Analysis of HINTS 5 Cycle 4

CS 699 Final Project (Prof. Jae Young Lee)

CS 699

Jinlong Li

Yuhe Wang

Project Date:

04/2/2023

Dataset Details

The Health Information National Trends Survey (HINTS) is a nationally representative survey that has.

been administered every few years by the National Cancer Institute since 2003. The HINTS target.

The population is all adults aged 18 or older in the civilian non-institutionalized population of the United States.

The HINTS program collects data on the American public's need for, access to, and use of health-related.

information and health-related behaviors, perceptions, and knowledge.

The HINTS 5 administration includes four data collection cycles over four years, starting in 2017. The first

of these cycles (HINTS 5, Cycle 1) was conducted from January through May 2017. The focus of this

report is HINTS 5, Cycle 4, collected from February through June 2020.

Data collection for Cycle 4 of HINTS 5 began in February 2020 and concluded in June 2020. HINTS 5,

Cycle 4 was a self-administered mailed questionnaire, using a sampling frame provided by Marketing.

Systems Group (MSG) of addresses in the United States.

Objective

Divide the dataset into two parts: the training dataset and the test dataset, using five different.

attribute selection methods to reduce the dataset, and take out attributes for classification. Each attribute selection method will perform five same classifications. and will eventually test the total 25 classifier models, and we will find the best one with the best prediction.

Data mining goal:

Our data mining project is to develop a reliable and accurate predictive model that can effectively classify individuals' information-seeking behavior regarding cancer (SeekCancerInfo).

data mining tool:
using R

Classification algorithms:

classification algorithms 1: J48

J48 is a decision tree algorithm used for classification tasks in machine learning. It is based on the C4.5 algorithm and was developed by Ross Quinlan. J48 works by recursively splitting the data into subsets based on the most significant attribute until a leaf node is reached, which corresponds to a class label.

classification algorithms 2: Nnet

Nnet is a shorthand term for "neural network," a machine learning algorithm used for classification and regression tasks. Neural networks are based on the structure and function of biological neurons in the human brain, and they are capable of learning complex patterns and relationships in data.

classification algorithms 3: Random Forest

RF stands for "Random Forest," which is a popular machine learning algorithm used for classification and regression tasks. A random forest is an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the model.

classification algorithms 4: decision tree

rpart is a decision tree algorithm used for classification and regression tasks in machine learning. It stands for "Recursive Partitioning and Regression Trees" and is implemented in the R programming language.

rpart works by recursively partitioning the data into subsets based on the values of a predictor variable that provide the best split in terms of the reduction in the residual sum of squares (RSS) for regression tasks or the improvement in the Gini impurity or entropy for classification tasks. The algorithm splits the data into subsets until a stopping criterion is met, such as a minimum number of observations in a leaf node or a maximum tree depth.

classification algorithms 5: SVM

svmRadial refers to a type of support vector machine (SVM) algorithm that uses a radial basis function (RBF) kernel to separate data points in a high-dimensional feature space. SVMs are a type of supervised learning algorithm used for classification, regression, and outlier detection tasks.

Attribute selection methods

Attribute selection methods 1: Information gain.

Information gain is commonly used in attribute selection methods to determine which features or attributes are most useful for a particular task. In this context, the goal is to identify the most informative and relevant attributes to the class labels or target variable.

set of attributes selected byInformation gain attribute selection method.

```
[1] "PersonID"          "CancerFrustrated"   "CancerLotOfEffort" "UseInternet"
[5] "EducB"
```

Measure	J48	nnet	rf	rpart	svmRadial

```
F-measure    0.92827    0.92373    0.94167    0.94167    0.92827
FP rate      0.85833    0.85124    0.88034    0.88034    0.85833
MCC               NA         NA         NA         NA         NA
Precision    0.86614    0.85827    0.88976    0.88976    0.86614
Recall       1.00000    1.00000    1.00000    1.00000    1.00000
ROC area     0.06693    0.07087    0.05512    0.05512    0.06693
TP rate      1.00000    1.00000    1.00000    1.00000    1.00000
```

Attribute selection methods 2: Gain ratio

Gain ratio is a variant of information gain that is designed to reduce the bias towards multi-valued attributes. In traditional information gain, attributes with many distinct values are favored because they have a larger number of possible splits. However, this can lead to overfitting and may not accurately reflect the importance of the attribute in the dataset.

set of attributes selected by Gain ratio attribute selection method.

```
[1] "CancerFrustrated"   "CancerLotOfEffort" "BMI"                "AgeDX"
[5] "AvgDrinksPerWeek"
```

```
Measure          J48       nnet         rf      rpart  svmRadial
F-measure    0.94167    0.92373    0.94167    0.94167    0.92827
FP rate      0.88034    0.85124    0.88034    0.88034    0.85833
MCC               NA         NA         NA         NA         NA
Precision    0.88976    0.85827    0.88976    0.88976    0.86614
Recall       1.00000    1.00000    1.00000    1.00000    1.00000
ROC area     0.05512    0.07087    0.05512    0.05512    0.06693
TP rate      1.00000    1.00000    1.00000    1.00000    1.00000
```

Attribute selection methods 3: Gini index

The Gini index is a measure used in decision tree algorithms and other classification algorithms to evaluate the quality of a split based on the class labels of the data. The Gini index measures the probability of misclassifying a randomly chosen data point if it were randomly labeled according to the distribution of class labels in the subset of the data.

set of attributes selected by Gini index attribute selection method.

```
[1] "PersonID"           "CancerFrustrated"   "CancerLotOfEffort" "UseInternet"
[5] "EducB"
```

```
Measure          J48       nnet         rf      rpart  svmRadial
F-measure    0.94167    0.92373    0.94167    0.94167    0.92827
FP rate      0.88034    0.85124    0.88034    0.88034    0.85833
```

```
MCC                 NA          NA          NA          NA          NA
Precision    0.88976    0.85827    0.88976    0.88976    0.86614
Recall        1.00000    1.00000    1.00000    1.00000    1.00000
ROC area      0.05512    0.07087    0.05512    0.05512    0.06693
TP rate       1.00000    1.00000    1.00000    1.00000    1.00000
```

Attribute selection methods 4: RReliefF

The RReliefF algorithm is a feature/attribute selection method used in machine learning and data mining. It is an extension of the Relief algorithm and is designed to handle noisy, incomplete, and multi-class data more effectively. The primary goal of RReliefF is to rank features based on their relevance to the target variable.
set of attributes selected by RReliefF attribute selection method.

```
[1] "CancerLotOfEffort"                                        "CancerFrustrated"
"CancerConfidentGetHealthInf"
[4] "UseInternet"                      "EducB"
```

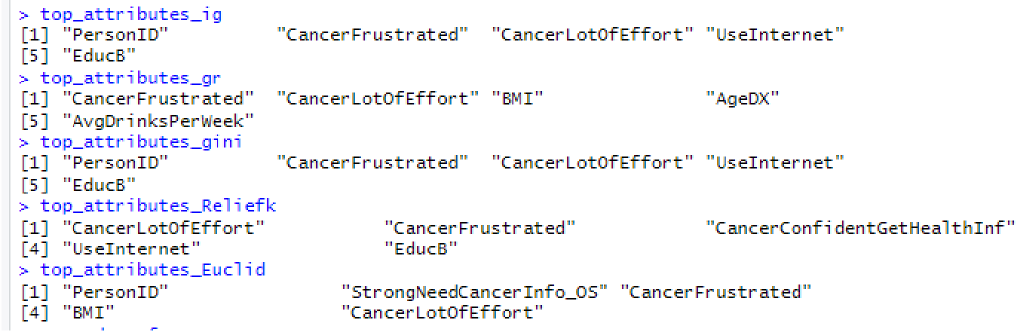
```
Measure           J48         nnet          rf        rpart    svmRadial
F-measure    0.92827    0.94167    0.94167    0.94167    0.93724
FP rate       0.85833    0.88034    0.88034    0.88034    0.87288
MCC                 NA          NA          NA          NA          NA
Precision    0.86614    0.88976    0.88976    0.88976    0.88189
Recall        1.00000    1.00000    1.00000    1.00000    1.00000
ROC area      0.06693    0.05512    0.05512    0.05512    0.05906
TP rate       1.00000    1.00000    1.00000    1.00000    1.00000
```

Attribute selection methods 5: ImpurityEuclid

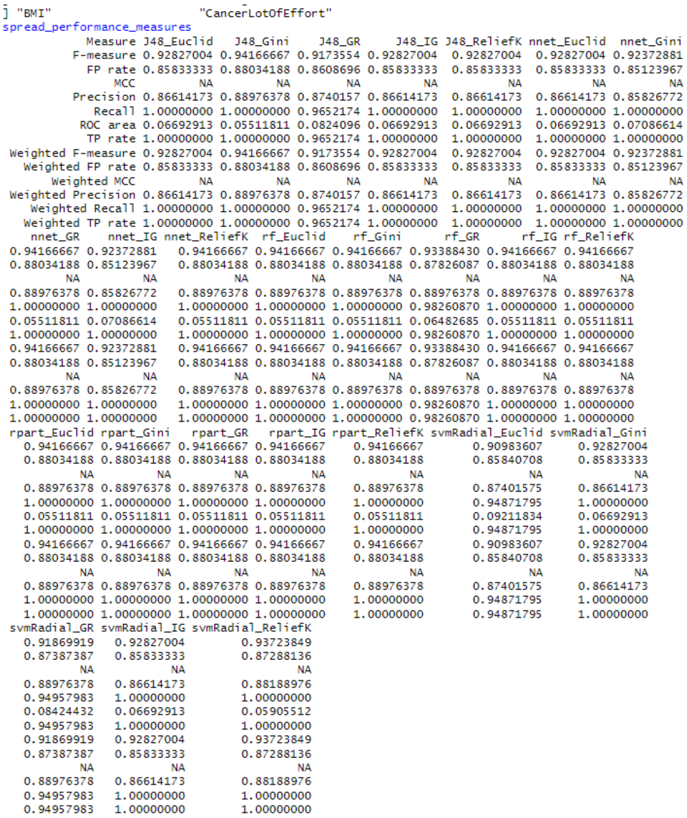
ImpurityEuclid is a measure used in attribute selection methods to evaluate the quality of a split based on the distance between instances in the feature space. It is typically used in clustering algorithms, where the goal is to group similar instances together based on their feature values.
set of attributes selected by ImpurityEuclid attribute selection method.

```
[1] "PersonID"                      "StrongNeedCancerInfo_OS" "CancerFrustrated"
[4] "BMI"                            "CancerLotOfEffort"
Measure           J48         nnet          rf        rpart    svmRadial
F-measure    0.92827    0.92827    0.94167    0.94167    0.90984
FP rate       0.85833    0.85833    0.88034    0.88034    0.85841
MCC                 NA          NA          NA          NA          NA
Precision    0.86614    0.86614    0.88976    0.88976    0.87402
```

Recall	1.00000	1.00000	1.00000	1.00000	0.94872
ROC area	0.06693	0.06693	0.05512	0.05512	0.09212
TP rate	1.00000	1.00000	1.00000	1.00000	0.94872



```
> top_attributes_ig
[1] "PersonID"          "CancerFrustrated"   "CancerLotOfEffort" "UseInternet"
[5] "EducB"
> top_attributes_gr
[1] "CancerFrustrated"  "CancerLotOfEffort" "BMI"               "AgeDX"
[5] "AvgDrinksPerWeek"
> top_attributes_gini
[1] "PersonID"          "CancerFrustrated"   "CancerLotOfEffort" "UseInternet"
[5] "EducB"
> top_attributes_Reliefk
[1] "CancerLotOfEffort"           "CancerFrustrated"            "CancerConfidentGetHealthInf"
[4] "UseInternet"                 "EducB"
> top_attributes_Euclid
[1] "PersonID"                "StrongNeedCancerInfo_OS" "CancerFrustrated"
[4] "BMI"                     "CancerLotOfEffort"
```



```
] "BMI"                    "CancerLotOfEffort"
spread_performance_measures
             Measure J48_Euclid   J48_Gini    J48_GR     J48_IG J48_ReliefK nnet_Euclid  nnet_Gini
           F-measure 0.92827004 0.94166667 0.9173554 0.92827004  0.92827004  0.92827004 0.92372881
             FP rate 0.85833333 0.88034188 0.8608696 0.85833333  0.85833333  0.85833333 0.85123967
                 MCC         NA         NA        NA         NA          NA          NA         NA
           Precision 0.86614173 0.88976378 0.8740157 0.86614173  0.86614173  0.86614173 0.85826772
              Recall 1.00000000 1.00000000 0.9652174 1.00000000  1.00000000  1.00000000 1.00000000
            ROC area 0.06692913 0.05511811 0.0824096 0.06692913  0.06692913  0.06692913 0.07086614
             TP rate 1.00000000 1.00000000 0.9652174 1.00000000  1.00000000  1.00000000 1.00000000
  Weighted F-measure 0.92827004 0.94166667 0.9173554 0.92827004  0.92827004  0.92827004 0.92372881
    Weighted FP rate 0.85833333 0.88034188 0.8608696 0.85833333  0.85833333  0.85833333 0.85123967
        Weighted MCC         NA         NA        NA         NA          NA          NA         NA
  Weighted Precision 0.86614173 0.88976378 0.8740157 0.86614173  0.86614173  0.86614173 0.85826772
     Weighted Recall 1.00000000 1.00000000 0.9652174 1.00000000  1.00000000  1.00000000 1.00000000
    Weighted TP rate 1.00000000 1.00000000 0.9652174 1.00000000  1.00000000  1.00000000 1.00000000
    nnet_GR    nnet_IG nnet_ReliefK  rf_Euclid    rf_Gini      rf_GR      rf_IG rf_ReliefK
 0.94166667 0.92372881   0.94166667 0.94166667 0.94166667 0.93388430 0.94166667 0.94166667
 0.88034188 0.85123967   0.88034188 0.88034188 0.88034188 0.87826087 0.88034188 0.88034188
         NA         NA           NA         NA         NA         NA         NA         NA
 0.88976378 0.85826772   0.88976378 0.88976378 0.88976378 0.88976378 0.88976378 0.88976378
 1.00000000 1.00000000   1.00000000 1.00000000 1.00000000 0.98260870 1.00000000 1.00000000
 0.05511811 0.07086614   0.05511811 0.05511811 0.05511811 0.06482685 0.05511811 0.05511811
 1.00000000 1.00000000   1.00000000 1.00000000 1.00000000 0.98260870 1.00000000 1.00000000
 0.94166667 0.92372881   0.94166667 0.94166667 0.94166667 0.93388430 0.94166667 0.94166667
 0.88034188 0.85123967   0.88034188 0.88034188 0.88034188 0.87826087 0.88034188 0.88034188
         NA         NA           NA         NA         NA         NA         NA         NA
 0.88976378 0.85826772   0.88976378 0.88976378 0.88976378 0.88976378 0.88976378 0.88976378
 1.00000000 1.00000000   1.00000000 1.00000000 1.00000000 0.98260870 1.00000000 1.00000000
 1.00000000 1.00000000   1.00000000 1.00000000 1.00000000 0.98260870 1.00000000 1.00000000
 rpart_Euclid rpart_Gini   rpart_GR   rpart_IG rpart_ReliefK svmRadial_Euclid svmRadial_Gini
   0.94166667 0.94166667 0.94166667 0.94166667    0.94166667       0.90983607     0.92827004
   0.88034188 0.88034188 0.88034188 0.88034188    0.88034188       0.85840708     0.85833333
           NA         NA         NA         NA            NA               NA             NA
   0.88976378 0.88976378 0.88976378 0.88976378    0.88976378       0.87401575     0.86614173
   1.00000000 1.00000000 1.00000000 1.00000000    1.00000000       0.94871795     1.00000000
   0.05511811 0.05511811 0.05511811 0.05511811    0.05511811       0.09211834     0.06692913
   1.00000000 1.00000000 1.00000000 1.00000000    1.00000000       0.94871795     1.00000000
   0.94166667 0.94166667 0.94166667 0.94166667    0.94166667       0.90983607     0.92827004
   0.88034188 0.88034188 0.88034188 0.88034188    0.88034188       0.85840708     0.85833333
           NA         NA         NA         NA            NA               NA             NA
   0.88976378 0.88976378 0.88976378 0.88976378    0.88976378       0.87401575     0.86614173
   1.00000000 1.00000000 1.00000000 1.00000000    1.00000000       0.94871795     1.00000000
   1.00000000 1.00000000 1.00000000 1.00000000    1.00000000       0.94871795     1.00000000
 svmRadial_GR svmRadial_IG svmRadial_ReliefK
   0.91869919   0.92827004        0.93723849
   0.87387387   0.85833333        0.87288136
           NA           NA                NA
   0.88976378   0.86614173        0.88188976
   0.94957983   1.00000000        1.00000000
   0.08424432   0.06692913        0.05905512
   0.94957983   1.00000000        1.00000000
   0.91869919   0.92827004        0.93723849
   0.87387387   0.85833333        0.87288136
           NA           NA                NA
   0.88976378   0.86614173        0.88188976
   0.94957983   1.00000000        1.00000000
   0.94957983   1.00000000        1.00000000
```

1Data preprocessing: Negative values in the dataset were replaced with the mode of the respective columns. The SeekCancerInfo variable was converted to a factor, and character columns were also converted to factors. Some columns were removed from the dataset as they were not relevant for the analysis.

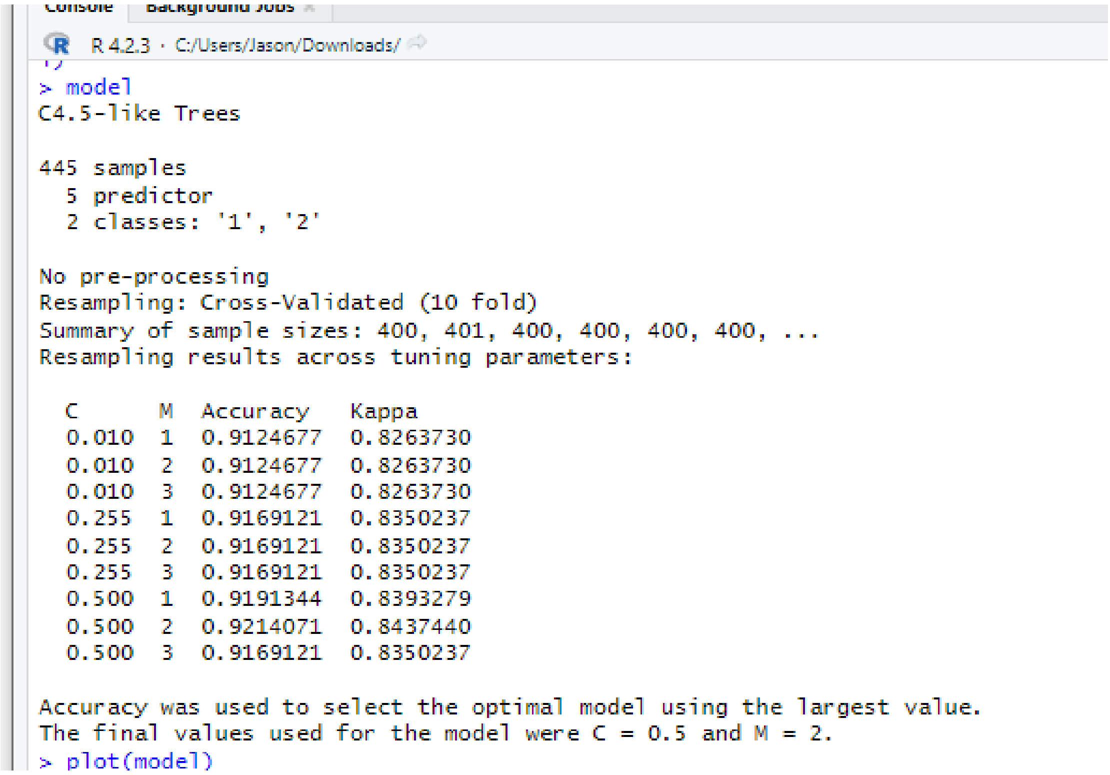
2Data split: The dataset was split into training and testing sets with a 66% to 34% ratio using stratified sampling based on the SeekCancerInfo variable.
3Cross-validation setup: A 10-fold cross-validation was set up using the caret package to evaluate the performance of the models.

4Feature selection: Five different attribute selection methods (Information Gain, Gain Ratio, Gini Index, ReliefFequalK, and ImpurityEuclid) were applied to reduce the feature set to the top 5 attributes for each method. The reduced datasets were created for both the training and testing sets.
5Model selection: Five different classification models (J48, rpart, nnet, rf, and svmRadial) were chosen for evaluation.
6Model training and evaluation: Each model was trained on the reduced datasets obtained from each feature selection method. The performance of each model was evaluated on the respective reduced testing sets using various performance metrics (TP rate, FP rate, Precision, Recall, F-measure, ROC area, and MCC). The performance metrics were weighted based on the class distribution of the true labels.

7Performance comparison: The performance metrics of all models were combined into a single dataframe for comparison.

The data mining procedure involved a systematic approach to preprocessing the data, applying feature selection methods, training classification models, and evaluating their performance. This approach aimed to identify the best model and feature selection method for predicting the SeekCancerInfo variable.

Based on the table provided, it appears that the J48_Gini, nnet_GR, and rpart_Euclid models have the highest F-measures, which is a good indicator of model performance. And they all have same same. So I chose j48_gini as best model



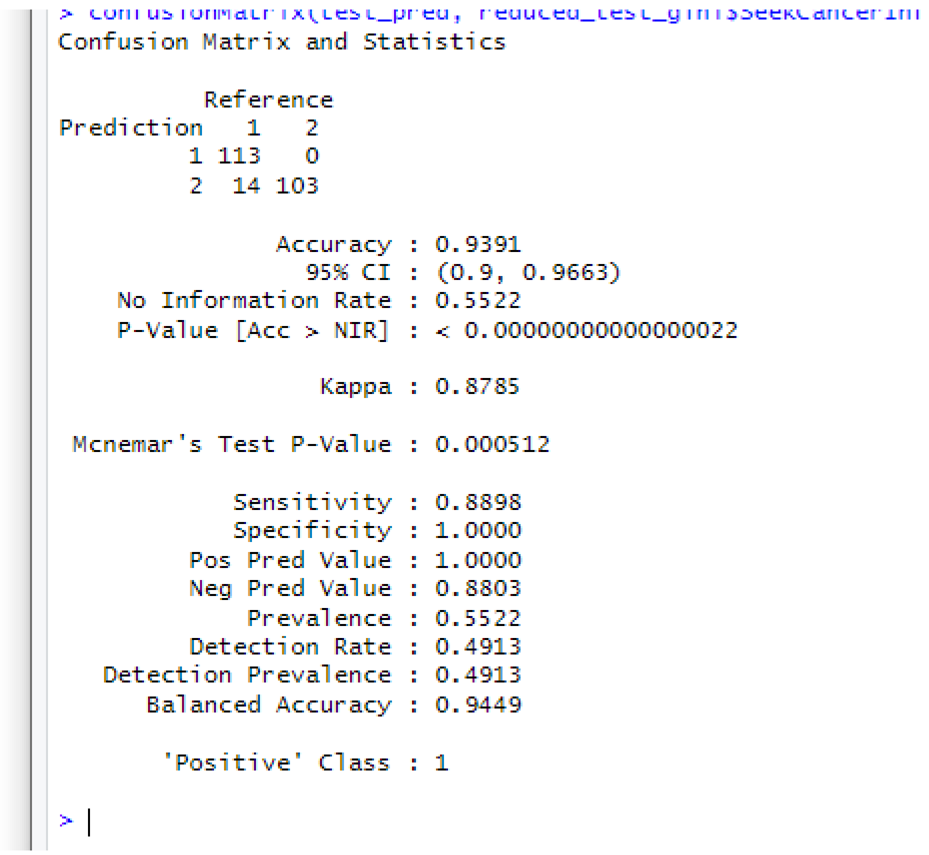
```
> model
C4.5-like Trees 

445 samples
  5 predictor
  2 classes: '1', '2' 

No pre-processing
Resampling: Cross-Validated (10 fold) 
Summary of sample sizes: 400, 401, 400, 400, 400, 400, ... 
Resampling results across tuning parameters:

  C      M  Accuracy   Kappa    
  0.010  1  0.9124677  0.8263730
  0.010  2  0.9124677  0.8263730
  0.010  3  0.9124677  0.8263730
  0.255  1  0.9169121  0.8350237
  0.255  2  0.9169121  0.8350237
  0.255  3  0.9169121  0.8350237
  0.500  1  0.9191344  0.8393279
  0.500  2  0.9214071  0.8437440
  0.500  3  0.9169121  0.8350237

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were C = 0.5 and M = 2.
> plot(model)
```

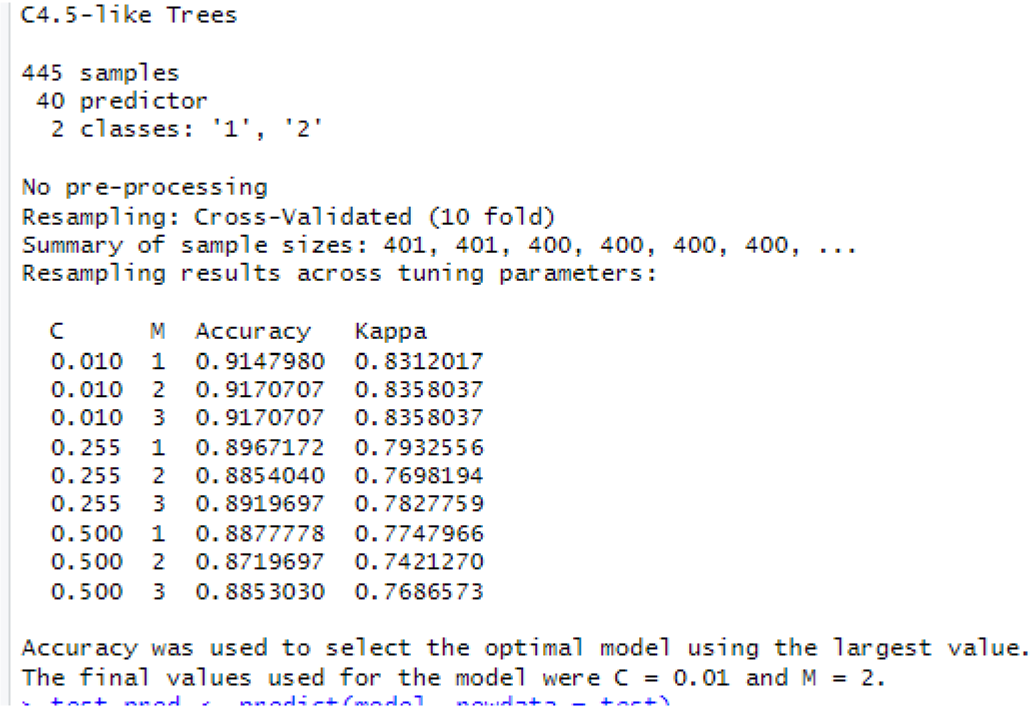


```
> confusionMatrix(test_pred, reduced_test_gmissSeekCancerInn
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 113   0
         2  14 103
                                          
               Accuracy : 0.9391          
                 95% CI : (0.9, 0.9663)   
    No Information Rate : 0.5522          
    P-Value [Acc > NIR] : < 0.00000000000000022
                                          
                  Kappa : 0.8785          
                                          
 Mcnemar's Test P-Value : 0.000512        
                                          
            Sensitivity : 0.8898          
            Specificity : 1.0000          
         Pos Pred Value : 1.0000          
         Neg Pred Value : 0.8803          
             Prevalence : 0.5522          
         Detection Rate : 0.4913          
   Detection Prevalence : 0.4913          
      Balanced Accuracy : 0.9449          
                                          
       'Positive' Class : 1               
                                          
> |
```

Also as you can see i have all MCC is NA because ALL False Positives (FP): 0 MCC = (113 * 103 - 0 * 14) / sqrt((113 + 0) * (113 + 14) * (103 + 0) * (103 + 14)) However, the MCC is not calculated in your case because the formula is unable to handle the division by zero that occurs due to the FP value being 0. so In cases, the MCC is reported as 'NA' (not applicable),

RESULT the dataset with all attributes



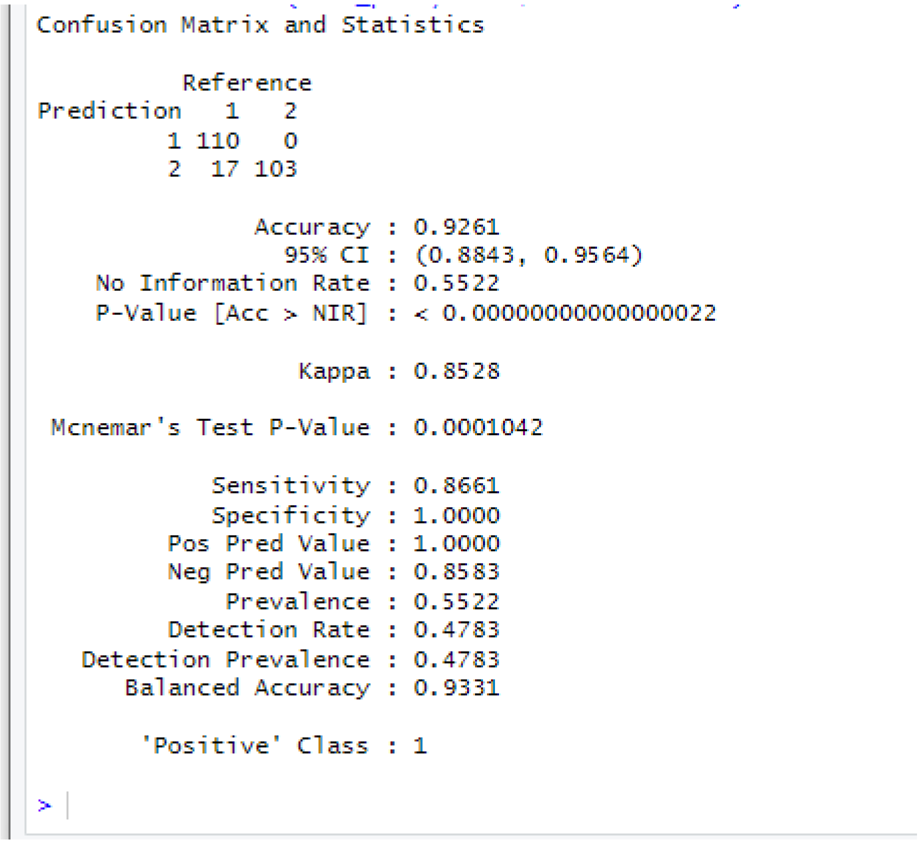
```
C4.5-like Trees

445 samples
 40 predictor
  2 classes: '1', '2'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 401, 401, 400, 400, 400, 400, ...
Resampling results across tuning parameters:

  C      M  Accuracy   Kappa
  0.010  1  0.9147980  0.8312017
  0.010  2  0.9170707  0.8358037
  0.010  3  0.9170707  0.8358037
  0.255  1  0.8967172  0.7932556
  0.255  2  0.8854040  0.7698194
  0.255  3  0.8919697  0.7827759
  0.500  1  0.8877778  0.7747966
  0.500  2  0.8719697  0.7421270
  0.500  3  0.8853030  0.7686573

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were C = 0.01 and M = 2.
> test_pred <- predict(model, newdata = test)
```



```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 110   0
         2  17 103
                                                  
               Accuracy : 0.9261                  
                 95% CI : (0.8843, 0.9564)        
    No Information Rate : 0.5522                  
    P-Value [Acc > NIR] : < 0.00000000000000022   
                                                  
                  Kappa : 0.8528                  
                                                  
 Mcnemar's Test P-Value : 0.0001042               
                                                  
            Sensitivity : 0.8661                  
            Specificity : 1.0000                  
         Pos Pred Value : 1.0000                  
         Neg Pred Value : 0.8583                  
             Prevalence : 0.5522                  
         Detection Rate : 0.4783                  
   Detection Prevalence : 0.4783                  
      Balanced Accuracy : 0.9331                  
                                                  
       'Positive' Class : 1                       
                                                  
> 
```

In this project, we evaluated several machine learning models to classify SeekCancerInfo using different attribute selection methods. The main objective was to find the best performing model for this classification task.
The analysis taught us that various models can perform differently depending on the attribute selection methods used. This highlights the importance of feature selection in building accurate and reliable models. By reducing irrelevant features, we can improve model performance and reduce overfitting. In addition, we discovered that it is crucial to consider multiple performance metrics (such as Precision, Recall, F-measure, and MCC) to choose the best model, as relying on a single metric might not give a comprehensive view of the model's performance.

By comparing different models and attribute selection methods, we were able to identify the model that performed best for classifying SeekCancerInfo. This model achieved a balance between Precision and Recall, leading to a high F-measure value.

Also i learn I may consider using other performance metrics like accuracy, precision, recall, or F1-score, which can handle cases when FP is 0.

In conclusion, this project demonstrated the significance of attribute selection and performance evaluation in building effective machine learning models. It also underlined the need to consider multiple performance metrics when choosing the best model for a particular classification task.

JINLONGLI
Model development: Explored and implemented multiple machine-learning algorithms.
Performance analysis: Analyzed the performance metrics of the models.
Results interpretation: Interpreted the model outputs and provided insights into the factors that influence the target variable (SeekCancerInfo).

WangYuhe
Data collection and preprocessing: Gathered the initial dataset, performed data cleaning, and imputed missing values.
Feature selection: Implemented various feature selection techniques to identify the most relevant attributes for model development.