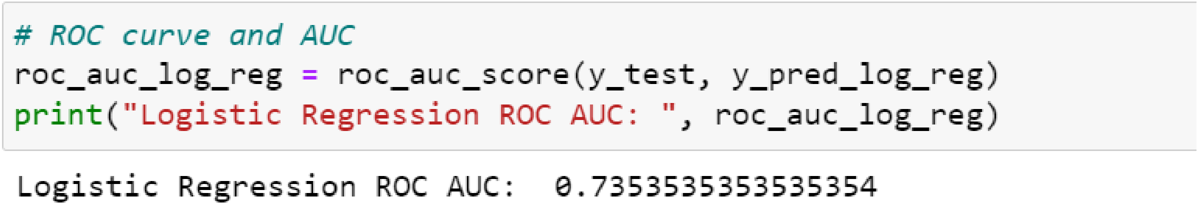
```
# Confusion matrix and classification report
print("Logistic Regression Confusion Matrix")
print(confusion_matrix(y_test, y_pred_log_reg))
print("Logistic Regression Classification Report")
print(classification_report(y_test, y_pred_log_reg))
```

```
Logistic Regression Confusion Matrix
[[79 20]
 [18 37]]
Logistic Regression Classification Report
              precision    recall  f1-score   support

           0       0.81      0.80      0.81        99
           1       0.65      0.67      0.66        55

    accuracy                           0.75       154
   macro avg       0.73      0.74      0.73       154
weighted avg       0.76      0.75      0.75       154
```

Interpretation

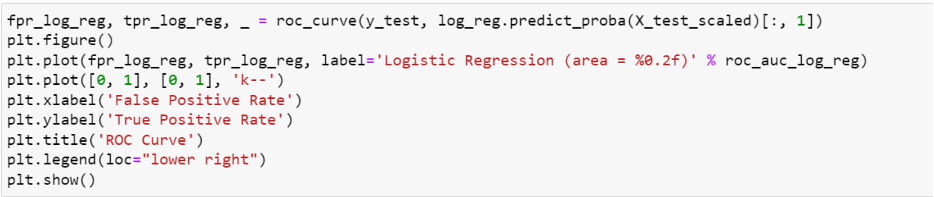
The confusion matrix for the logistic regression model reveals that the model correctly predicted 79 instances of no diabetes (true negatives) and 37 instances of diabetes (true positives). However, it incorrectly predicted diabetes in 20 cases where there was none (false positives) and failed to predict diabetes in 18 cases where it was present (false negatives). These results indicate the model's performance in distinguishing between the two classes, showing a reasonable ability to correctly identify both diabetic and non-diabetic cases, though with some misclassifications.

The classification report provides a detailed performance evaluation of the logistic regression model. For the class representing no diabetes (0), the model achieved a precision of 81% and a recall of 80%, resulting in an F1-score of 81%. For the diabetes class (1), the precision was 65% and the recall was 67%, leading to an F1-score of 66%. The overall accuracy of the model was 75%, indicating that 75% of the predictions were correct. The macro and weighted averages provide a balanced view of performance across both classes. These metrics suggest that while the model performs well overall, it is more accurate in predicting non-diabetic cases compared to diabetic ones, highlighting areas for potential improvement.

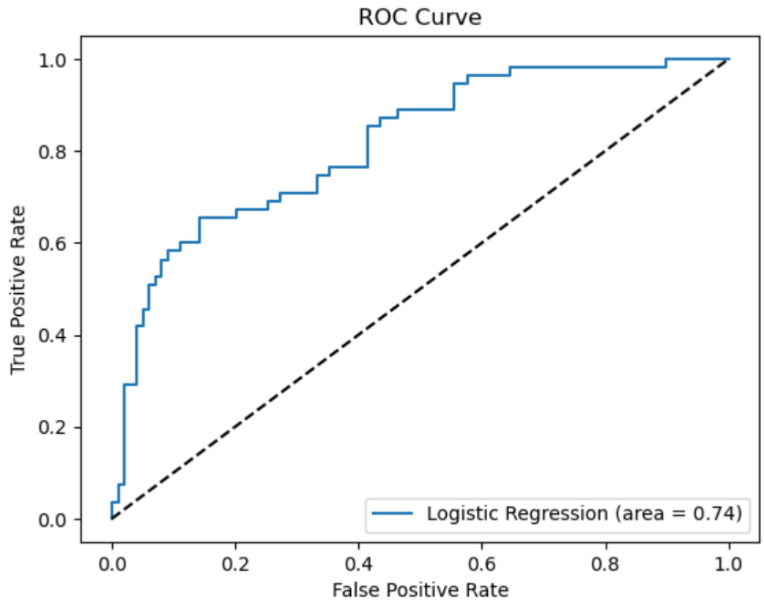


```
# ROC curve and AUC
roc_auc_log_reg = roc_auc_score(y_test, y_pred_log_reg)
print("Logistic Regression ROC AUC: ", roc_auc_log_reg)
```

```
Logistic Regression ROC AUC:  0.7353535353535354
```



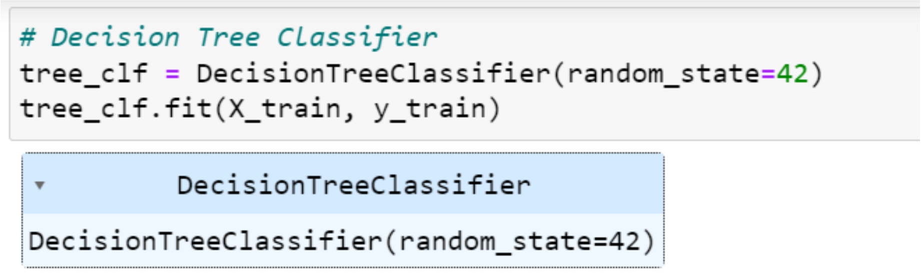
```
fpr_log_reg, tpr_log_reg, _ = roc_curve(y_test, log_reg.predict_proba(X_test_scaled)[:, 1])
plt.figure()
plt.plot(fpr_log_reg, tpr_log_reg, label='Logistic Regression (area = %0.2f)' % roc_auc_log_reg)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend(loc="lower right")
plt.show()
```



Interpretation

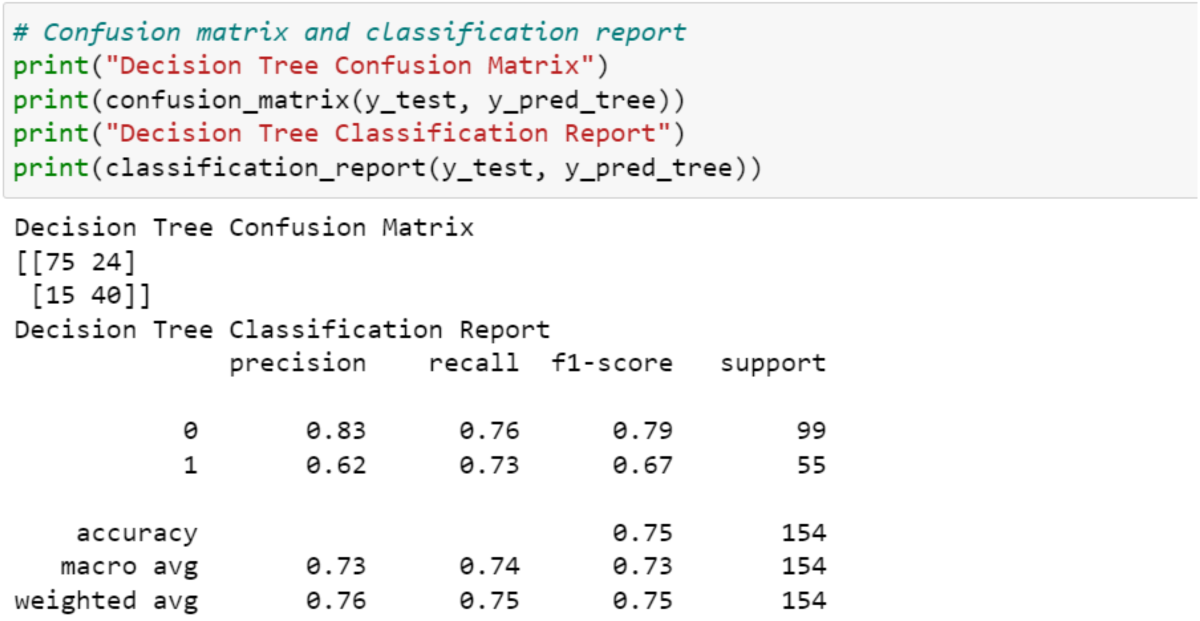
This code calculates the Area Under the Curve (AUC) for the Receiver Operating Characteristic (ROC) curve of the logistic regression model. The roc_auc_score function from the sklearn.metrics module is used to compute this metric, which evaluates the model's ability to distinguish between the positive and negative classes (diabetes and no diabetes, respectively). The y_test variable contains the true labels, while y_pred_log_reg contains the predicted labels from the logistic regression model.

The ROC AUC value of approximately 0.74 indicates the overall performance of the logistic regression model in distinguishing between diabetic and non-diabetic cases. An AUC value of 0.74 suggests that the model has a good but not perfect ability to discriminate between the two classes. In general, an AUC value closer to 1.0 indicates excellent model performance, while a value closer to 0.5 indicates performance no better than random chance.



```
# Decision Tree Classifier
tree_clf = DecisionTreeClassifier(random_state=42)
tree_clf.fit(X_train, y_train)
```

```
▾          DecisionTreeClassifier
DecisionTreeClassifier(random_state=42)
```



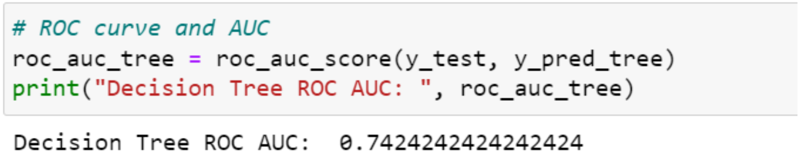
```python
# Confusion matrix and classification report
print("Decision Tree Confusion Matrix")
print(confusion_matrix(y_test, y_pred_tree))
print("Decision Tree Classification Report")
print(classification_report(y_test, y_pred_tree))
```

```
Decision Tree Confusion Matrix
[[75 24]
 [15 40]]
Decision Tree Classification Report
              precision    recall  f1-score   support

           0       0.83      0.76      0.79        99
           1       0.62      0.73      0.67        55

    accuracy                           0.75       154
   macro avg       0.73      0.74      0.73       154
weighted avg       0.76      0.75      0.75       154
```

Interpretation

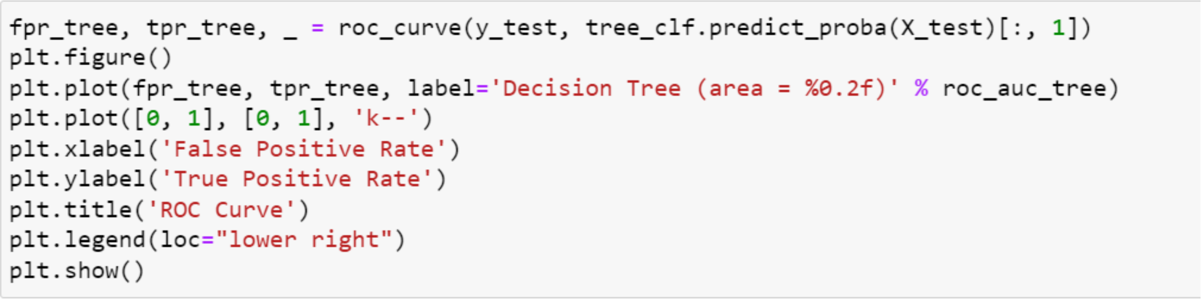
The confusion matrix for the decision tree model indicates that the model correctly predicted 75 instances of no diabetes (true negatives) and 40 instances of diabetes (true positives). However, it incorrectly predicted diabetes in 24 cases where there was none (false positives) and failed to predict diabetes in 15 cases where it was present (false negatives). These results show that while the decision tree model is effective at identifying both diabetic and non-diabetic cases, there are notable misclassifications, particularly with a higher number of false positives compared to the logistic regression model.

The classification report provides a detailed performance evaluation of the decision tree model. For the class representing no diabetes (0), the model achieved a precision of 83% and a recall of 76%, resulting in an F1-score of 79%. For the diabetes class (1), the precision was 62% and the recall was 73%, leading to an F1-score of 67%. The overall accuracy of the model was 75%, indicating that 75% of the predictions were correct. The macro and weighted averages provide a balanced view of performance across both classes. These metrics suggest that while the decision tree model performs well overall, it has a higher precision and slightly lower recall for non-diabetic cases, and it better captures diabetes cases than the logistic regression model, highlighting its effectiveness in identifying true positives.

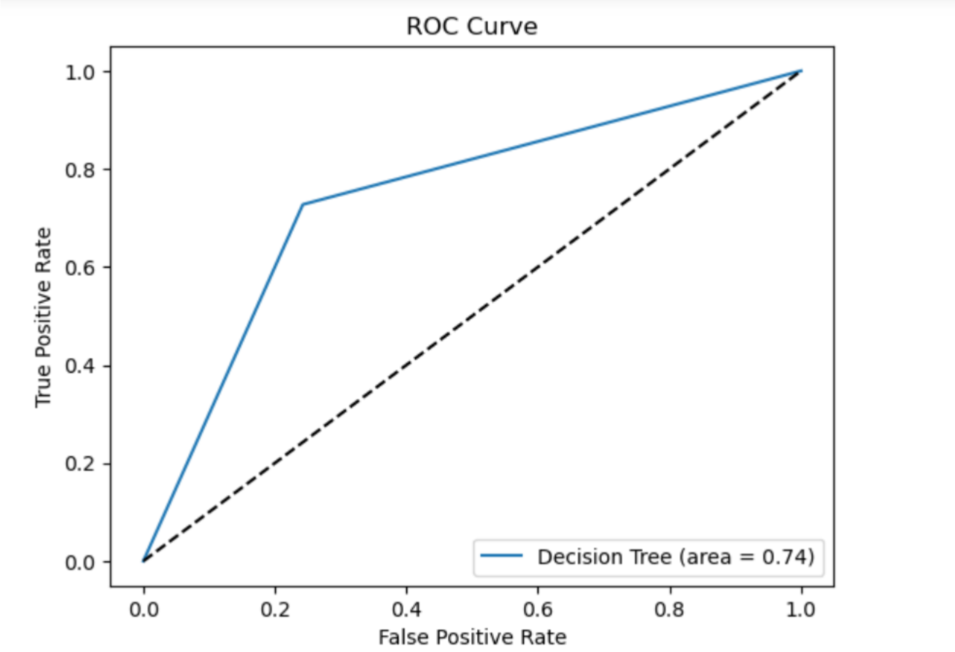


```python
# ROC curve and AUC
roc_auc_tree = roc_auc_score(y_test, y_pred_tree)
print("Decision Tree ROC AUC: ", roc_auc_tree)
```

```
Decision Tree ROC AUC:  0.7424242424242424
```

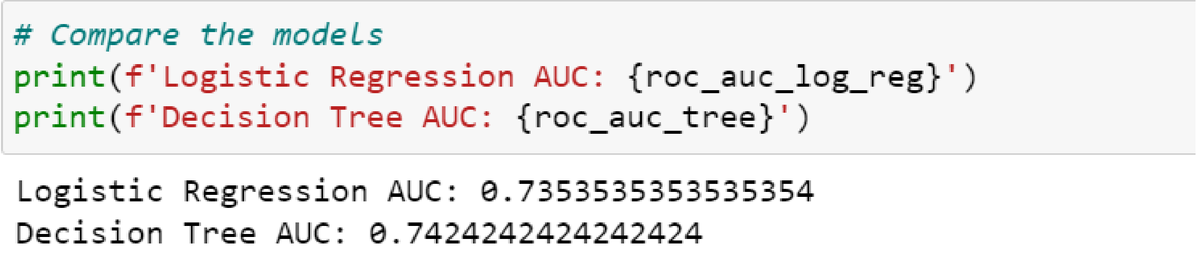


```
fpr_tree, tpr_tree, _ = roc_curve(y_test, tree_clf.predict_proba(X_test)[:, 1])
plt.figure()
plt.plot(fpr_tree, tpr_tree, label='Decision Tree (area = %0.2f)' % roc_auc_tree)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend(loc="lower right")
plt.show()
```



Interpretation

The decision tree model demonstrates a reasonable level of performance with an AUC of 0.74, indicating it is fairly effective at distinguishing between patients with and without diabetes. This metric complements the previously discussed confusion matrix and classification report, providing a more comprehensive understanding of the model's predictive capabilities.



```
# Compare the models
print(f'Logistic Regression AUC: {roc_auc_log_reg}')
print(f'Decision Tree AUC: {roc_auc_tree}')
```

```
Logistic Regression AUC: 0.7353535353535354
Decision Tree AUC: 0.7424242424242424
```

Interpretation

Both models exhibit similar performance in distinguishing between diabetic and non-diabetic cases, with AUC values slightly above 0.73, indicating good but not perfect discriminatory power.

Logistic Regression

The AUC of approximately 0.73 suggests that the logistic regression model has a good ability to differentiate between the two classes.

Decision Tree

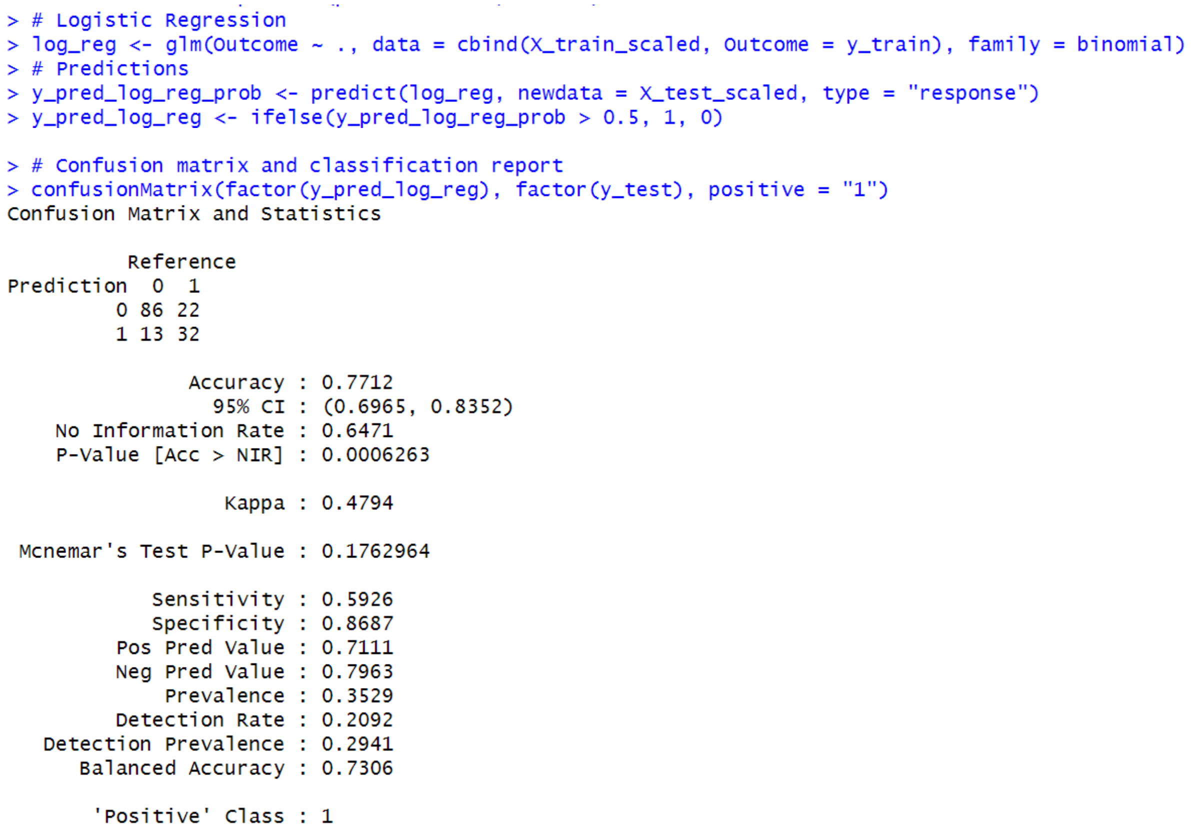
The decision tree model shows a slightly higher AUC of approximately 0.74, indicating marginally better performance compared to logistic regression.

Overall Comparison

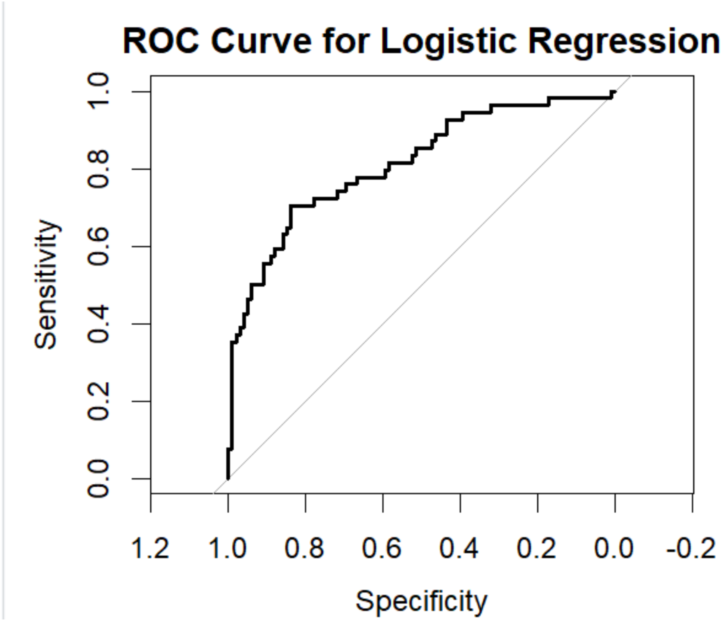
The marginal difference in AUC values (0.74 for the decision tree vs. 0.73 for logistic regression) suggests that both models perform similarly in terms of overall predictive accuracy.

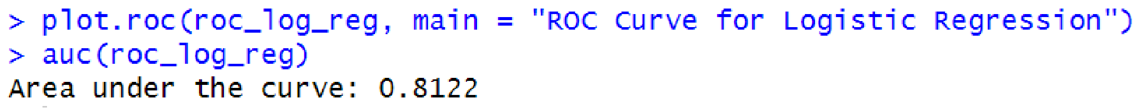
The choice between the two models may depend on other factors such as the importance of model interpretability, the ability to capture non-linear relationships, and the specific application context. For instance, if interpretability and understanding the influence of individual predictors are crucial, logistic regression might be preferred. Conversely, if the goal is to capture complex interactions and provide a straightforward decision-making process, a decision tree could be more suitable.

⇨ R

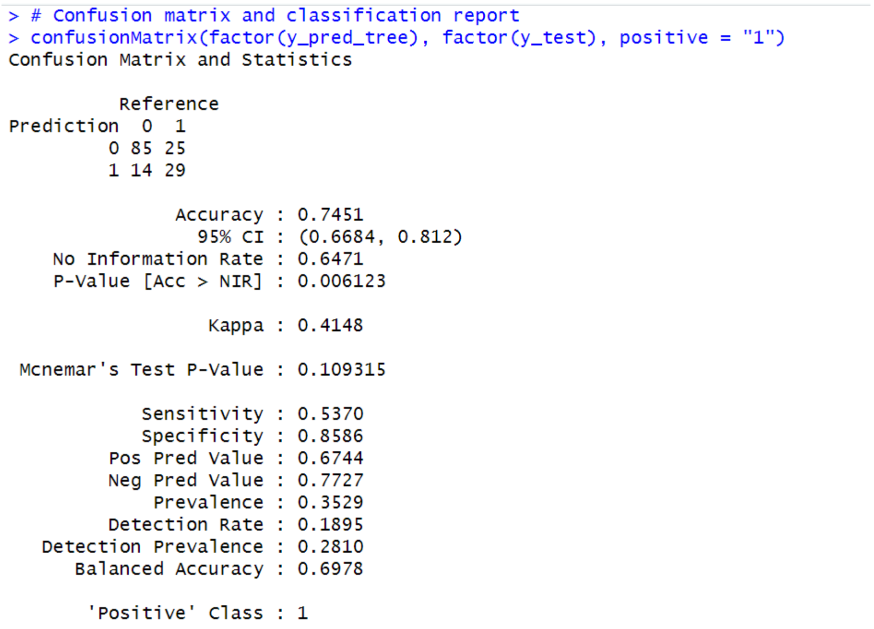


```
> # Logistic Regression
> log_reg <- glm(Outcome ~ ., data = cbind(X_train_scaled, Outcome = y_train), family = binomial)
> # Predictions
> y_pred_log_reg_prob <- predict(log_reg, newdata = X_test_scaled, type = "response")
> y_pred_log_reg <- ifelse(y_pred_log_reg_prob > 0.5, 1, 0)

> # Confusion matrix and classification report
> confusionMatrix(factor(y_pred_log_reg), factor(y_test), positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 86 22
         1 13 32

               Accuracy : 0.7712
                 95% CI : (0.6965, 0.8352)
    No Information Rate : 0.6471
    P-Value [Acc > NIR] : 0.0006263

                  Kappa : 0.4794

 Mcnemar's Test P-Value : 0.1762964

            Sensitivity : 0.5926
            Specificity : 0.8687
         Pos Pred Value : 0.7111
         Neg Pred Value : 0.7963
             Prevalence : 0.3529
         Detection Rate : 0.2092
   Detection Prevalence : 0.2941
      Balanced Accuracy : 0.7306

       'Positive' Class : 1
```

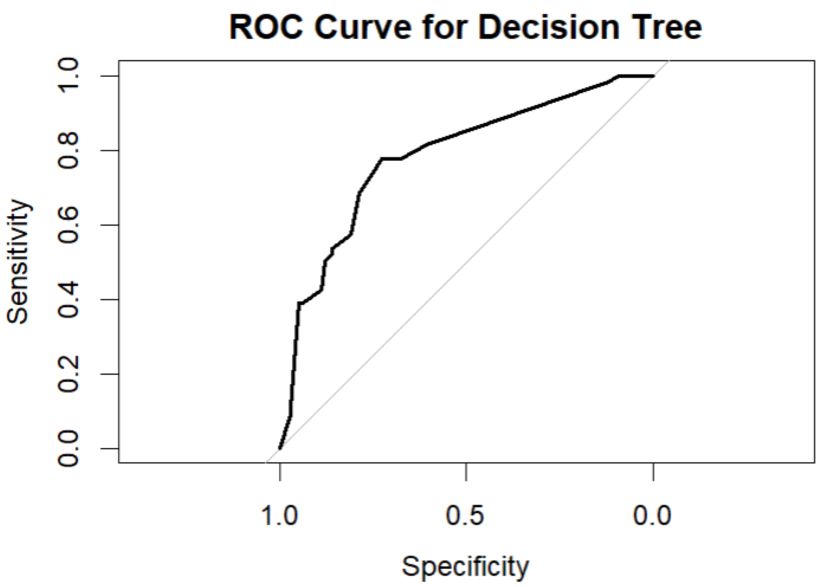


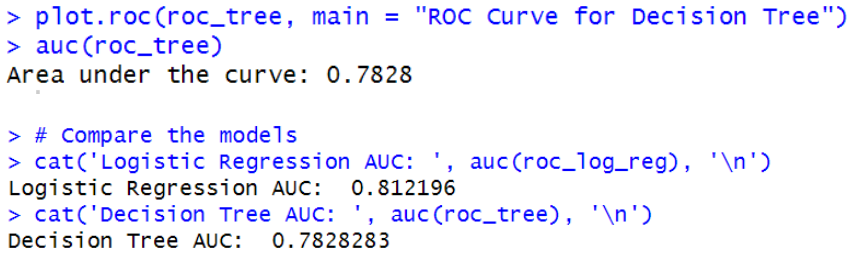


```
> plot.roc(roc_log_reg, main = "ROC Curve for Logistic Regression")
> auc(roc_log_reg)
Area under the curve: 0.8122
```



```
> # Confusion matrix and classification report
> confusionMatrix(factor(y_pred_tree), factor(y_test), positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 85 25
         1 14 29
                                          
               Accuracy : 0.7451          
                 95% CI : (0.6684, 0.812) 
    No Information Rate : 0.6471          
    P-Value [Acc > NIR] : 0.006123        
                                          
                  Kappa : 0.4148          
                                          
 Mcnemar's Test P-Value : 0.109315        
                                          
            Sensitivity : 0.5370          
            Specificity : 0.8586          
         Pos Pred Value : 0.6744          
         Neg Pred Value : 0.7727          
             Prevalence : 0.3529          
         Detection Rate : 0.1895          
   Detection Prevalence : 0.2810          
      Balanced Accuracy : 0.6978          
                                          
       'Positive' Class : 1               
```





```
> plot.roc(roc_tree, main = "ROC Curve for Decision Tree")
> auc(roc_tree)
Area under the curve: 0.7828

> # Compare the models
> cat('Logistic Regression AUC: ', auc(roc_log_reg), '\n')
Logistic Regression AUC:  0.812196
> cat('Decision Tree AUC: ', auc(roc_tree), '\n')
Decision Tree AUC:  0.7828283
```

Interpretation

The provided AUC values for the logistic regression and decision tree models are 0.812196 and 0.7828283 respectively.

Logistic Regression

- The AUC of approximately 0.81 indicates that the logistic regression model has a strong ability to differentiate between diabetic and non-diabetic cases.
- This higher AUC value suggests that the logistic regression model is effective in predicting the likelihood of diabetes, making it a reliable choice for classification tasks where model interpretability and understanding of individual predictors are important.
- Logistic regression provides clear insights into the relationship between each predictor variable and the outcome, which can be valuable for healthcare providers in understanding risk factors for diabetes.

Decision Tree

- The AUC of approximately 0.78 for the decision tree model indicates good performance but slightly lower than that of the logistic regression model.
- Decision trees are useful for capturing complex, non-linear relationships between variables and offer a visual representation of the decision-making process. This can be advantageous for intuitive understanding and explanation.

- Despite its slightly lower AUC, the decision tree model still demonstrates reasonable accuracy and may be preferred in scenarios where interpretability and capturing interactions between variables are crucial.

Overall Comparison

- The logistic regression model outperforms the decision tree model with a higher AUC (0.81 vs. 0.78), suggesting that it has a better overall ability to correctly classify diabetic and non-diabetic cases.
- The logistic regression model's higher AUC reflects its effectiveness in prediction and its robustness in handling linear relationships between the predictors and the outcome.
- On the other hand, the decision tree model, while slightly less accurate in terms of AUC, offers advantages in terms of visual interpretability and the ability to model complex relationships.