SOCIOECONOMIC PROFILING VIA CENSUS DATA ANALYSIS

BUAN 6356.005-BUSINESS ANALYTICS WITH R

GROUP 1 MEMBERS

- Sushma, Reddy Vutla
- Bakare, Ife Funmilola
- Kilaru, Priyankka
- Subbgari, Shreeyesh R

PROJECT OVERVIEW

OBJECTIVE

- Income Prediction: Build a model to predict whether an individual's income exceeds a certain threshold (e.g., $50,000) based on their demographic and employment-related features.
- Explore the relationship between income levels and other factors, such as race, age, education, and gender, to identify patterns

PROJECT MOTIVATION

- Understanding socioeconomic factors is crucial for policy-making and resource allocation.

- Utilizing census data to predict income levels based on various features.

DATA DESCRIPTION

DATASET DETAILS

- Source: Adult Dataset from UCI ML Repository (1994 Census database)
- Size: 48,842 instances, 14 features (6 numerical, 8 categorical)
- Training Set: 32,561 instances, Test Set: 16,281 instances

KEY FEATURES FOR ANALYSIS

- Age (continuous)
- Education (categorical)
- Occupation (categorical)
- Marital Status (categorical)
- Capital Gain (continuous)
- Capital Loss (continuous)

DATA CLEANING AND QUALITY CHECK

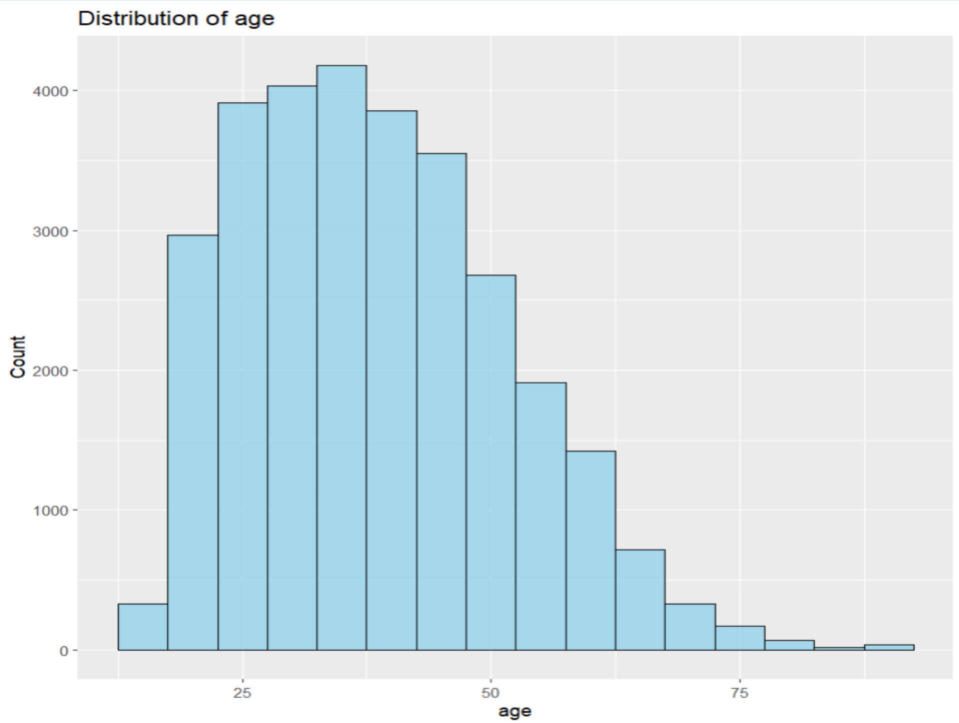
Steps Taken:

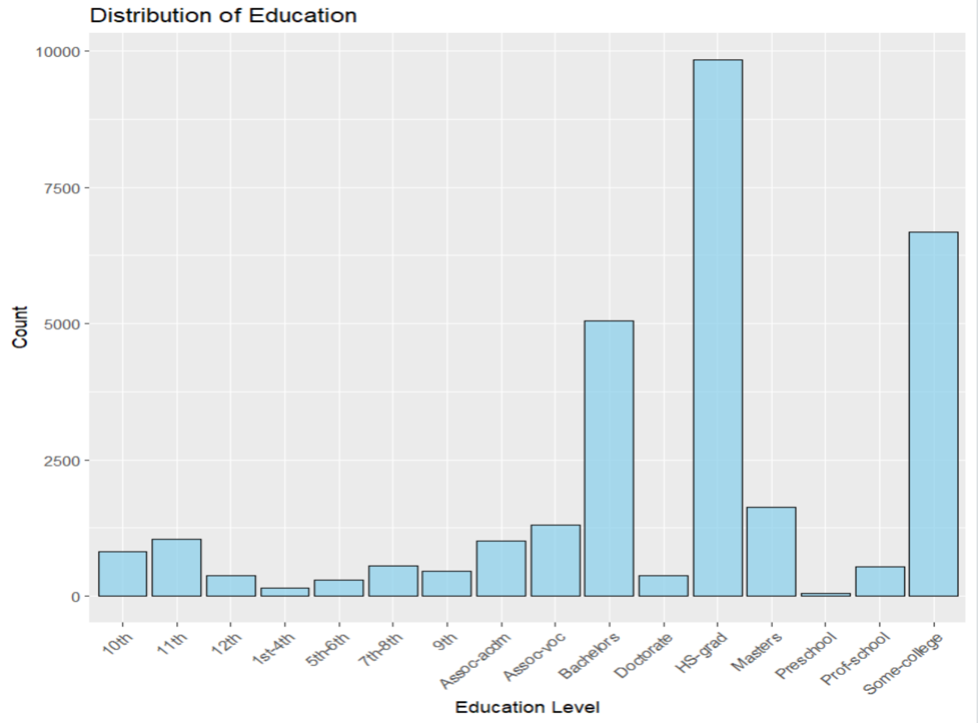
- Handling Missing Values:
  - Replaced '?' with NA in both training and testing sets.

- Identifying and Converting Character Columns:
  - Converted common character columns to factors.

- Imputing Missing Values:
  - Mean for numerical columns, mode for categorical columns.

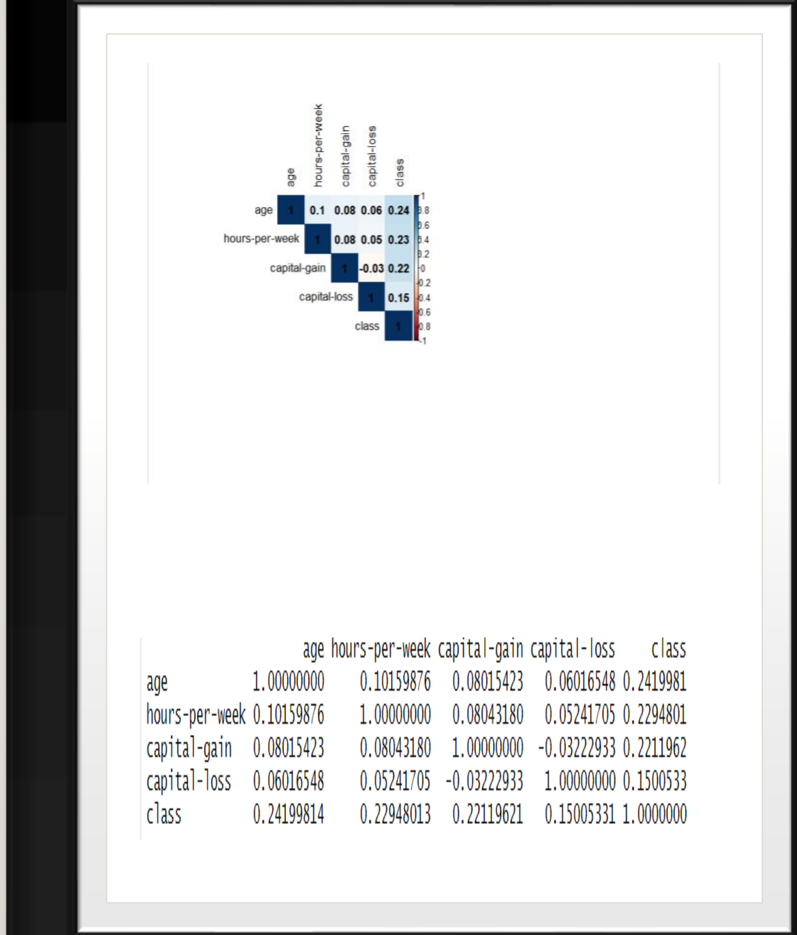
HISTOGRAM DISTRIBUTION





CORRELATION ANALYSIS

- Key Insights:
- Positive correlation (0.24) between age and income class.
- Moderate positive correlation (0.23) between hours worked per week and income class.
- Positive correlation (0.22) between capital gain and income class.
- Positive correlation (0.15) between capital loss and income class.
- The heatmap visually represents the correlation matrix. Darker colors indicate stronger correlations, enhancing the understanding of relationships between numerical



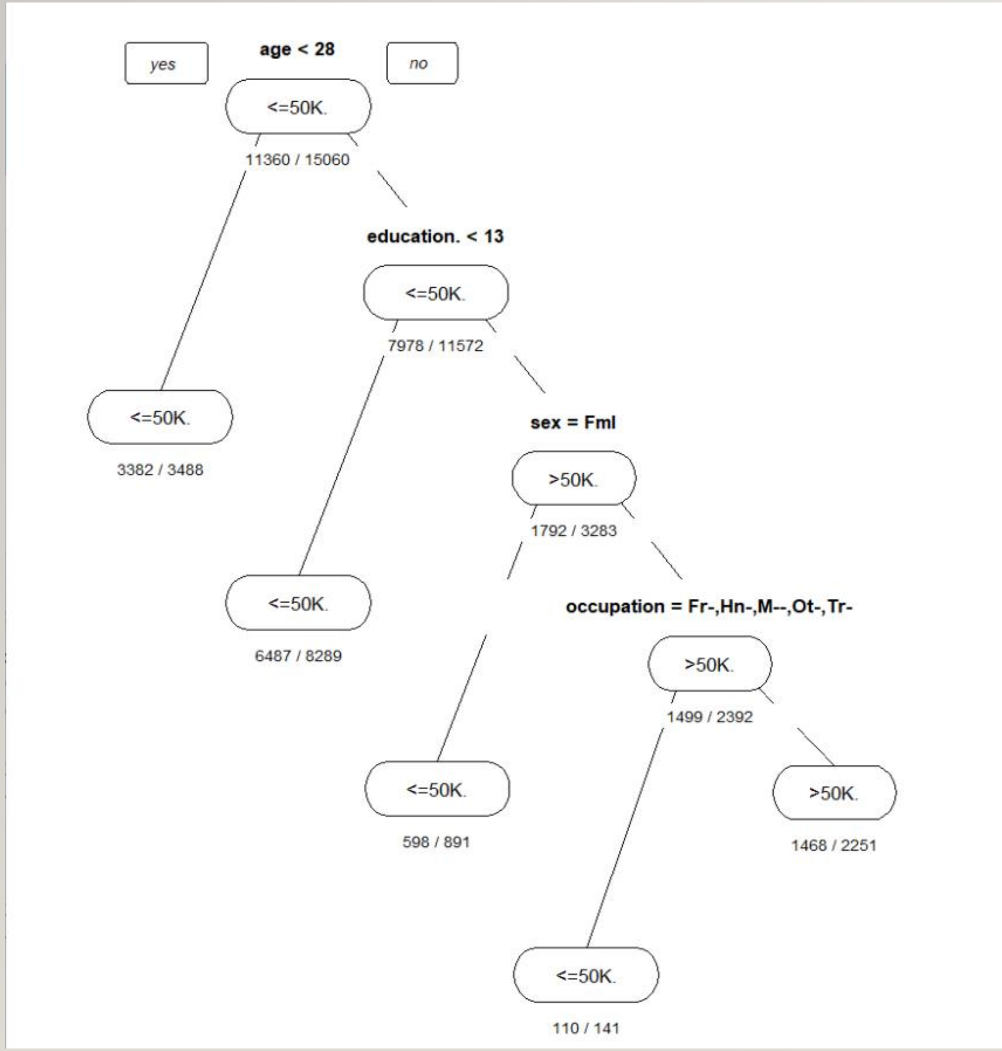
```
                     age hours-per-week capital-gain capital-loss     class
age           1.00000000     0.10159876   0.08015423   0.06016548 0.2419981
hours-per-week 0.10159876     1.00000000   0.08043180   0.05241705 0.2294801
capital-gain  0.08015423     0.08043180   1.00000000  -0.03222933 0.2211962
capital-loss  0.06016548     0.05241705  -0.03222933   1.00000000 0.1500533
class         0.24199814     0.22948013   0.22119621   0.15005331 1.0000000
```

Data Preprocessing

Objective:

- ➢ Remove unnecessary columns
- ➢ Encode categorical variables
- ✓ 1. Remove Unnecessary Columns:(*Enhances efficiency by focusing on relevant features)*
- Eliminated columns with indices fnlwgt, education, relationship, marital_status, capital_gain, and capital_loss and from both training and validation datasets.
- ✓ 2. Encode Categorical Variables:
- Converted categorical variables: workclass, occupation, race, sex, country, and class.

Decision Tree Model



❑ *Shows the income levels*

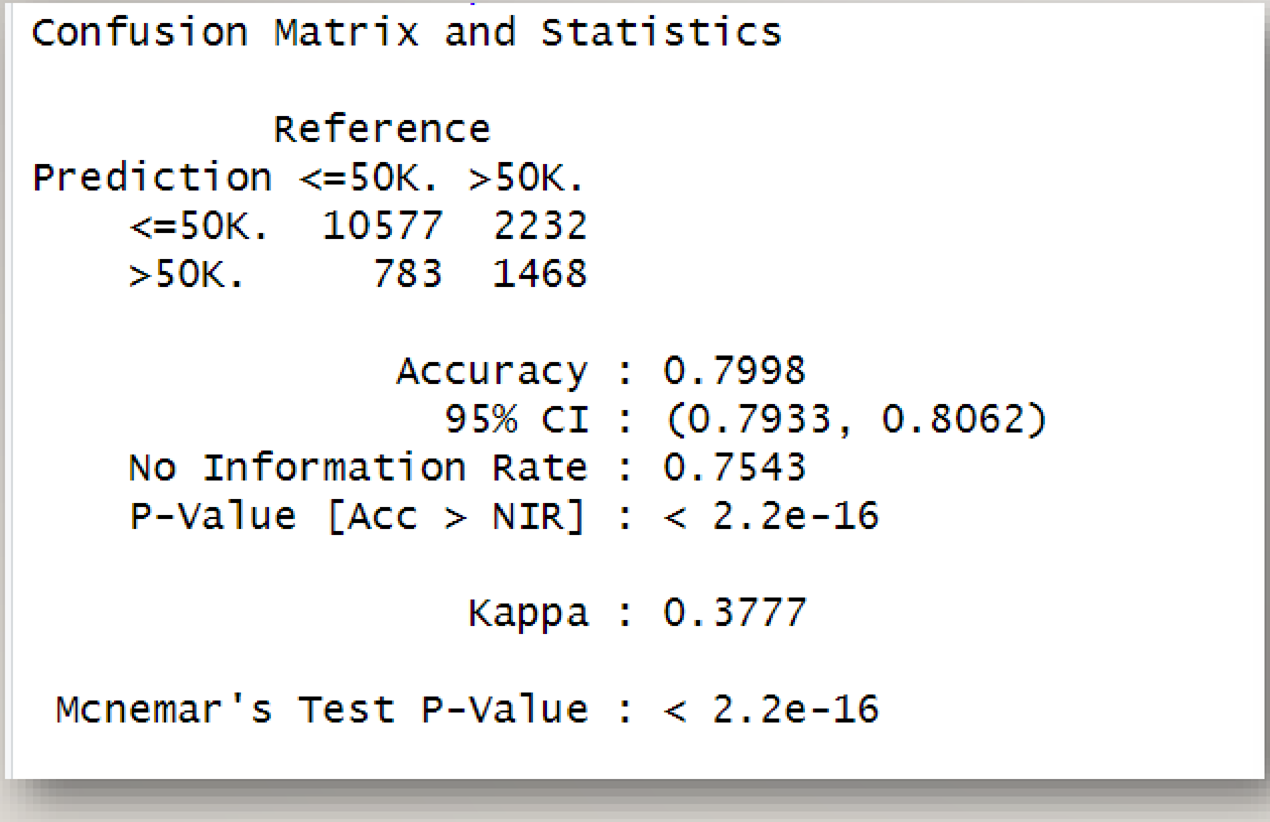
❑ No of Leaves: 5

❑ Split based on:

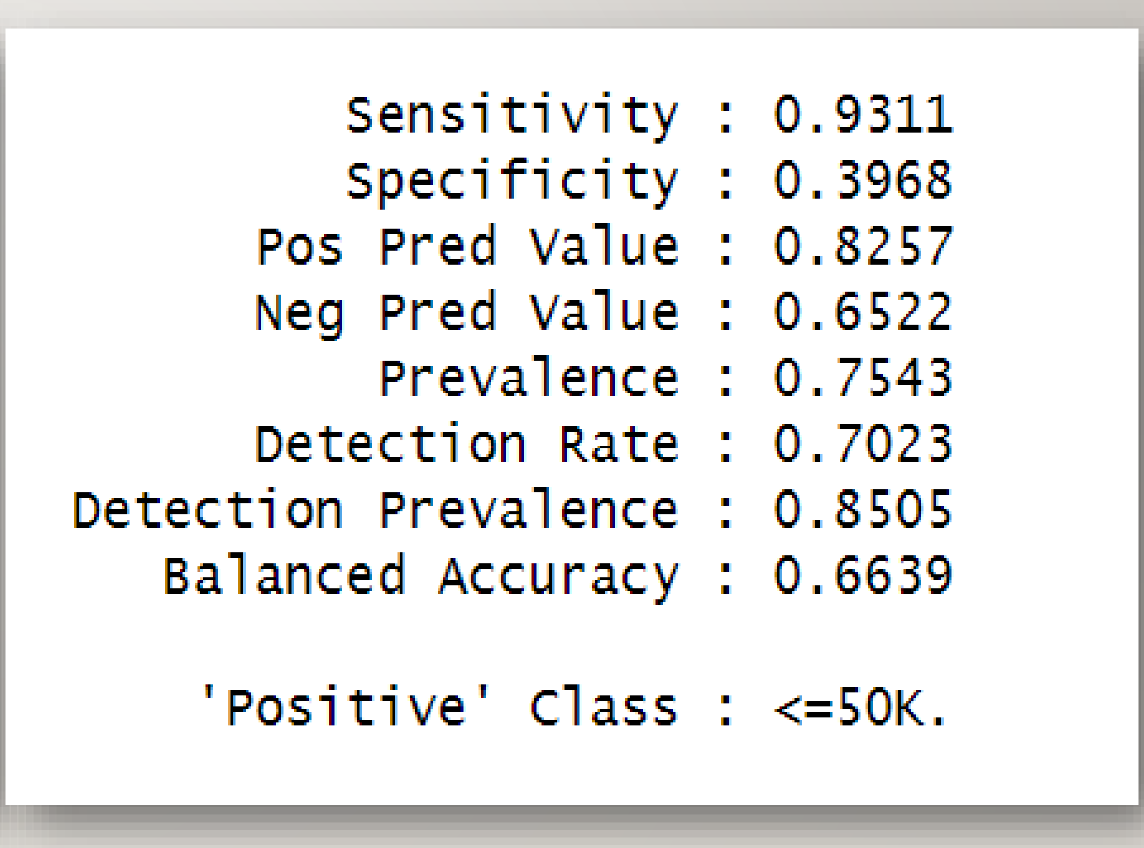
- Age
- Workclass
- Education_level
- Occupation

Decision Tree Model

Confusion Matrix for Training Data - (Summary)



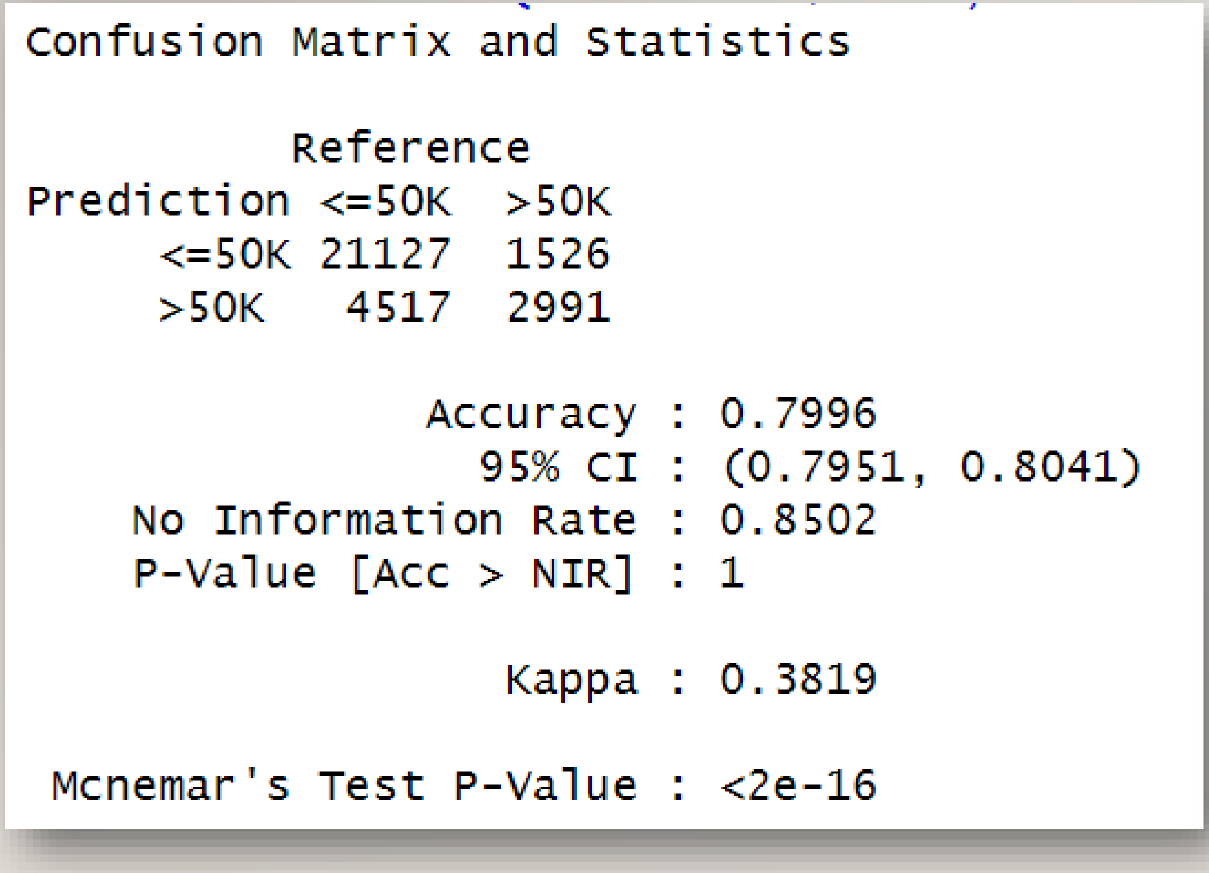
```
Confusion Matrix and Statistics

          Reference
Prediction <=50K. >50K.
    <=50K.  10577  2232
    >50K.     783  1468

               Accuracy : 0.7998
                 95% CI : (0.7933, 0.8062)
    No Information Rate : 0.7543
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3777

 Mcnemar's Test P-Value : < 2.2e-16
```



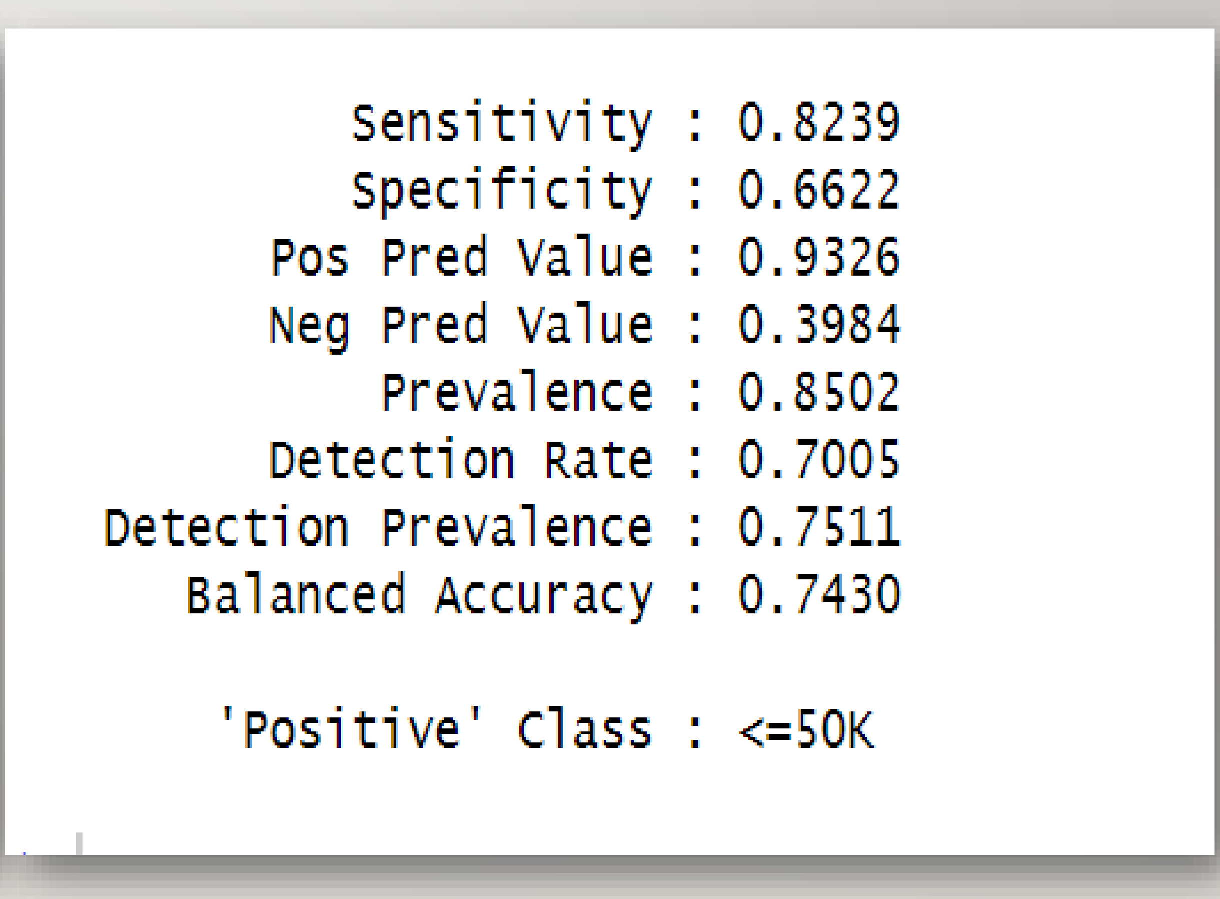
```
            Sensitivity : 0.9311
            Specificity : 0.3968
         Pos Pred Value : 0.8257
         Neg Pred Value : 0.6522
             Prevalence : 0.7543
         Detection Rate : 0.7023
   Detection Prevalence : 0.8505
      Balanced Accuracy : 0.6639

       'Positive' Class : <=50K.
```

Decision Tree Model

Confusion Matrix for Validation Data - (Summary)

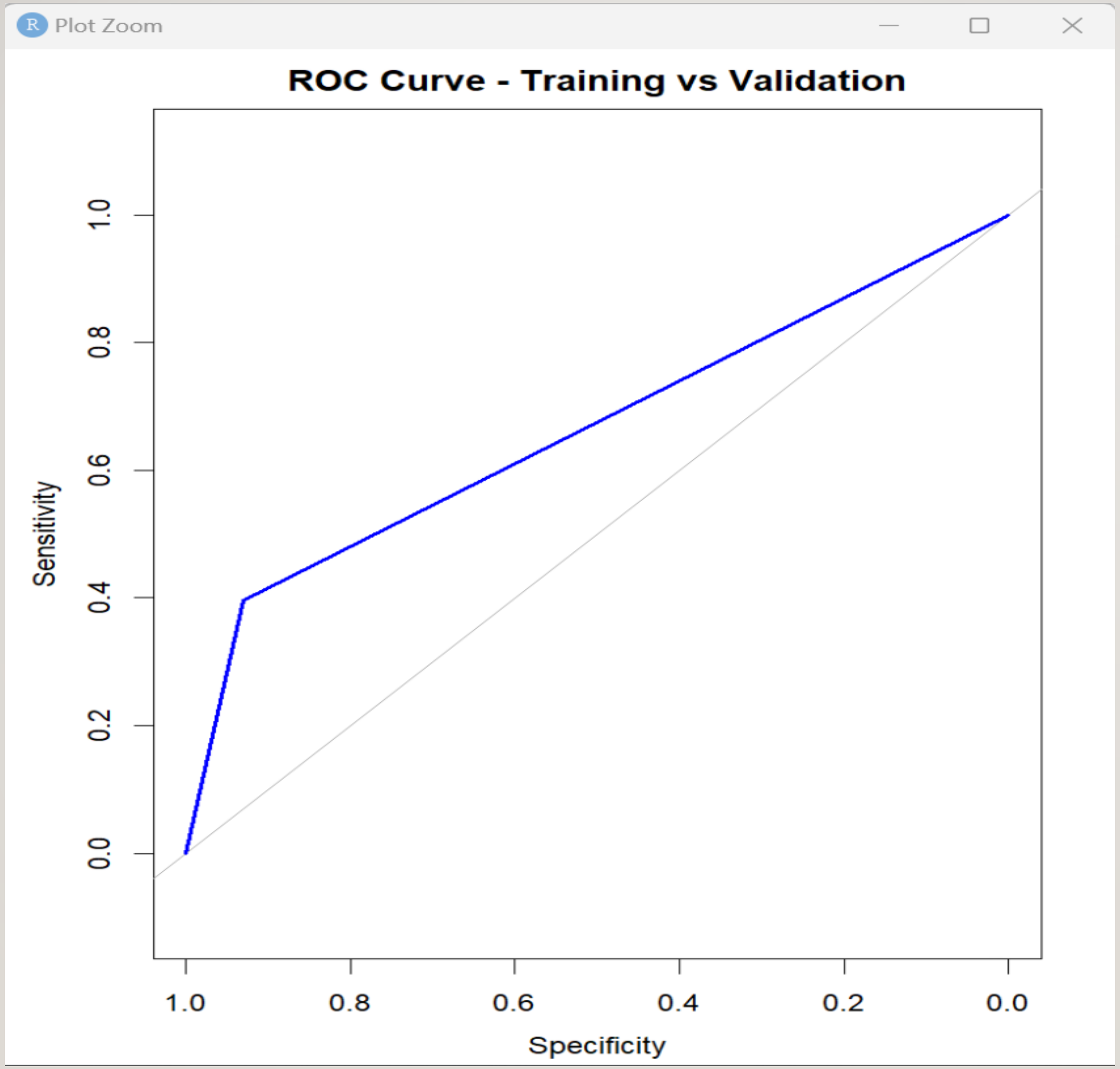


```
Confusion Matrix and Statistics

          Reference
Prediction <=50K  >50K
     <=50K 21127  1526
     >50K   4517  2991

               Accuracy : 0.7996
                 95% CI : (0.7951, 0.8041)
    No Information Rate : 0.8502
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3819

 Mcnemar's Test P-Value : <2e-16
```



```
            Sensitivity : 0.8239
            Specificity : 0.6622
         Pos Pred Value : 0.9326
         Neg Pred Value : 0.3984
             Prevalence : 0.8502
         Detection Rate : 0.7005
   Detection Prevalence : 0.7511
      Balanced Accuracy : 0.7430

       'Positive' Class : <=50K
```

Actual and Predicted Records for Decision Tree - ROC

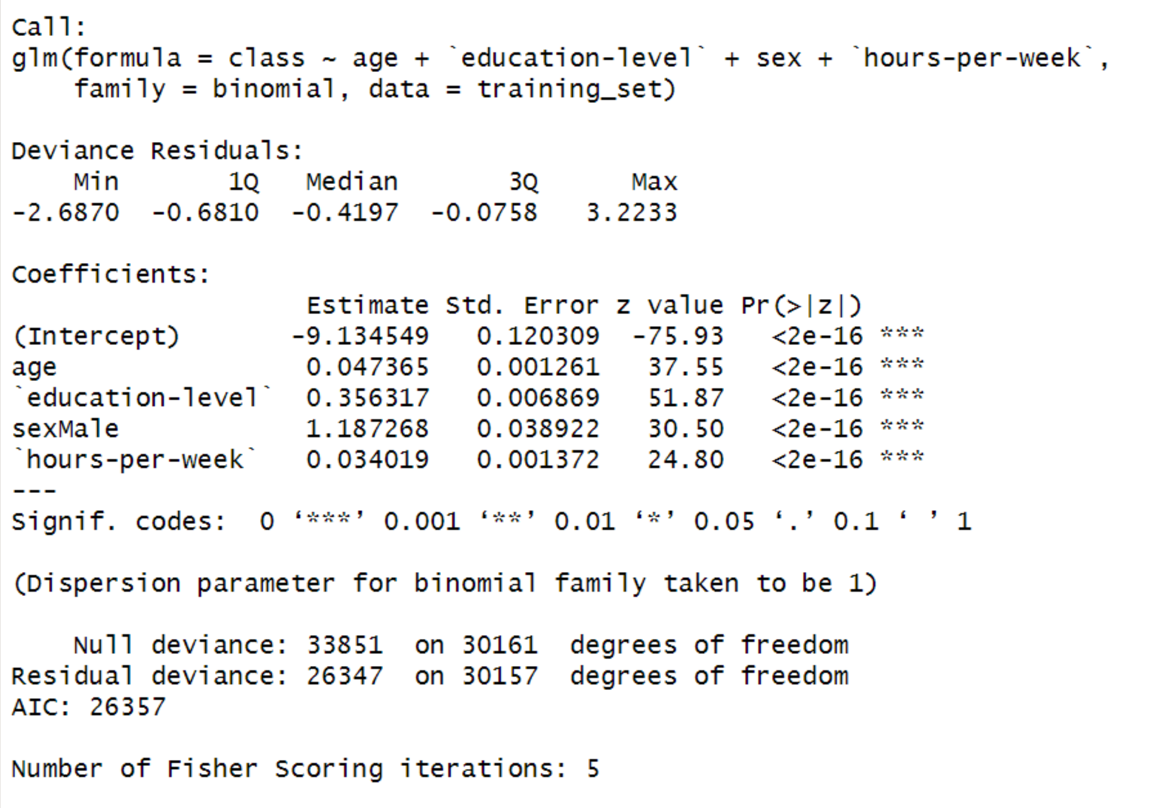


- Area under the curve: 0.7639

LOGISTIC REGRESSION MODEL

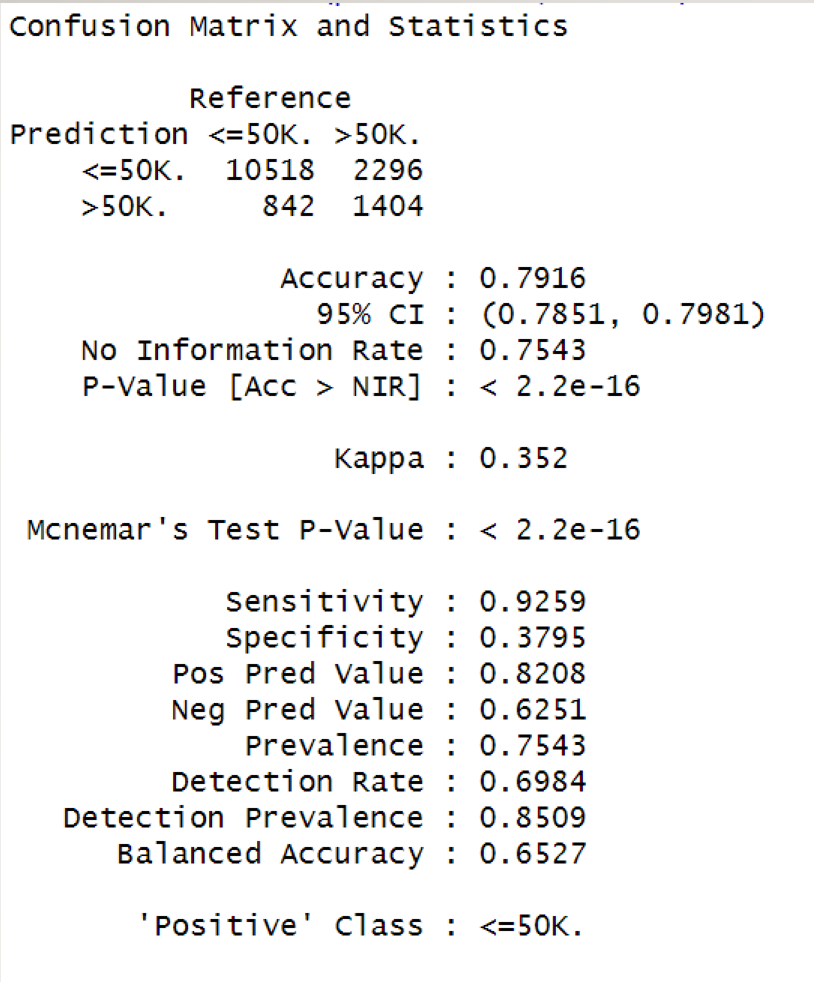
- Logistic regression is used to predict the class (or category) of individuals based on one or multiple predictor variables (x). It is used to model a binary outcome, that is a variable, which can have only two possible values: 0 or 1, yes or no, diseased or non-diseased

LOGISTIC REGRESSION MODEL SUMMARY



```
Call:
glm(formula = class ~ age + `education-level` + sex + `hours-per-week`, 
    family = binomial, data = training_set)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.6870  -0.6810  -0.4197  -0.0758   3.2233  

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -9.134549   0.120309  -75.93   <2e-16 ***
age                0.047365   0.001261   37.55   <2e-16 ***
`education-level`  0.356317   0.006869   51.87   <2e-16 ***
sexMale            1.187268   0.038922   30.50   <2e-16 ***
`hours-per-week`   0.034019   0.001372   24.80   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33851  on 30161  degrees of freedom
Residual deviance: 26347  on 30157  degrees of freedom
AIC: 26357

Number of Fisher Scoring iterations: 5
```

LOGISTIC REGRESSION MODEL CONFUSION MATRIX



```
Confusion Matrix and Statistics

          Reference
Prediction <=50K. >50K.
    <=50K.  10518  2296
    >50K.     842  1404

               Accuracy : 0.7916
                 95% CI : (0.7851, 0.7981)
    No Information Rate : 0.7543
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.352

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9259
            Specificity : 0.3795
         Pos Pred Value : 0.8208
         Neg Pred Value : 0.6251
             Prevalence : 0.7543
         Detection Rate : 0.6984
   Detection Prevalence : 0.8509
      Balanced Accuracy : 0.6527

       'Positive' Class : <=50K.
```

COMPARISION AND CONCLUSION

- The accuracy values you provided for the three models are as follows:

  Decision Tree: 0.7996

  Logistic Regression: 0.7916

  Neural Network: 0.8108

- In general, accuracy alone may not provide a complete picture, and it's advisable to consider other metrics depending on the specific characteristics of your classification problem. However, if we only consider accuracy, the neural network has the highest accuracy (81.08%), making it the best performer among the models mentioned.
- Therefore, decision tree is the optimal model out of the three considering, ROC curve, precision, recall etc.