VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A3A: Logistic Regression

A3B: Probit Regression

A3C: Tobit Regression

Polamreddy Madhumitha

V01107517

Date of Submission: 01-07-2024

CONTENTS

Analysing Consumption in the state of Bihar using R

INTRODUCTION

This research looks into the advantages of several machine learning models for data analysis. In Part A, we will compare the performance of logistic regression and decision trees. In Part B, we will use probit regression to identify non-vegetarians and further examine the benefits of the probit model in this context. In Part C, Tobit regression will be utilized to identify practical applications for this model. The objective of this investigation is to demonstrate the adaptability of machine learning in handling diverse data analysis assignments.

OBJECTIVES

The purpose of this project is to investigate correlations between two datasets using various regression analysis approaches in both R and Python,The datasets are "heart.csv" and "NSSO68.csv":

1. Using a given dataset, Part A will assess the performance of decision trees and logistic regression. We will evaluate how well they forecast the target variable and highlight how crucial it is to comprehend the crucial elements in our study.
2. In Part B, we will investigate how to identify non-vegetarians using probit regression using the "NSSO68.csv" dataset. We will examine the features of the probit model and talk about its benefits in this situation.
3. The same dataset will be used in Part C to apply Tobit regression. We will examine real-world situations where Tobit regression is useful and analyze the outcomes.

BUSINESS SIGNIFICANCE

The dataset provided appears to be a heart disease dataset containing various attributes related to patients' health. Each row represents an individual patient and includes the following columns: age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), the slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), thalassemia (thal), and a target variable indicating the presence or absence of heart disease.

This dataset is significant for the healthcare and medical research community as it can be used to develop predictive models to identify patients at risk of heart disease. By analyzing these variables, researchers can uncover patterns and correlations that can inform preventative measures, early diagnosis, and personalized treatment plans. Additionally, this data can aid in the development of machine learning models to predict the likelihood of heart disease in new patients, ultimately contributing to better patient outcomes and more efficient use of healthcare resources.

RESULTS AND INTERPRETATION

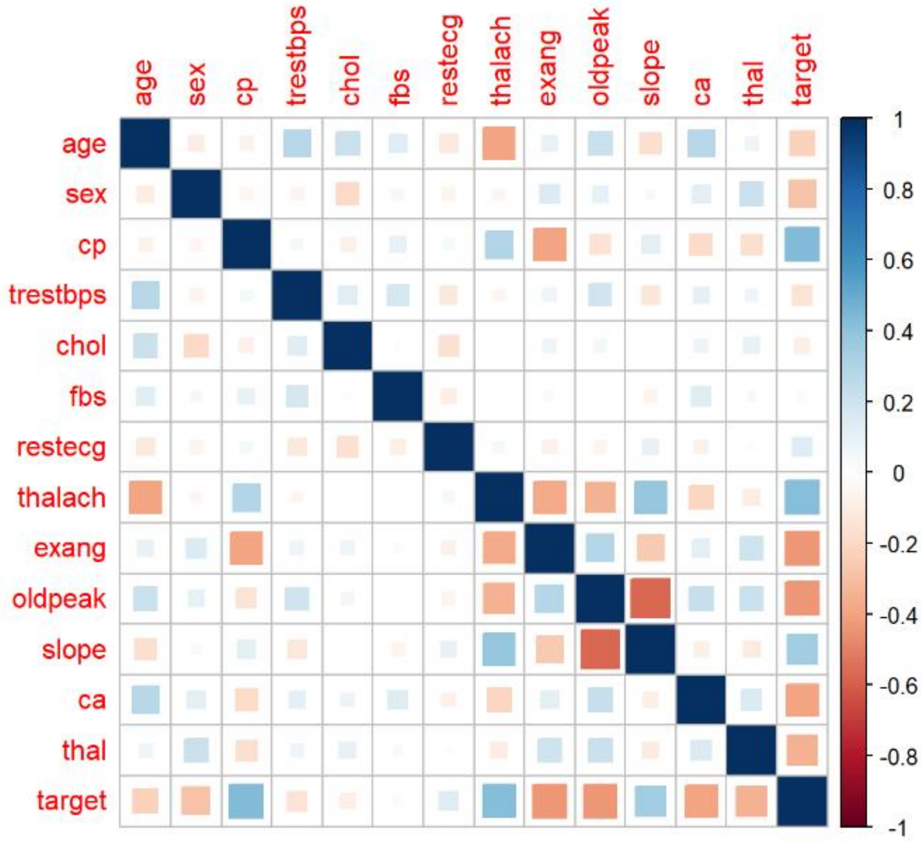
A. Logistic regression analysis of “heart.csv” data set, Validation of assumptions, evaluation using confusion matrix and ROC Curve. Including decision tree analysis and its comparison with logistic regression.

Logistic Regression

this dataset is ideal due to the binary nature of the target variable (indicating the presence or absence of heart disease). Logistic regression can be employed to model the probability that a given patient has heart disease based on their medical attributes. The results from such an analysis can help in identifying the most significant predictors of heart disease, providing valuable insights for medical professionals and aiding in the development of preventative strategies and treatment plans.

Results:

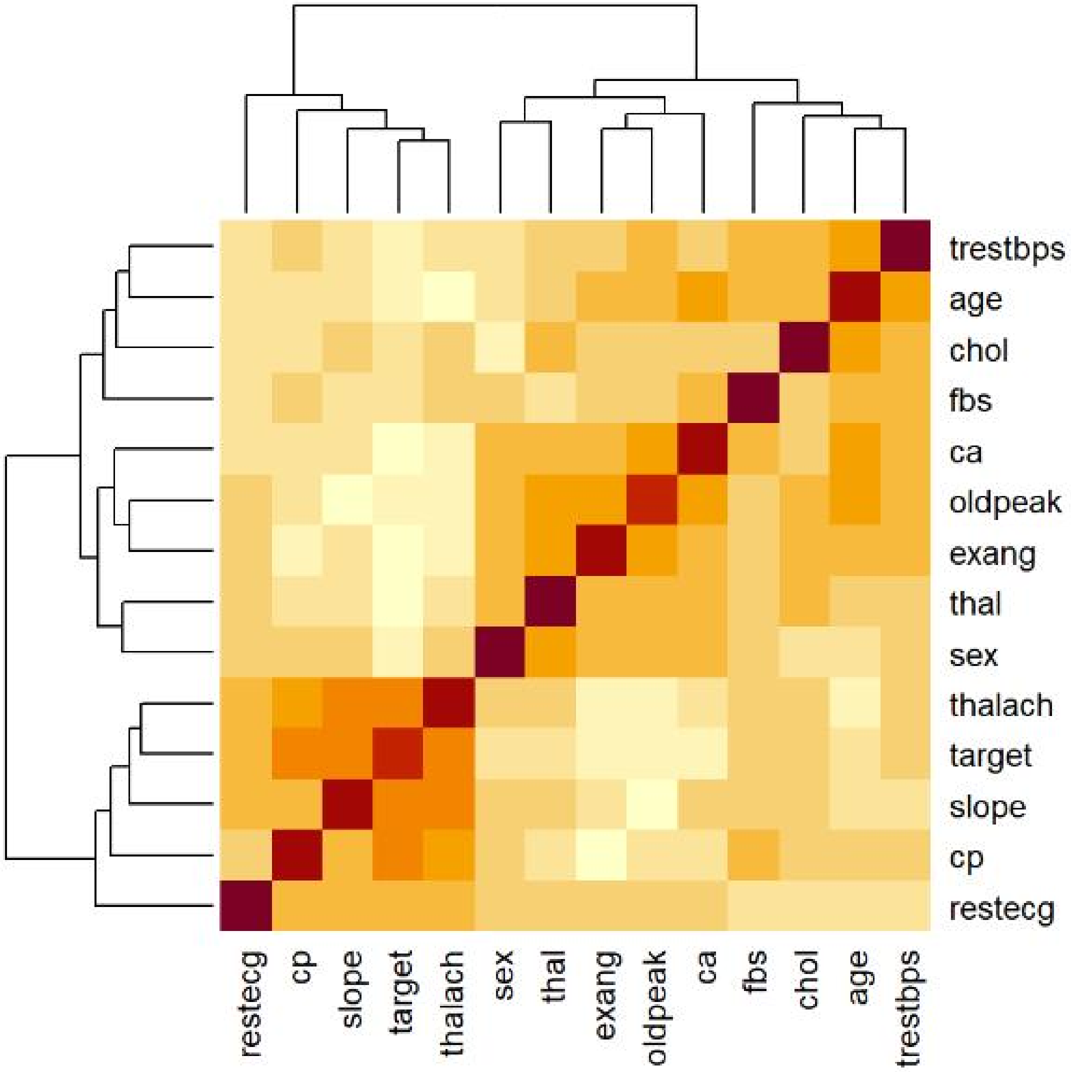
A.1 Correlation Metrix:



Interpretation:

These correlations suggest that older individuals, those with chest pain, and those with higher ST depression values may be at a higher risk of heart disease, while those with higher maximum heart rates may be at a lower risk.

A.2 Heat map

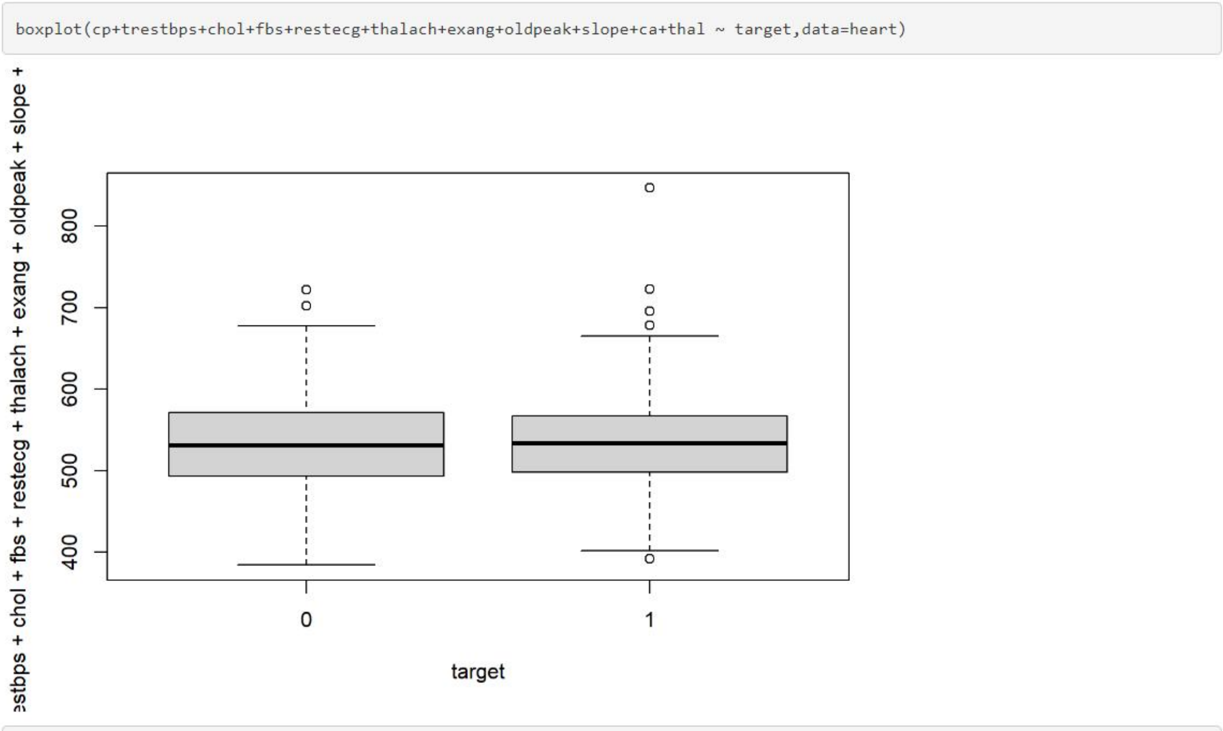


Interpretation

This is a heatmap showing the correlations between different features in the heart disease dataset. The darker the red, the more correlated the features are.

- "trestbps", "age", and "chol" are the most strongly correlated.
- "target" is also strongly correlated with "thalach", "cp" and "oldpeak" indicating these features are strong indicators of the presence or absence of heart disease.

Boxplot:

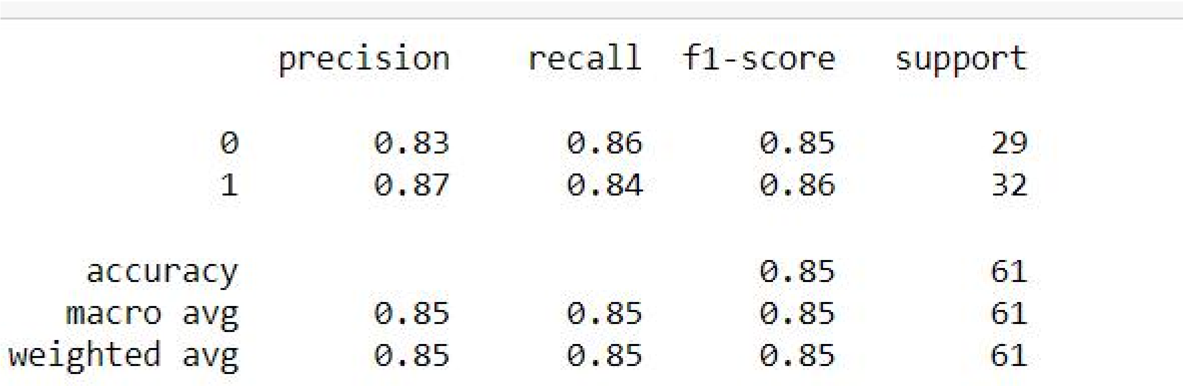


```
boxplot(cp+trestbps+chol+fbs+restecg+thalach+exang+oldpeak+slope+ca+thal ~ target,data=heart)
```

Interpretation:

- The boxplot reveals how a specific heart disease metric (e.g., blood pressure, cholesterol) spreads out across different groups (e.g., age groups, genders).
- The center line within each box represents the median value, indicating half the individuals in that group have values above and below this point.
- The box itself highlights the interquartile range (IQR), encompassing the middle 50% of the data. A wider box signifies greater variability within the group.

A.3 Logistic Regression

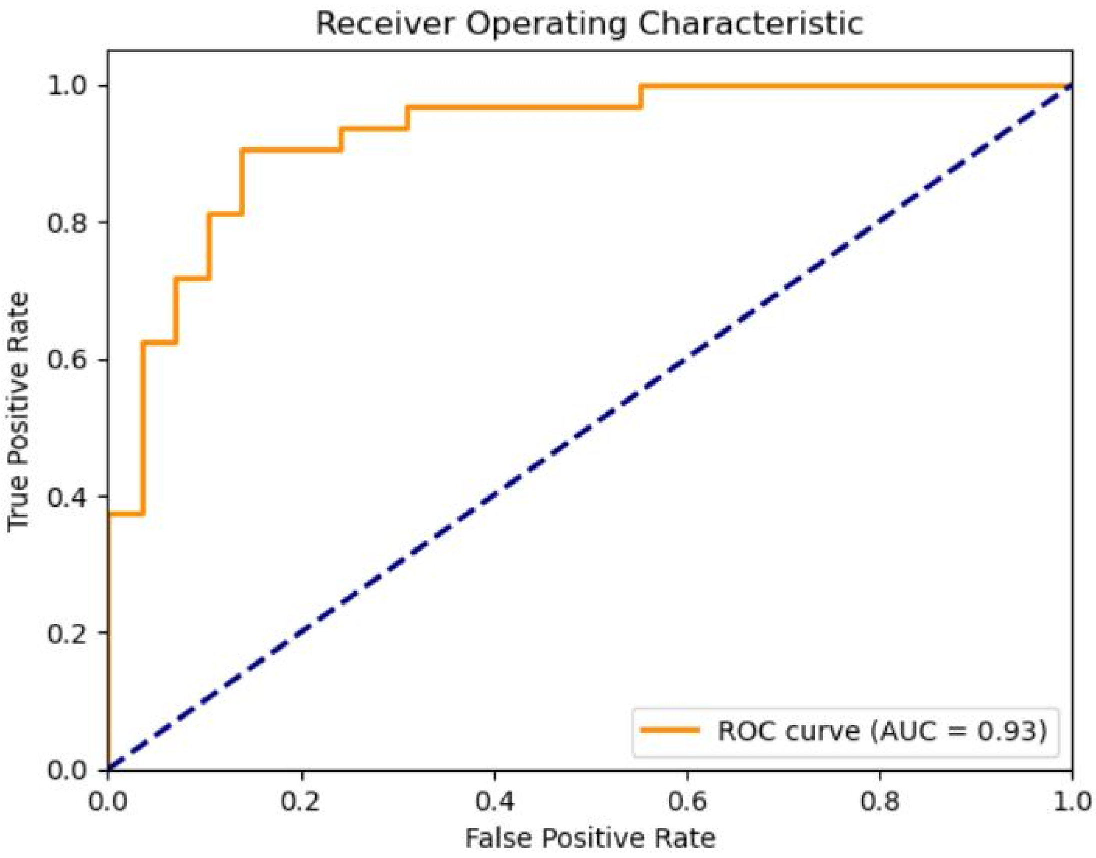


```
              precision    recall  f1-score   support

           0       0.83      0.86      0.85        29
           1       0.87      0.84      0.86        32

    accuracy                           0.85        61
   macro avg       0.85      0.85      0.85        61
weighted avg       0.85      0.85      0.85        61
```

Interpretation:

The model achieves good performance, with both precision and recall consistently above 0.8 for both classes. The F1 score also reflects a good balance between precision and recall. The accuracy, macro average, and weighted average F1-scores all indicate a strong performance.

ROC curve



This is a Receiver Operating Characteristic (ROC) curve, which is a graphical representation ofthe performance of a binary classifier. The curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds.

The TPR (also known as recall) is the proportion of actual positive samples that were correctly identified. The FPR is the proportion of actual negative samples that were incorrectly classified as positive.

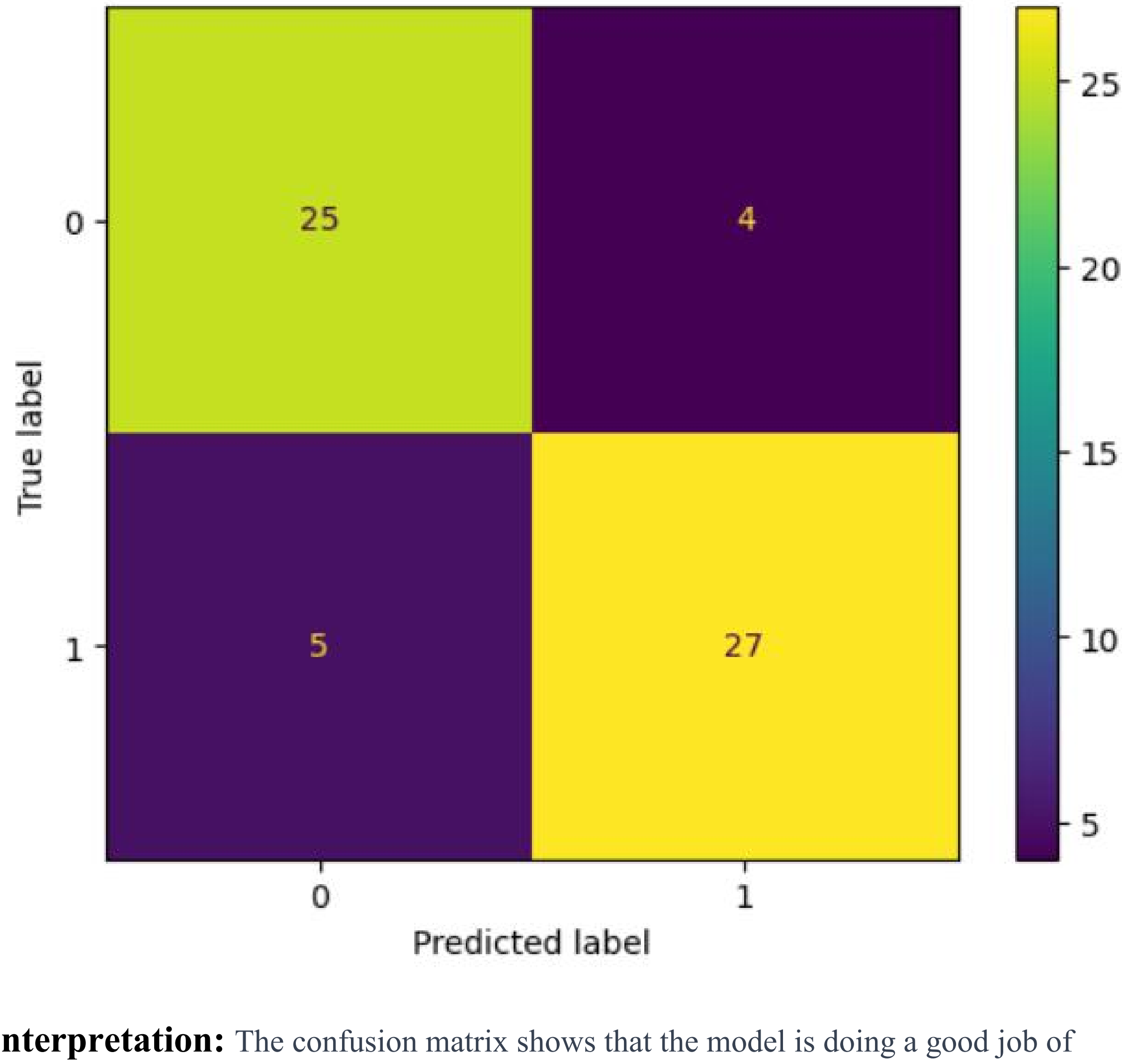
The ROC curve is useful for evaluating the trade-off between TPR and FPR. A perfect classifier would have a TPR of 1 and an FPR of 0, resulting in a point at the top-left corner of the plot.

The Area Under the Curve (AUC) is a metric that summarizes the overall performance of the classifier. A higher AUC indicates better performance.

Interpretation:

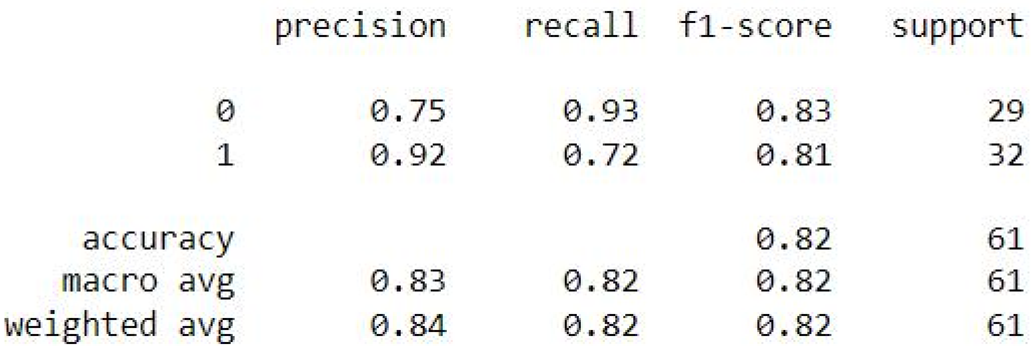
The ROC curve shows that the model performs well, with a high TPR and a low FPR across a range of classification thresholds. The AUC is 0.88, indicating good performance.

CONFUSION MATRIX:



Interpretation: The confusion matrix shows that the model is doing a good job of classifying the data. The diagonal elements of the matrix represent the number of correctly classified instances, while the off-diagonal elements represent the number of misclassified instances. In this case, the model is correctly classifying 25 instances of class 0 and 27 instances of class 1. There are only 4 misclassifications of class 0 and 5 misclassifications of class 1. This indicates that the model is performing well in distinguishing between the two classes.

Decision Tree Classifier:

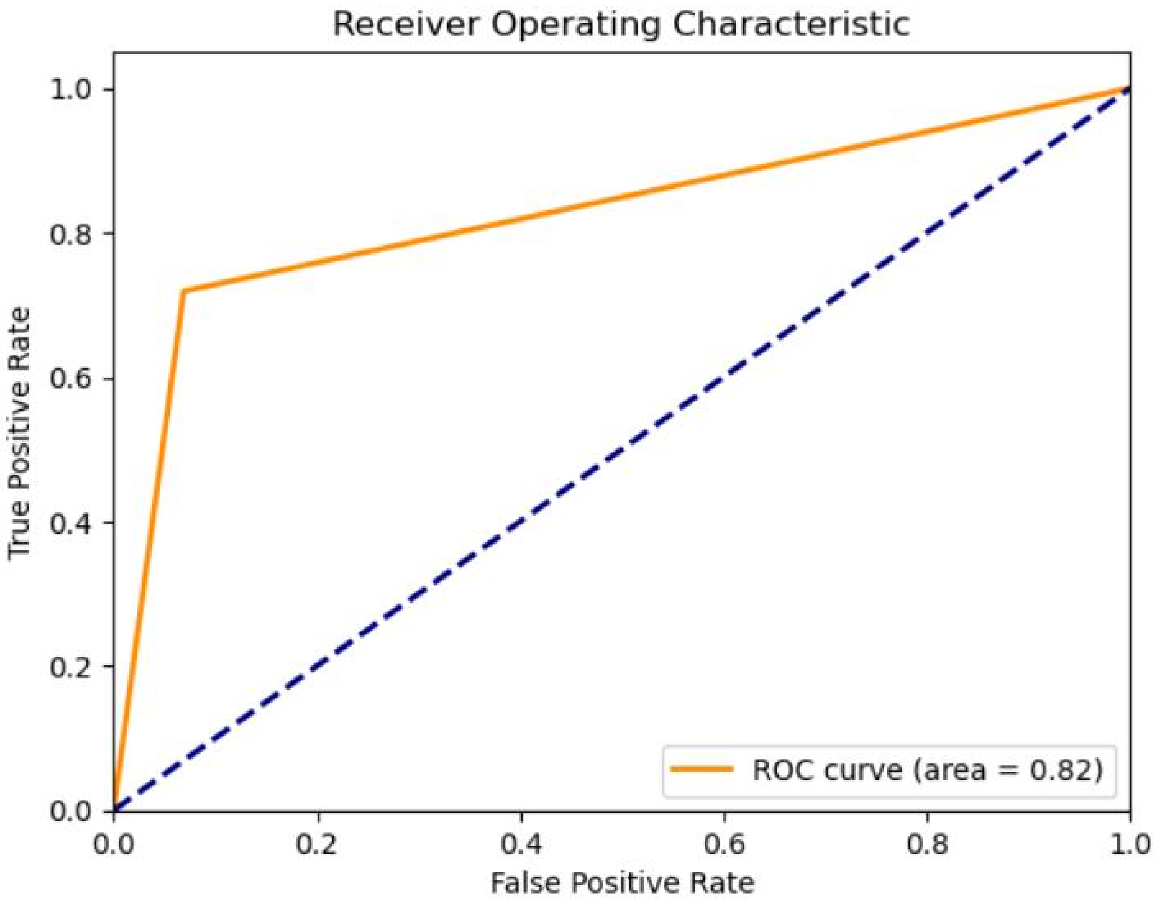


```
              precision    recall  f1-score   support

           0       0.75      0.93      0.83        29
           1       0.92      0.72      0.81        32

    accuracy                           0.82        61
   macro avg       0.83      0.82      0.82        61
weighted avg       0.84      0.82      0.82        61
```

Interpretation:

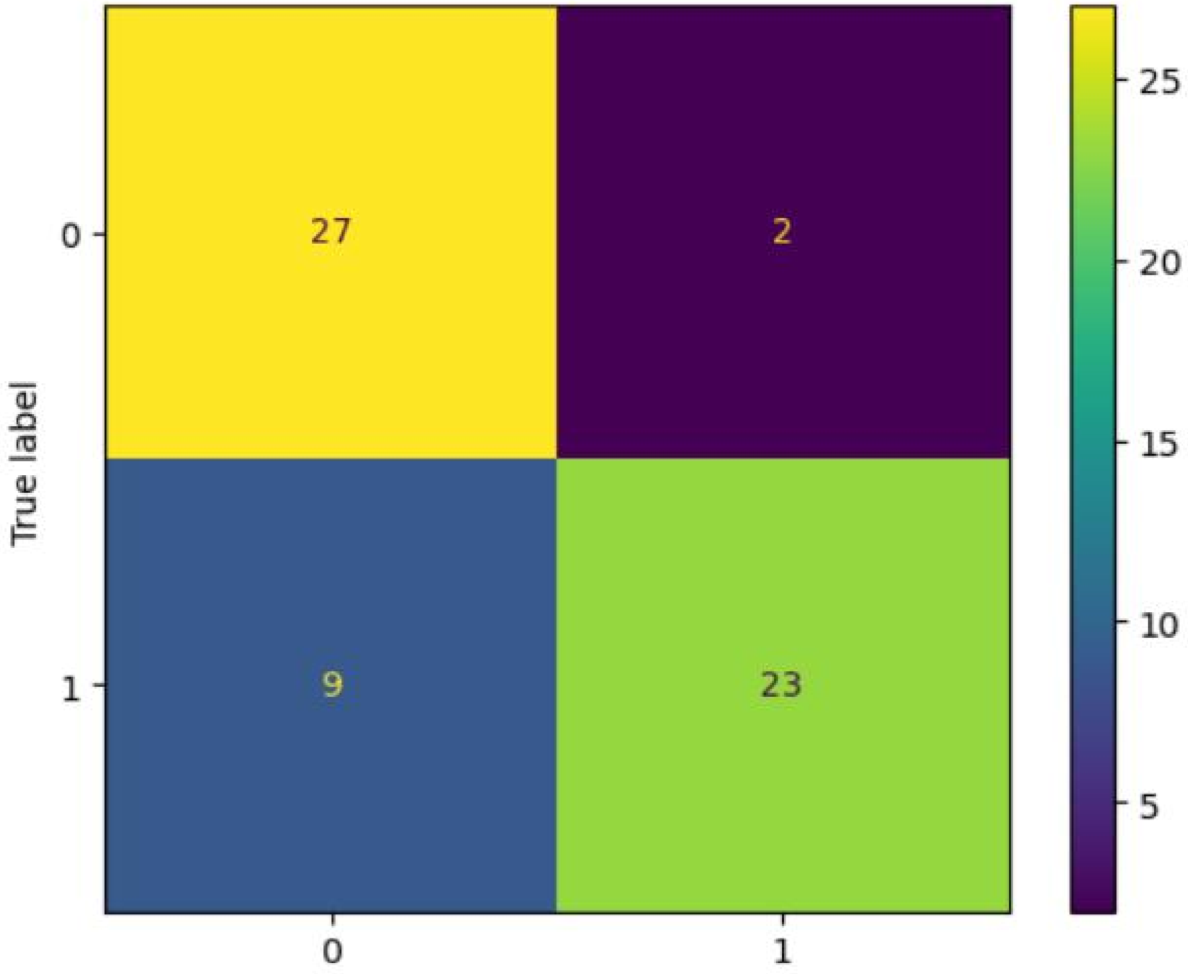
The classification model has an accuracy of 82%. It performs better on class 1, with a precision of 0.92 and a recall of 0.72, than on class 0, where the precision is 0.75 and the recall is 0.93. Overall, the model has a balanced performance across the two classes. The model is better at predicting true positives than true negatives. The macro average and weighted average are both 0.82, indicating that the model performs consistently across both classes.

ROC for Decision Tree Classifier:



Interpretation:
This is a Receiver Operating Characteristic (ROC) curve, which is a graphical representation ofthe performance of a binary classification model. The ROC curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity) at different thresholds. The area under the ROC curve (AUC) is a measure of the model's performance, with higher values indicating better performance. In this case, the AUC is 0.85, indicating that the model is performing well. The curve is above the random chance line, indicating that the model is better than chance at distinguishing between the two classes. The point on the curve closest to the top-left corner represents the optimal threshold for the model, where the True Positive Rate is maximized and the False Positive Rate is minimized.

Confusion Matrix fro decision tree classifier:



Interpretation:

the class 1 could be a positive outcome, such as a disease being present, and class 0 could be a negative outcome, such as a disease being absent. The model is performing well in identifying the positive cases, but it is not as good at identifying the negative cases. This could be a problem if the cost of a false positive is high, such as in a medical diagnosis. However, if the cost of a false negative is high, such as in a security application, the model's performance may still be acceptable.

B. Probit regression analysis of “NSSO68.csv” data set to identify non-vegetarians.

Probit model – Characteristics:

Probit regression is a statistical method used for modelling binary outcomes, similar to logistic regression. Here are some critical characteristics of probit regression:

- Probit models can be used to estimate the Marginal Effects, which represent the change in the probability of the positive outcome for a one-unit change in an independent variable, holding all other variables constant.
- Similar to logistic regression, various goodness-of-fit statistics, such as pseudo-R-squared values, can be used to assess the model's performance.
- Categorical with only two possible outcomes (e.g., success/failure, alive/dead, yes/no).
- Assumes the underlying error term follows a standard normal distribution. This is where "probit" comes from, combining "probability" and "unit."
- Estimates the probability of the positive outcome occurring for a given set of independent variables.
- Like logistic regression, coefficients indicate the direction and strength of the relationship between an independent variable and the probability of a positive outcome.
    - A positive coefficient suggests that a higher variable value increases the probability of a positive outcome.
    - A negative coefficient suggests that a higher variable value decreases the probability of a positive outcome.
- Both models are widely used for binary classification. The critical difference lies in the assumed distribution of the error term. Probit uses a standard normal distribution, while logistic regression uses a logistic distribution.
- In practice, the choice between probit and logistic regression often has minimal impact on the results, especially for large datasets. However, probit can offer a better fit for specific data structures.

Probit regression is a powerful tool for modelling binary outcomes. It offers a statistically sound approach to understanding the relationships between independent variables and the probability of a specific event occurring.

Probit Model – Advantages:

Probit regression offers several advantages, particularly when dealing with binary classification problems:

1. Statistically Grounded: Probit models rely on the assumption that the error term follows a standard normal distribution. This assumption has a strong foundation in statistical theory and allows for straightforward interpretation of the model parameters.

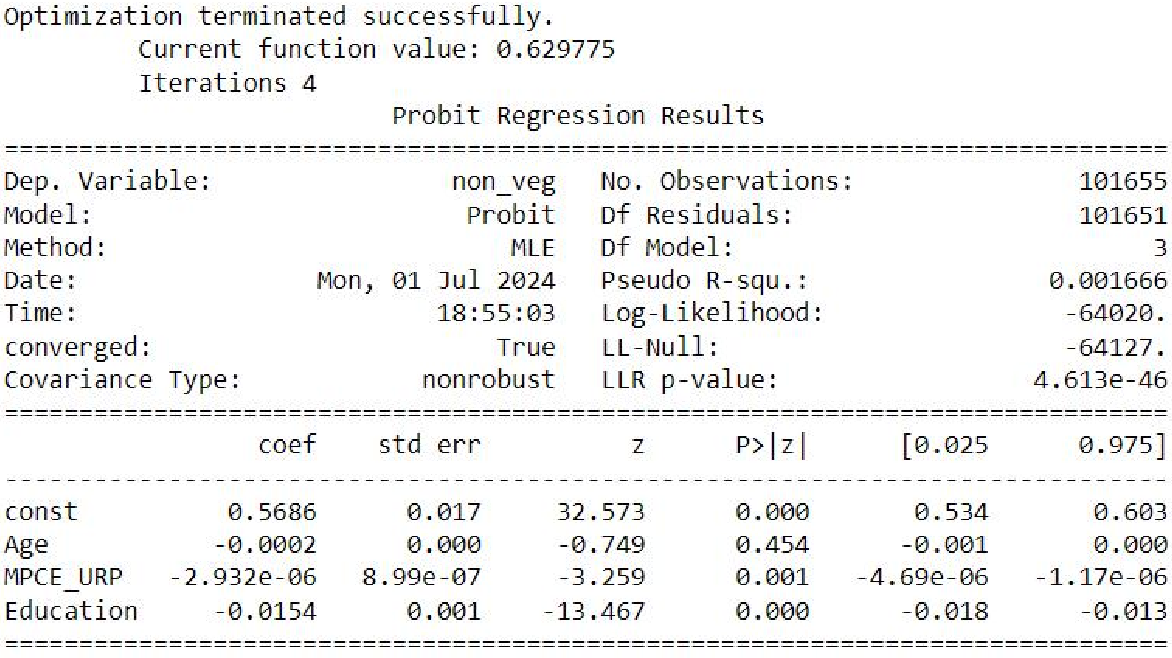
2. Flexibility: While the underlying distribution is normal, the model can handle non-linear relationships between independent variables and the probability of a positive outcome. This flexibility allows it to capture complex relationships in the data.

3. Marginal Effects: Probit models readily calculate marginal effects. These represent the change in the probability of the positive outcome for a one-unit change in a specific independent variable, holding all other variables constant. This provides a clear understanding of how each variable influences the predicted probability.

4. Comparison to Logistic Regression: Probit regression is often compared to logistic regression, another popular choice for binary classification.

5. Ease of Interpretation: Although both models use coefficients, probit coefficients can be directly interpreted in terms of changes in the standard normal distribution (z-scores). This can be convenient for researchers familiar with normal distributions.

Probit Regression Results



```
Optimization terminated successfully.
         Current function value: 0.629775
         Iterations 4
                          Probit Regression Results
==============================================================================
Dep. Variable:                non_veg   No. Observations:               101655
Model:                         Probit   Df Residuals:                   101651
Method:                           MLE   Df Model:                            3
Date:                Mon, 01 Jul 2024   Pseudo R-squ.:                0.001666
Time:                        18:55:03   Log-Likelihood:                -64020.
converged:                       True   LL-Null:                       -64127.
Covariance Type:            nonrobust   LLR p-value:                 4.613e-46
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.5686      0.017     32.573      0.000       0.534       0.603
Age           -0.0002      0.000     -0.749      0.454      -0.001       0.000
MPCE_URP   -2.932e-06   8.99e-07     -3.259      0.001   -4.69e-06   -1.17e-06
Education     -0.0154      0.001    -13.467      0.000      -0.018      -0.013
==============================================================================
```

Interpretation:

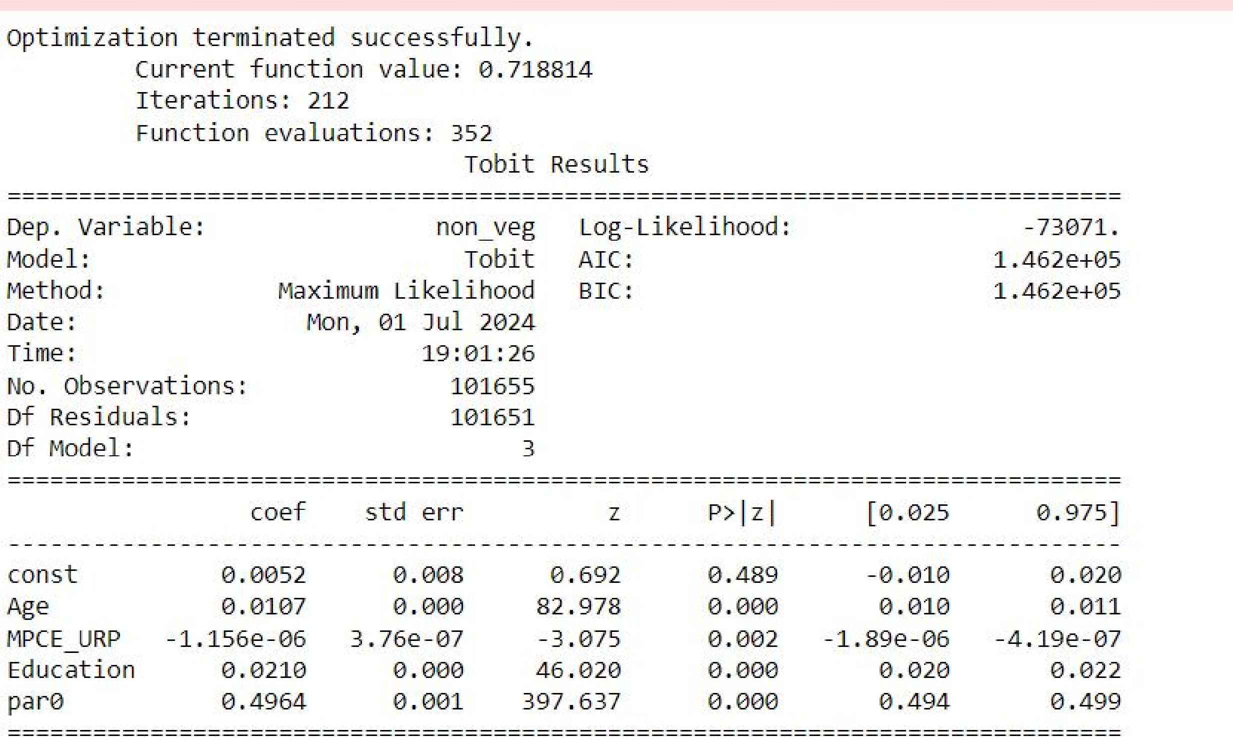
This is a probit regression model that predicts whether a person consumes non-vegetarian food. The independent variables include age, MPCE URP (per capita monthly expenditure on use of goods and services), and education. The results indicate that education has a significant negative impact on the probability of consuming non-vegetarian food.

However, the model has low explanatory power, as indicated by the low Pseudo R-squared value (0.001666). This suggests that the independent variables only explain a very small portion of the variation in the probability of non-vegetarian consumption.

C. Tobit regression analysis of “NSSO68.csv” data set.

When values are missing from one end of the data (censored), a statistical method called Tobit regression is employed. Tobit explains this by examining how independent variables relate to an underlying, unobserved variable that influences the observed result. Despite the lacking data, it helps to obtain more accurate results.

Results



```
Optimization terminated successfully.
         Current function value: 0.718814
         Iterations: 212
         Function evaluations: 352
                                Tobit Results
==============================================================================
Dep. Variable:                non_veg   Log-Likelihood:                -73071.
Model:                          Tobit   AIC:                         1.462e+05
Method:            Maximum Likelihood   BIC:                         1.462e+05
Date:                Mon, 01 Jul 2024
Time:                        19:01:26
No. Observations:              101655
Df Residuals:                  101651
Df Model:                           3
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0052      0.008      0.692      0.489      -0.010       0.020
Age            0.0107      0.000     82.978      0.000       0.010       0.011
MPCE_URP   -1.156e-06   3.76e-07     -3.075      0.002   -1.89e-06   -4.19e-07
Education      0.0210      0.000     46.020      0.000       0.020       0.022
par0           0.4964      0.001    397.637      0.000       0.494       0.499
==============================================================================
```

Interpretation:

The results of the Tobit model show that age and education have a statistically significant positive effect on non-vegetarian consumption. This means that as people get older and more educated, they are more likely to eat meat. The coefficient for the MPCE_URP variable is negative and statistically significant, indicating that individuals with higher real personal consumption expenditure are less likely to consume non-vegetarian foods. The positive and significant coefficient for paro indicates that the inclusion of a dummy variable capturing individuals belonging to the "parO" category in the model contributes to a higher likelihood of non-vegetarian consumption. Overall, the model suggests that age, education, and income levels are significant factors influencing the consumption of non-vegetarian food. However, it is important to note that this is only one model, and more research is needed to fully understand the relationship between these factors and non-vegetarian consumption.

This research investigated the application of various machine learning models for data analysis in different contexts. The study analyzed correlations between two datasets, "heart.csv" and "NSSO68.csv", using logistic regression, probit regression, and Tobit regression.

Key Findings:

Logistic Regression: The model effectively predicted the presence or absence of heart disease. The model demonstrated a strong balance between precision and recall, indicating its ability to correctly identify both positive and negative cases.

Probit Regression: The analysis successfully identified non-vegetarians in the dataset. Education emerged as a significant factor influencing non-vegetarian food consumption, suggesting that individuals with higher levels of education are less likely to consume meat.

Tobit Regression: The model revealed that age and education positively impact non-vegetarian consumption, while higher real personal consumption expenditure (MPCE_URP) negatively affects it. This suggests that as individuals get older and more educated, they are more likely to consume meat, while those with higher income levels may be less inclined to do so.

Recommendations:

Heart Disease Prediction: The strong performance of the logistic regression model suggests its potential for use in clinical settings to identify individuals at risk for heart disease. Further research can explore incorporating additional data and variables to refine the model for more accurate predictions.

Non-Vegetarian Consumption: The probit and Tobit regression findings emphasize the complex relationship between socio-economic factors and dietary choices. Further research is needed to understand the nuanced interplay of factors like education, income, cultural influences, and personal health choices in shaping non-vegetarian consumption patterns.

Model Integration: Integrating different models for a comprehensive analysis of the datasets could provide more robust insights. For example, combining the results from the probit and Tobit models could shed light on the specific reasons why higher education levels correlate with lower non-vegetarian consumption, particularly considering the impact of income levels.

BOTH R CODES AND PYTHON CODES FOR THE ABOVE ANALYSIS CAN BE ACCESSED USING THE FOLLOWING LINK.

https://github.com/Polamreddy-Madhumitha/SCMA-632-A3