ML1 Data Mining Project Final Report: Mushroom Edibility Classification Model

Team Members: Anieesh Saravanan, Pranav Akiri

10/23/2024

Table of Contents

Part 1 – Statement/Project Goal

The Audubon Society Field Guide provides a classification of mushrooms based on a variety of characteristics that help determine whether a mushroom is edible or poisonous. These traits span across 22 features, which describe physical properties of the mushroom, such as:

1. Cap Shape
2. Cap Surface
3. Cap Color
4. Bruises
5. Odor
6. Gill Size
7. Stalk Surface

   …

(Refer to Part 2 for a full list of features)

Based on these features, mushrooms are classified as either edible or poisonous. This dataset combines mushrooms labeled as "unknown edibility" with the poisonous category to ensure the classification process exerts more caution when classifying mushrooms as edible. The most common classification overall is edible.

This classification is important for ensuring public safety, as some mushrooms that appear edible may actually be poisonous. Each mushroom is labeled based on its edibility, but the process of determining this label based on the various physical features is not always straightforward and well defined. Unlike other plants, there is no simple rule to classify mushrooms as edible or poisonous, such as the "leaflets three, let it be" rule for identifying Poison Ivy or Poison Oak. For this project, we will explore the features provided in the dataset and use them to train and test a binary classification model that predicts the edibility of a mushroom. After determining which classification models perform best, we can identify characteristics which contribute most to the model's predictions and determine which attributes are the most statistically significant in identifying mushroom edibility.

Part 2 – Description of Dataset

The dataset we used includes information regarding various features of mushrooms and was sourced from the *The Audubon Society Field Guide to North American Mushrooms*, authored by mycologist Gary Lincoff. The dimension of the dataset before any preprocessing occurred was 22 and it includes 8,124 unique instances. The attributes represent certain characteristics of the mushrooms that can assist in determining whether or not a particular mushroom is safe to consume. Each data value consists of a single letter instead of a phrase in order to ensure that the data appears clean and neat. The placeholders for each attribute can be interpreted as follows:

1. **class: edible = e, poisonous = p**
2. cap-shape: bell = b, conical = c, convex = x, flat = f, knobbed = k, sunken = s
3. cap-surface: fibrous = f, grooves = g, scaly = y, smooth = s
4. cap-color: brown = n, buff = b, cinnamon = c, gray = g, green = r, pink = p, purple = u, red = e, white = w, yellow = y
5. bruises: true = t, false = f
6. odor: almond = a, anise = l, creosote = c, fishy = y, foul = f, musty = m, none = n, pungent = p, spicy = s
7. gill-attachment: attached = a, descending = d, free = f, notched = n
8. gill-spacing: close = c, crowded = w, distant = d
9. gill-size: broad = b, narrow = n
10. gill-color: black = k, brown = n, buff = b, chocolate = h, gray = g, green = r, orange = o, pink = p, purple = u, red = e, white = w, yellow = y
11. stalk-shape: enlarging = e, tapering = t
12. stalk-root: bulbous = b, club = c, cup = u, equal = e, rhizomorphs = z, rooted = r
13. stalk-surface-above-ring: fibrous = f, scaly = y, silky = k, smooth = s
14. stalk-surface-below-ring: fibrous = f, scaly = y, silky = k, smooth = s
15. stalk-color-above-ring: brown = n, buff = b, cinnamon = c, gray = g, orange = o, pink = p, red = e, white = w, yellow = y
16. stalk-color-below-ring: brown = n, buff = b, cinnamon = c, gray = g, orange = o, pink = p, red = e, white = w, yellow = y
17. veil-type: partial = p, universal = u
18. veil-color: brown = n, orange = o, white = w, yellow = y
19. ring-number: none = n, one = o, two = t
20. ring-type: cobwebby = c, evanescent = e, flaring = f, large = l, none = n, pendant = p, sheathing = s, zone = z
21. spore-print-color: black = k, brown = n, buff = b, chocolate = h, green = r, orange = o, purple = u, white = w, yellow = y
22. population: abundant = a, clustered = c, numerous = n, scattered = s, several = v, solitary = y
23. habitat: grasses = g, leaves = l, meadows = m, paths = p, urban = u, waste = w, woods = d

After scanning the dataset, we noticed that there were 2,480 missing values, all of which fell under the stalk-root attribute. Although most of the attributes are uniform, some are quite skewed. For example, 98% of the values for the attribute veil-color are w and only 1% of the values are n. Another more extreme example occurs for the attribute veil-type where u does not appear at all in the dataset and every instance has p under veil-type. Attributes like ring-number and gill-spacing are similarly skewed, with one value being present for 92% and 84% of the instances in the dataset, respectively. On the other hand, the class values are pretty evenly distributed; 52% of the mushrooms in the dataset are edible while the other 48% are poisonous.

Part 3 – Preprocessing

The preprocessing of the mushroom dataset is carried out using Python in a Jupyter Notebook environment. Given that the dataset consists primarily of categorical variables, there are several steps necessary to prepare the data for attribute selection and modeling.

Part 3.1 – Handling Missing Values

The dataset contains 22 features, one of which, *stalk-root*, has missing values denoted by the symbol “?”. Rather than implementing a simpler approach such as imputing missing values using the mode of the attribute, we have implemented the K-Nearest-Neighbors (KNN) imputation method[1] from the scikit-learn library. The KNN imputation algorithm utilizes the similarity between instances in the dataset, estimating missing values by analyzing the closest neighbors based on feature values.

This method is chosen instead of mode as it allows the imputed values to reflect underlying patterns in the data more effectively. By utilizing this approach instead of mode or complete attribute deletion, we preserve the relationship between attributes and prevent potential bias in analysis and classification.

Part 3.2 – Encoding Categorical Variables into Numerical Data

Given that the features contain categorical data, it is necessary to convert them into numerical form to ensure compatibility with the machine learning models we plan to implement. Each feature is transformed into a numerical representation using label encoding, with each unique category within a feature assigned to a distinct integer starting from 0.

For example, the *cap-shape* and *odor* attributes have been arbitrarily encoded as follows:

- cap-shape
    - bell (b) → 0
    - conical (c) → 1
    - flat (f) → 2
    - knobbed (k) → 3
    - sunken (s) → 4
    - convex (x) → 5

- odor
    - almond (a) → 0
    - creosote (c) → 1
    - foul (f) → 2
    - anise (l) → 3
    - musty (m) → 4
    - none (n) → 5
    - pungent (p) → 6
    - spicy (s) → 7
    - fishy (y) → 8

[1] https://drive.google.com/drive/folders/1hnUSwRu4W7yZf5P-Dw0NbJQ8kPvf-cbH?usp=sharing

Part 3.3 – Dataset Splitting

To prepare the dataset for model training and evaluation, we are dividing the transformed dataset into three subsets):

- Training set (80% of the data): Trains the classification models to learn data patterns
- Validation set (10% of the data): Tunes hyperparameters and optimizes the models
- Testing set (10% of the data): Gives a final and unbiased evaluation of the model

This split ensures the model has sufficient data to learn patterns while leaving enough unseen data to evaluate the model holistically. It also maintains the original class distribution within subsets to prevent biases in model evaluation. This split resulted in 6,500 instances/mushrooms in the training subset, 812 instances/mushrooms in the validation subset, and 812 instances/mushrooms in the test subset.

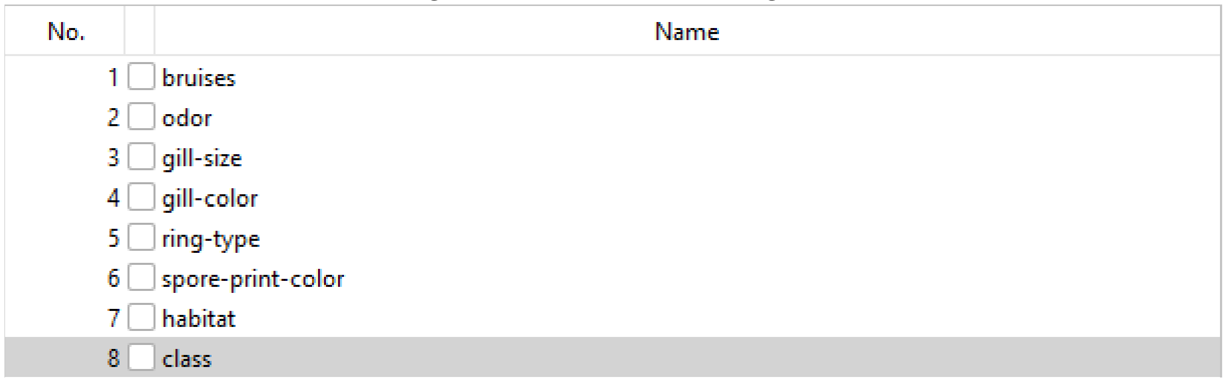
Part 4 – Attribute Selection Algorithms & Model Classifiers Used

The raw dataset, after preprocessing, consists of 22 attributes. To evaluate the performance of classifiers, we utilize several attribute selection methods to consider possible outcomes.

Part 4.1 – Attribute Selection Algorithms

A. *Intuition Based Selection* (non-Weka)
For this attribute selection algorithm, we did not use Weka.
The attributes that we intuitively thought were the most important characteristics when determining whether or not a mushroom is poisonous were chosen.

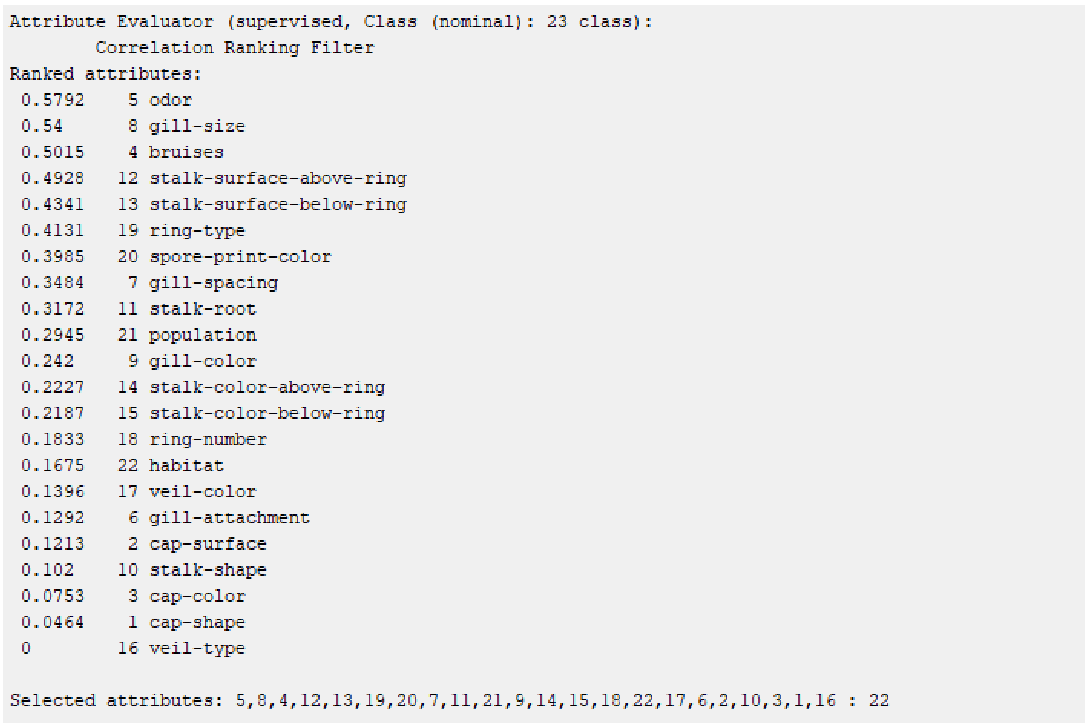
The retained attributes when using this attribute selection algorithm are:



No.		Name
1	☐	bruises
2	☐	odor
3	☐	gill-size
4	☐	gill-color
5	☐	ring-type
6	☐	spore-print-color
7	☐	habitat
8	☐	class

B. *CorrelationAttributeEval*
For this attribute selection algorithm, we used Weka. This method measures the linear correlation between attributes and class labels (“Package”, n.d.).

The image below displays the output from Weka when running *CorrelationAttributeEval*:



```
Attribute Evaluator (supervised, Class (nominal): 23 class):
        Correlation Ranking Filter
Ranked attributes:
 0.5792    5 odor
 0.54      8 gill-size
 0.5015    4 bruises
 0.4928   12 stalk-surface-above-ring
 0.4341   13 stalk-surface-below-ring
 0.4131   19 ring-type
 0.3985   20 spore-print-color
 0.3484    7 gill-spacing
 0.3172   11 stalk-root
 0.2945   21 population
 0.242     9 gill-color
 0.2227   14 stalk-color-above-ring
 0.2187   15 stalk-color-below-ring
 0.1833   18 ring-number
 0.1675   22 habitat
 0.1396   17 veil-color
 0.1292    6 gill-attachment
 0.1213    2 cap-surface
 0.102    10 stalk-shape
 0.0753    3 cap-color
 0.0464    1 cap-shape
 0         16 veil-type

Selected attributes: 5,8,4,12,13,19,20,7,11,21,9,14,15,18,22,17,6,2,10,3,1,16 : 22
```

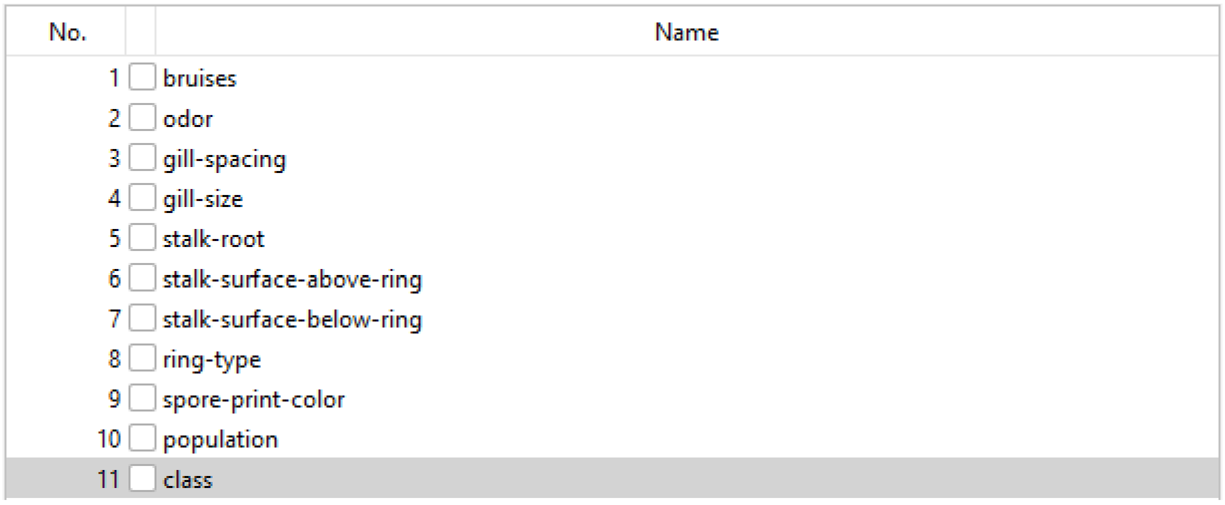
The attributes that had a **Pearson Correlation Coefficient absolute value greater than or equal to 0.25** were chosen.

The Pearson Correlation Coefficient (also denoted as r) is calculated as follows:

Given an attribute X and class Y:

$$r = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma(X_i - \overline{X})^2 \Sigma(Y_i - \overline{Y})^2}}$$

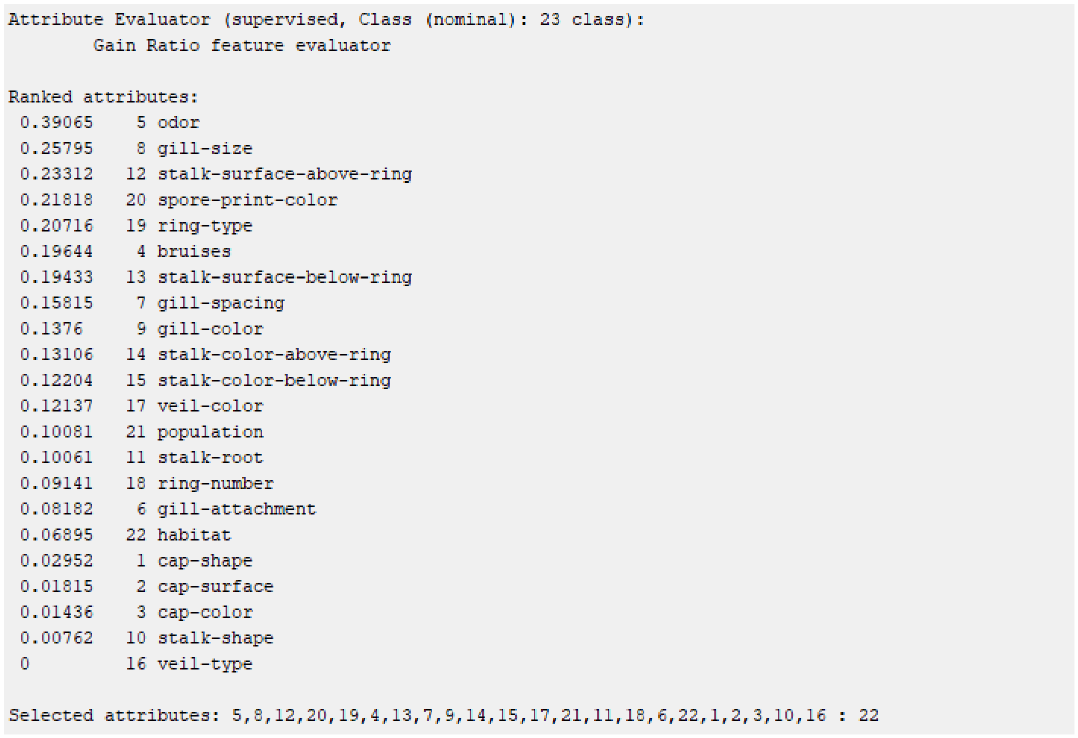
The retained attributes when using this attribute selection algorithm are:



No.		Name
1	☐	bruises
2	☐	odor
3	☐	gill-spacing
4	☐	gill-size
5	☐	stalk-root
6	☐	stalk-surface-above-ring
7	☐	stalk-surface-below-ring
8	☐	ring-type
9	☐	spore-print-color
10	☐	population
11	☐	class

C. *GainRatioAttributeEval*
For this attribute selection algorithm, we used Weka. This method ranks attributes based on their gain ratio with respect to the class. Gain ratio is designed to reduce the bias towards attributes that have many distinct values ("Package", n.d.).

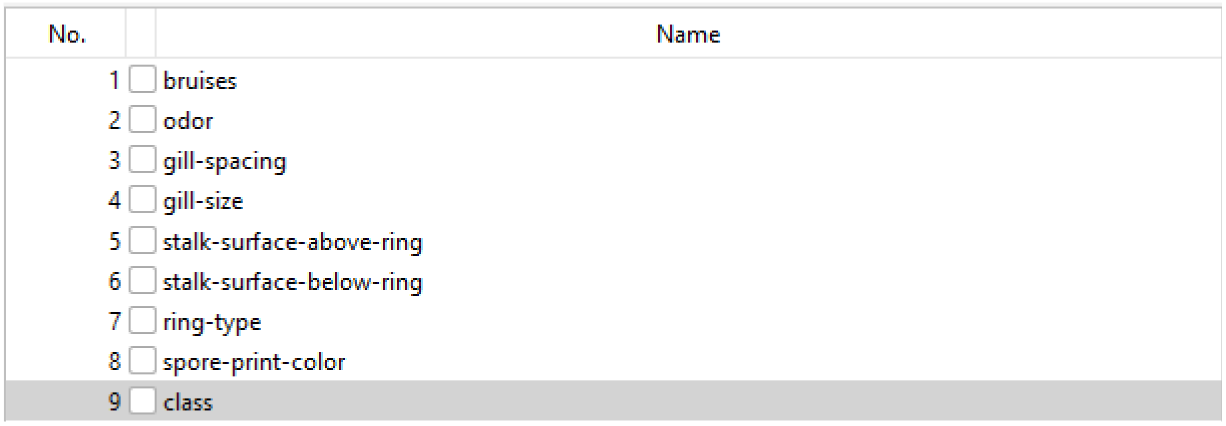
The image below displays the output from Weka when running *GainRatioAttributeEval*:



```
Attribute Evaluator (supervised, Class (nominal): 23 class):
        Gain Ratio feature evaluator

Ranked attributes:
 0.39065    5 odor
 0.25795    8 gill-size
 0.23312   12 stalk-surface-above-ring
 0.21818   20 spore-print-color
 0.20716   19 ring-type
 0.19644    4 bruises
 0.19433   13 stalk-surface-below-ring
 0.15815    7 gill-spacing
 0.1376     9 gill-color
 0.13106   14 stalk-color-above-ring
 0.12204   15 stalk-color-below-ring
 0.12137   17 veil-color
 0.10081   21 population
 0.10061   11 stalk-root
 0.09141   18 ring-number
 0.08182    6 gill-attachment
 0.06895   22 habitat
 0.02952    1 cap-shape
 0.01815    2 cap-surface
 0.01436    3 cap-color
 0.00762   10 stalk-shape
 0         16 veil-type

Selected attributes: 5,8,12,20,19,4,13,7,9,14,15,17,21,11,18,6,22,1,2,3,10,16 : 22
```

The attributes that had a **Gain Ratio absolute value greater than or equal to 0.15** were chosen.

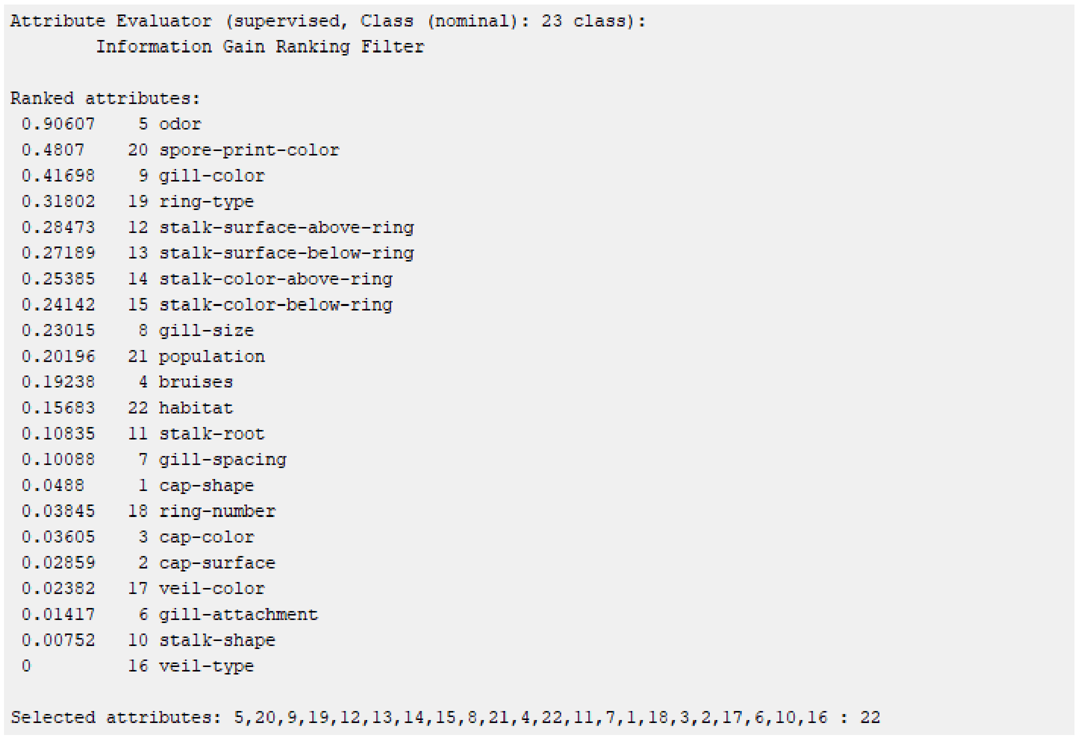
The retained attributes when using this attribute selection algorithm are:



No.		Name
1	☐	bruises
2	☐	odor
3	☐	gill-spacing
4	☐	gill-size
5	☐	stalk-surface-above-ring
6	☐	stalk-surface-below-ring
7	☐	ring-type
8	☐	spore-print-color
9	☐	class

D. *InfoGainAttributeEval*
For this attribute selection algorithm, we used Weka. This method ranks attributes based on their information gain with respect to the class. Information gain measures the reduction in entropy or uncertainty about the class when an attribute is known ("Package", n.d.).

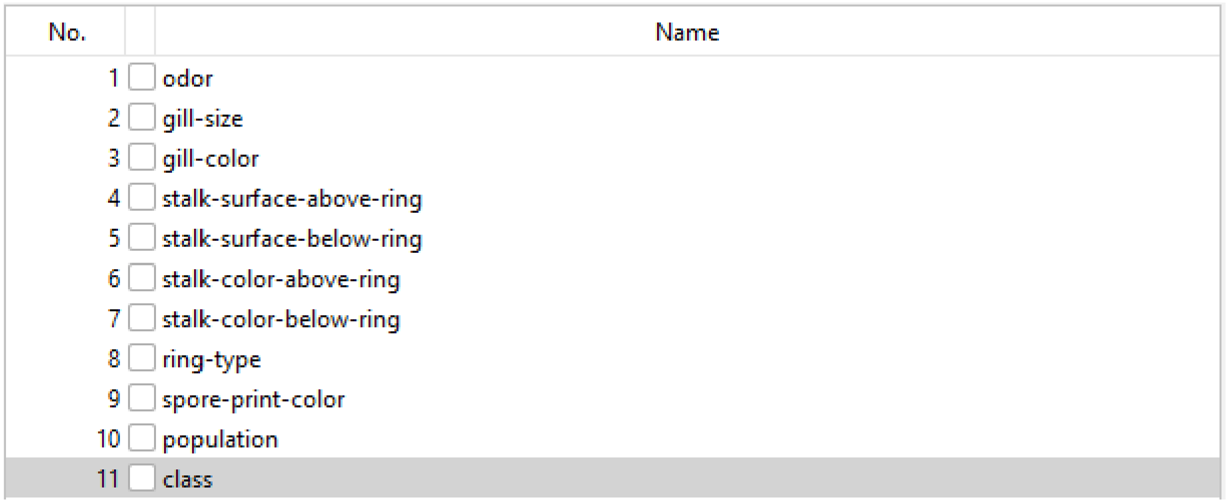
The image below displays the output from Weka when running *InfoGainAttributeEval*:



```
Attribute Evaluator (supervised, Class (nominal): 23 class):
        Information Gain Ranking Filter

Ranked attributes:
 0.90607    5 odor
 0.4807    20 spore-print-color
 0.41698    9 gill-color
 0.31802   19 ring-type
 0.28473   12 stalk-surface-above-ring
 0.27189   13 stalk-surface-below-ring
 0.25385   14 stalk-color-above-ring
 0.24142   15 stalk-color-below-ring
 0.23015    8 gill-size
 0.20196   21 population
 0.19238    4 bruises
 0.15683   22 habitat
 0.10835   11 stalk-root
 0.10088    7 gill-spacing
 0.0488     1 cap-shape
 0.03845   18 ring-number
 0.03605    3 cap-color
 0.02859    2 cap-surface
 0.02382   17 veil-color
 0.01417    6 gill-attachment
 0.00752   10 stalk-shape
 0         16 veil-type

Selected attributes: 5,20,9,19,12,13,14,15,8,21,4,22,11,7,1,18,3,2,17,6,10,16 : 22
```

The attributes that had an **Information Gain absolute value greater than or equal to 0.2** were chosen.

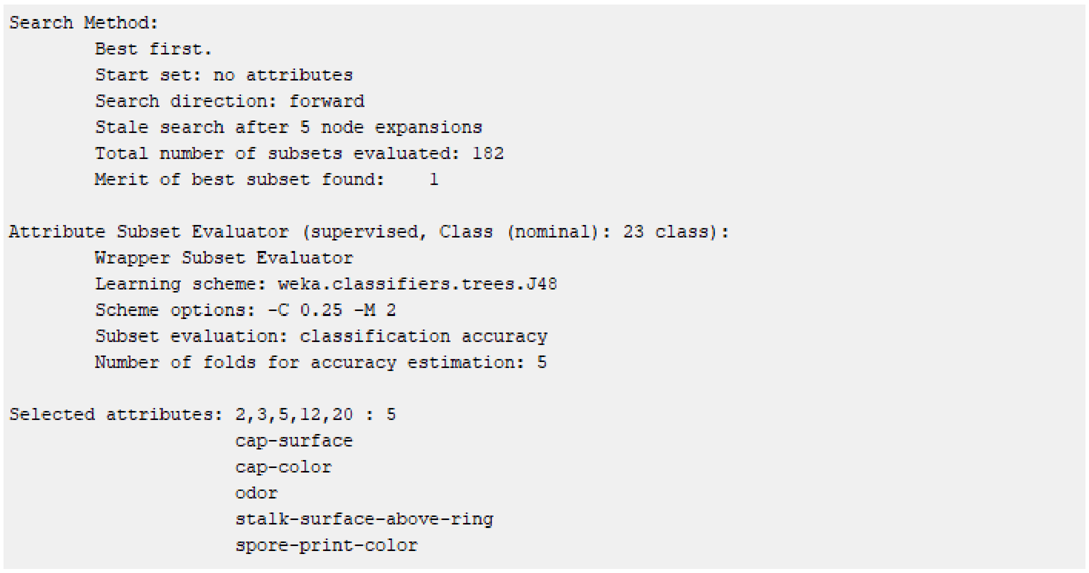
The retained attributes when using this attribute selection algorithm are:



No.	Name
1	odor
2	gill-size
3	gill-color
4	stalk-surface-above-ring
5	stalk-surface-below-ring
6	stalk-color-above-ring
7	stalk-color-below-ring
8	ring-type
9	spore-print-color
10	population
11	class

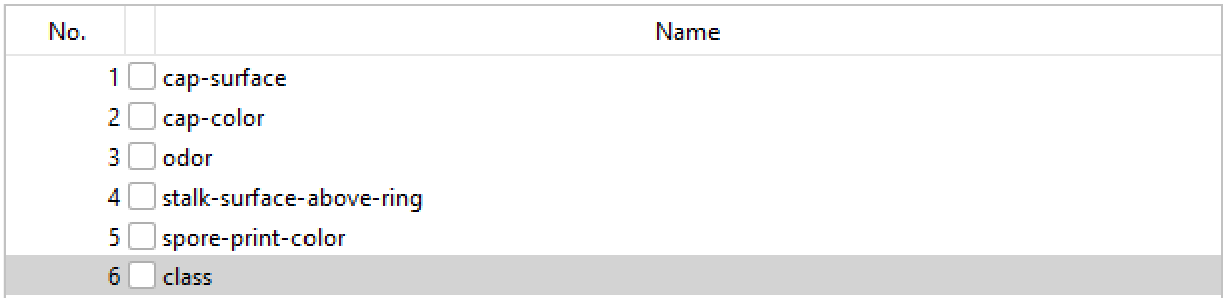
E. *WrapperSubsetEval (+ Best First)*
For this attribute selection algorithm, we used Weka. This attribute selection algorithm evaluates the performance of a subset of attributes using a classifier and selects the subset of attributes that performs the best ("Package", n.d.). We used the J48 Decision Tree as our classifier.

The image below displays the output from Weka when running *WrapperSubsetEval* with a J48 Decision Tree classifier and using the *BestFirst* search method.



```
Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 182
        Merit of best subset found:    1

Attribute Subset Evaluator (supervised, Class (nominal): 23 class):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.J48
        Scheme options: -C 0.25 -M 2
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 2,3,5,12,20 : 5
                     cap-surface
                     cap-color
                     odor
                     stalk-surface-above-ring
                     spore-print-color
```

The retained attributes when using this attribute selection algorithm are:



No.		Name
1	☐	cap-surface
2	☐	cap-color
3	☐	odor
4	☐	stalk-surface-above-ring
5	☐	spore-print-color
6	☐	class

Part 4.2 – Classifier Models

For this study, we utilized models from the scikit-learn Python library instead of Weka. This allowed us to manually control our preprocessing, derive training metrics for the validation set, and compare models in a single location. The following classifiers were selected for their relevance to binary classification problems, such as determining mushroom edibility:

A. Decision Tree (J48)
This classifier builds a decision tree that extrapolates data to approximate a sine curve with a set of decision rules ("DecisionTreeClassifier", n.d.). When the depth of the decision tree increases, these decision rules become more complex and the model is fitter. After the model is fitted, it can then be used to predict the class for a certain instance.

B. *Quadratic Discriminant Analysis (QDA)*
This classifier generates a quadratic decision boundary by fitting class conditional densities to the dataset, specifically fitting a Gaussian density to each class, and utilizing Bayes' rule ("QuadraticDiscriminantAnalysis", n.d.). More specifically, QDA selects the class k that maximizes the probability $P(y = k|x)$, which is used when creating a class conditional distribution.

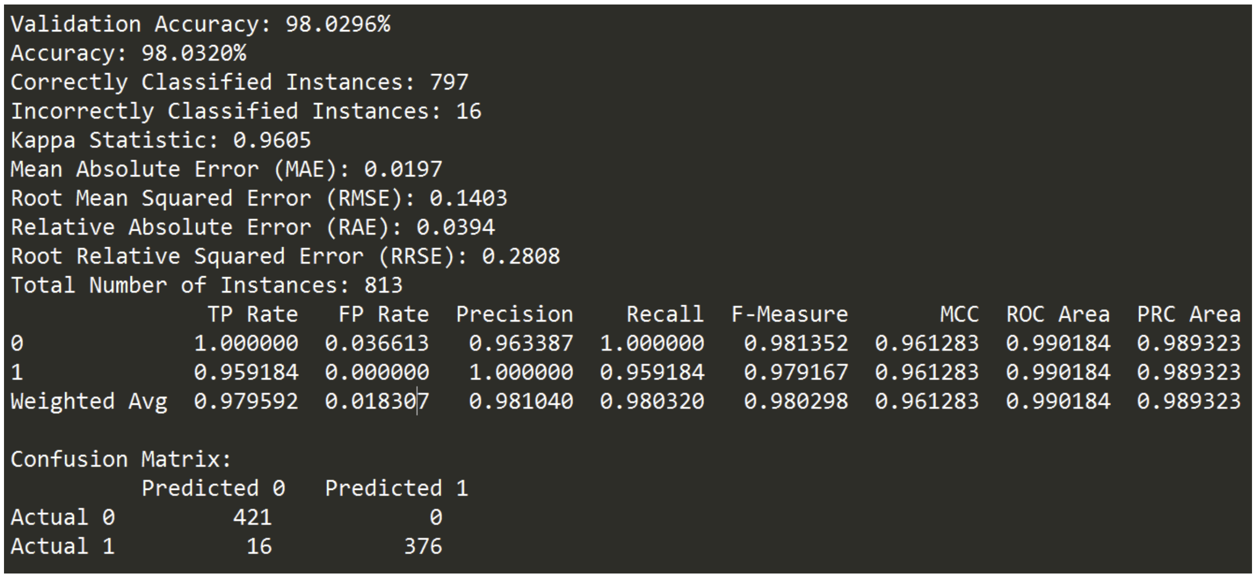
C. *Logistic Regression*
This classifier, as a special case of a generalized linear model (GLM), utilizes a Bernoulli distribution to model the log-odds as a linear combination of one or more independent variables ("LogisticRegression", n.d.). The logistic regression outputs a numerical value that represents the predicted probability for binary outcomes and can be used as a classifier by applying a threshold to it.

D. Support Vector Classifier (SVC)
This classifier uses support vector machines, which are supervised max-margin models that can efficiently perform non-linear classification using the kernel trick ("1.4. Support Vector Machines", n.d.). Support vector machines are also resilient to data that is noisy (i.e. where examples are misclassified) since they find an optimal hyperplane for the data, which enhances their performance.

Part 5 – Results and Analysis

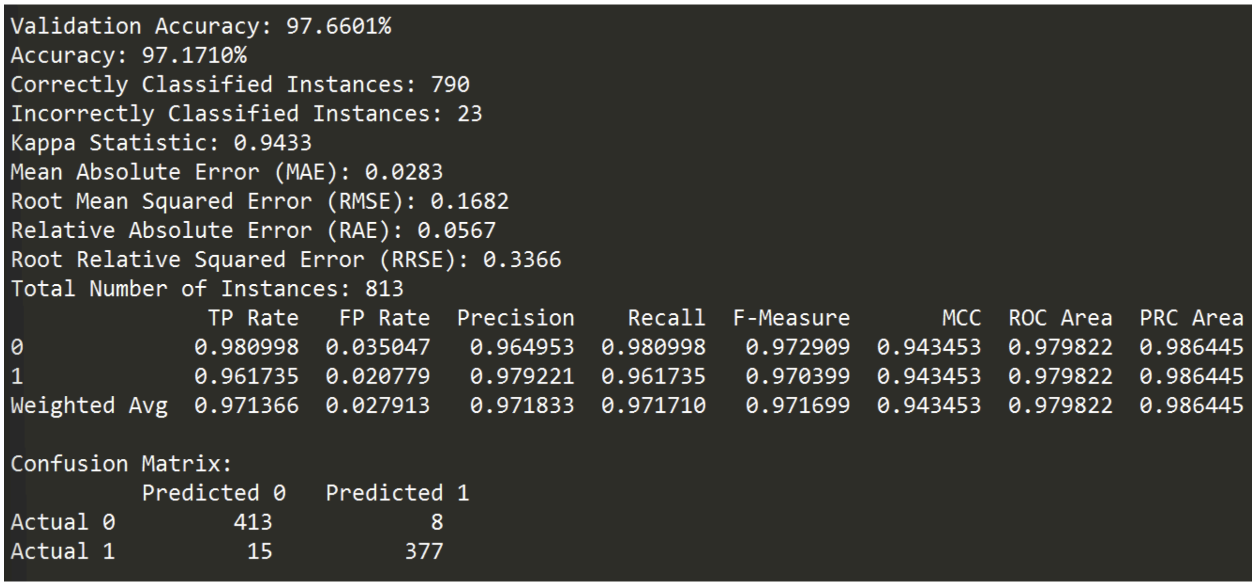
Part 5.1 – Results

<u>Correlation with J48 Decision Tree</u>



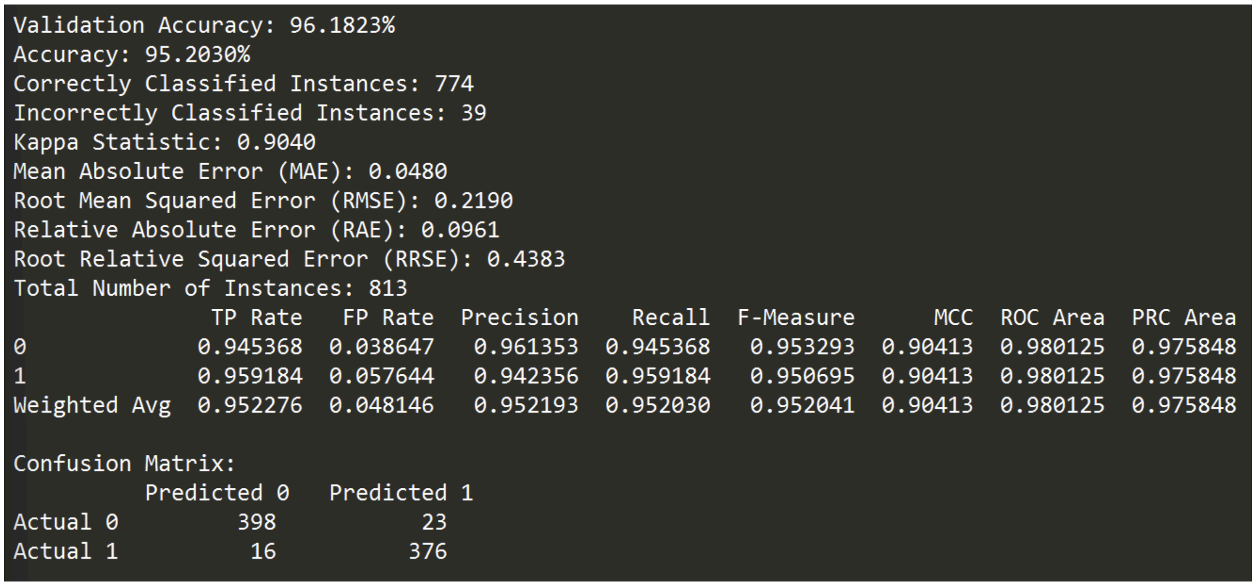
```
Validation Accuracy: 98.0296%
Accuracy: 98.0320%
Correctly Classified Instances: 797
Incorrectly Classified Instances: 16
Kappa Statistic: 0.9605
Mean Absolute Error (MAE): 0.0197
Root Mean Squared Error (RMSE): 0.1403
Relative Absolute Error (RAE): 0.0394
Root Relative Squared Error (RRSE): 0.2808
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             1.000000  0.036613   0.963387  1.000000   0.981352  0.961283  0.990184  0.989323
1             0.959184  0.000000   1.000000  0.959184   0.979167  0.961283  0.990184  0.989323
Weighted Avg  0.979592  0.018307   0.981040  0.980320   0.980298  0.961283  0.990184  0.989323

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         421            0
Actual 1          16          376
```

<u>Correlation with Quadratic Discriminant Analysis (QDA)</u>

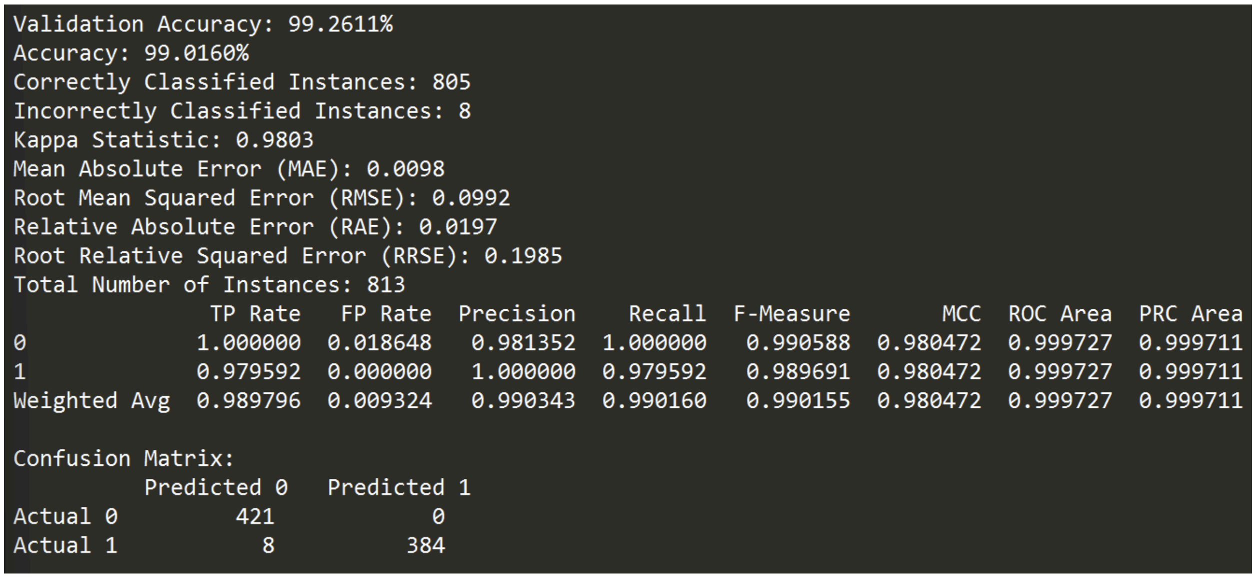


```
Validation Accuracy: 97.6601%
Accuracy: 97.1710%
Correctly Classified Instances: 790
Incorrectly Classified Instances: 23
Kappa Statistic: 0.9433
Mean Absolute Error (MAE): 0.0283
Root Mean Squared Error (RMSE): 0.1682
Relative Absolute Error (RAE): 0.0567
Root Relative Squared Error (RRSE): 0.3366
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             0.980998  0.035047   0.964953  0.980998   0.972909  0.943453  0.979822  0.986445
1             0.961735  0.020779   0.979221  0.961735   0.970399  0.943453  0.979822  0.986445
Weighted Avg  0.971366  0.027913   0.971833  0.971710   0.971699  0.943453  0.979822  0.986445

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         413            8
Actual 1          15          377
```

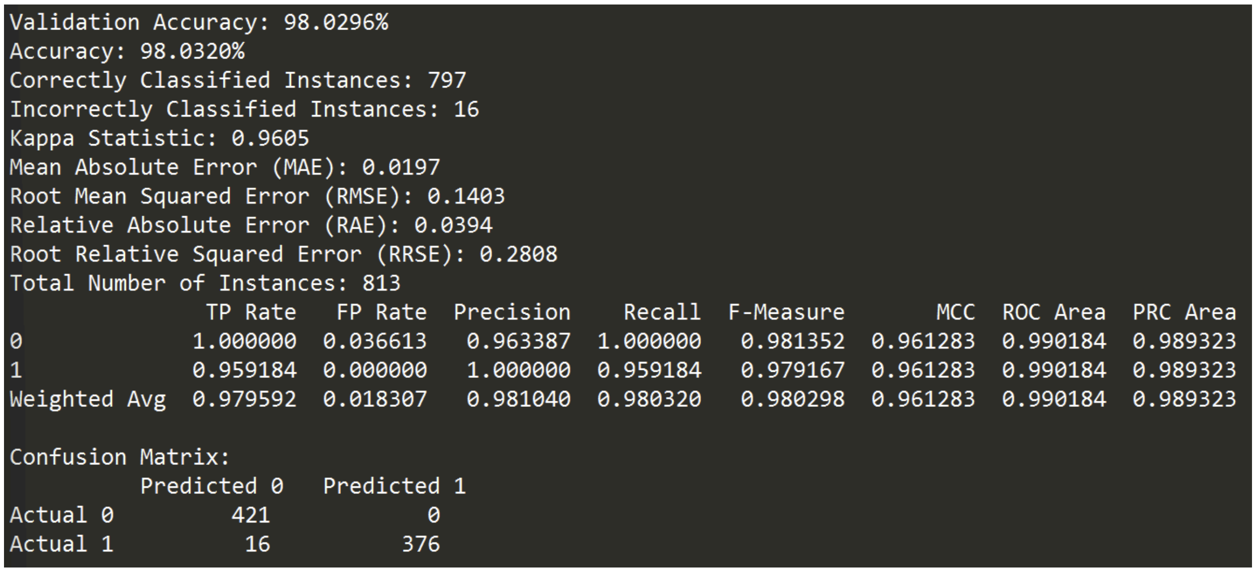
Correlation with Logistic Regression


```
Validation Accuracy: 96.1823%
Accuracy: 95.2030%
Correctly Classified Instances: 774
Incorrectly Classified Instances: 39
Kappa Statistic: 0.9040
Mean Absolute Error (MAE): 0.0480
Root Mean Squared Error (RMSE): 0.2190
Relative Absolute Error (RAE): 0.0961
Root Relative Squared Error (RRSE): 0.4383
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure      MCC  ROC Area  PRC Area
0             0.945368  0.038647   0.961353  0.945368   0.953293  0.90413  0.980125  0.975848
1             0.959184  0.057644   0.942356  0.959184   0.950695  0.90413  0.980125  0.975848
Weighted Avg  0.952276  0.048146   0.952193  0.952030   0.952041  0.90413  0.980125  0.975848

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         398            23
Actual 1          16           376
```

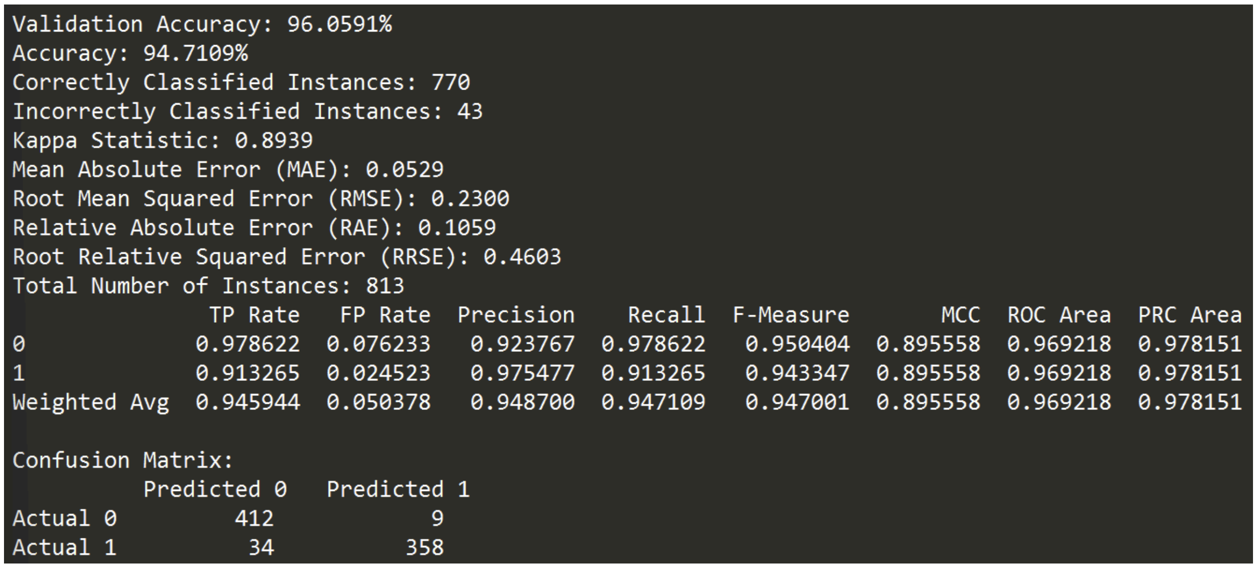
Correlation with Support Vector Classifier (SVC)


```
Validation Accuracy: 99.2611%
Accuracy: 99.0160%
Correctly Classified Instances: 805
Incorrectly Classified Instances: 8
Kappa Statistic: 0.9803
Mean Absolute Error (MAE): 0.0098
Root Mean Squared Error (RMSE): 0.0992
Relative Absolute Error (RAE): 0.0197
Root Relative Squared Error (RRSE): 0.1985
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             1.000000  0.018648   0.981352  1.000000   0.990588  0.980472  0.999727  0.999711
1             0.979592  0.000000   1.000000  0.979592   0.989691  0.980472  0.999727  0.999711
Weighted Avg  0.989796  0.009324   0.990343  0.990160   0.990155  0.980472  0.999727  0.999711

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         421             0
Actual 1           8           384
```

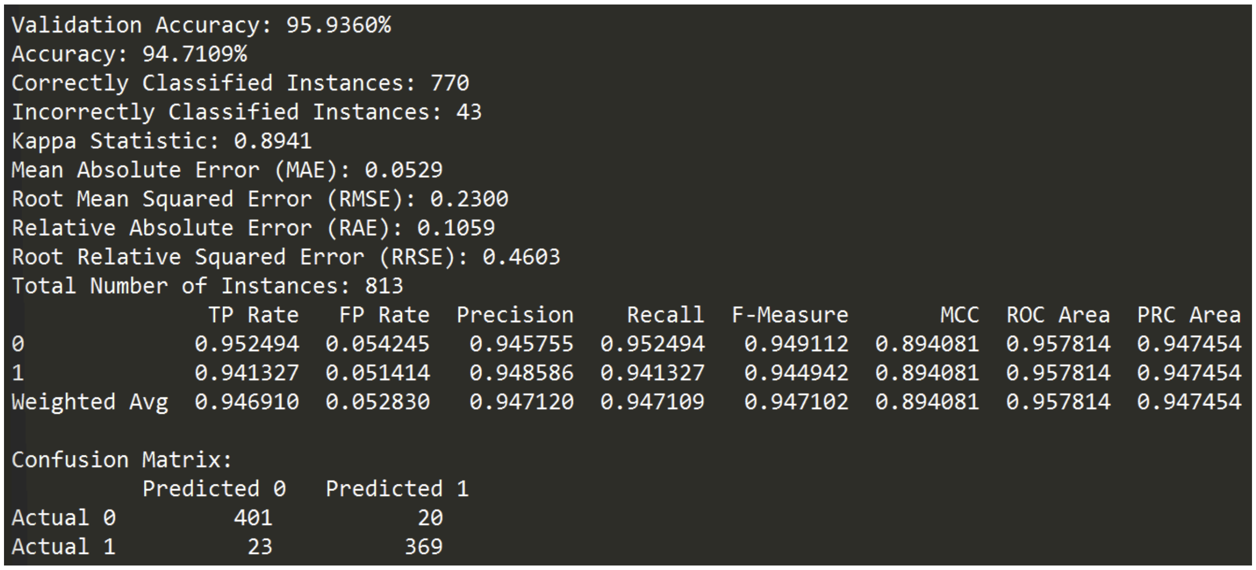
Gain-Ratio with J48 Decision Tree


```
Validation Accuracy: 98.0296%
Accuracy: 98.0320%
Correctly Classified Instances: 797
Incorrectly Classified Instances: 16
Kappa Statistic: 0.9605
Mean Absolute Error (MAE): 0.0197
Root Mean Squared Error (RMSE): 0.1403
Relative Absolute Error (RAE): 0.0394
Root Relative Squared Error (RRSE): 0.2808
Total Number of Instances: 813
              TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0            1.000000  0.036613   0.963387  1.000000   0.981352  0.961283  0.990184  0.989323
1            0.959184  0.000000   1.000000  0.959184   0.979167  0.961283  0.990184  0.989323
Weighted Avg  0.979592  0.018307   0.981040  0.980320   0.980298  0.961283  0.990184  0.989323

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0        421             0
Actual 1         16           376
```

Gain-Ratio with Quadratic Discriminant Analysis (QDA)

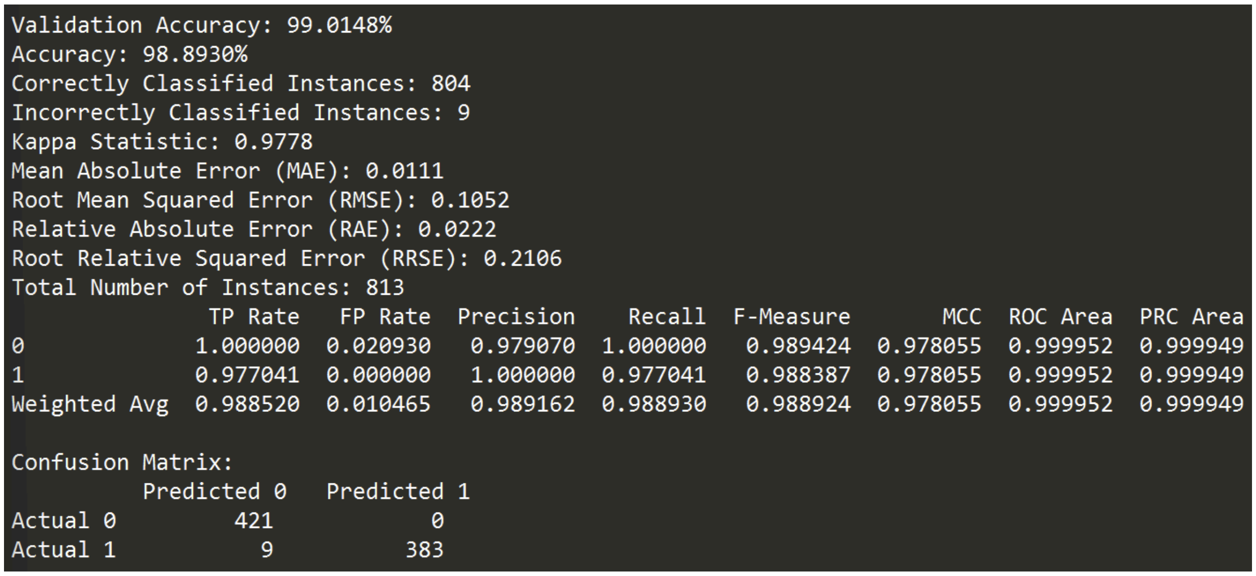

```
Validation Accuracy: 96.0591%
Accuracy: 94.7109%
Correctly Classified Instances: 770
Incorrectly Classified Instances: 43
Kappa Statistic: 0.8939
Mean Absolute Error (MAE): 0.0529
Root Mean Squared Error (RMSE): 0.2300
Relative Absolute Error (RAE): 0.1059
Root Relative Squared Error (RRSE): 0.4603
Total Number of Instances: 813
              TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0            0.978622  0.076233   0.923767  0.978622   0.950404  0.895558  0.969218  0.978151
1            0.913265  0.024523   0.975477  0.913265   0.943347  0.895558  0.969218  0.978151
Weighted Avg  0.945944  0.050378   0.948700  0.947109   0.947001  0.895558  0.969218  0.978151

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0        412             9
Actual 1         34           358
```

Gain-Ratio with Logistic Regression



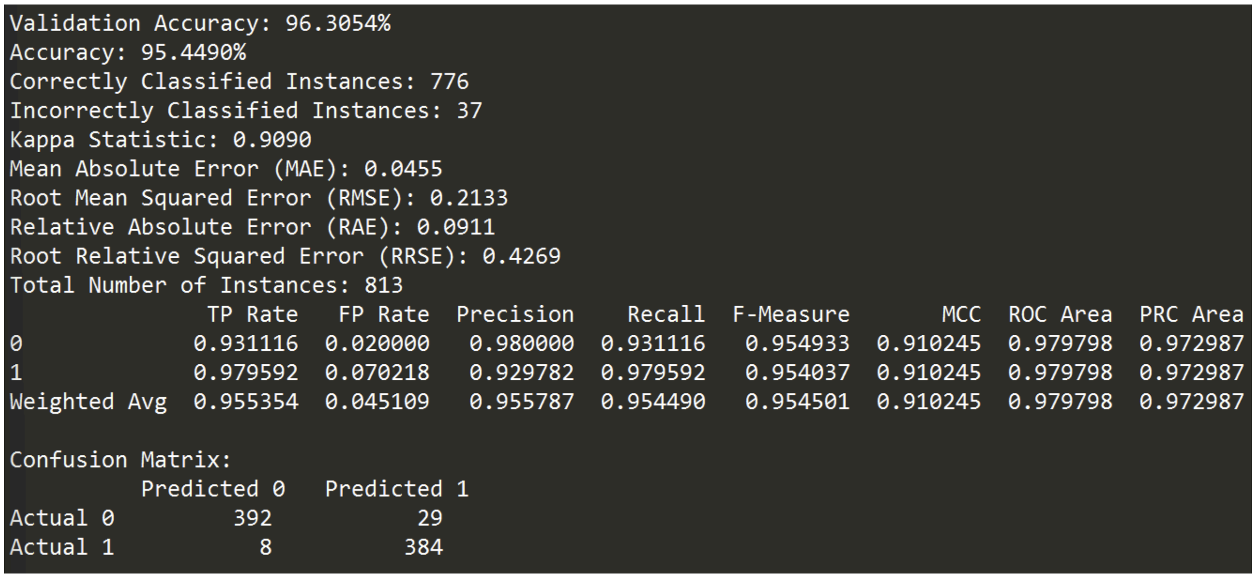
```
Validation Accuracy: 95.9360%
Accuracy: 94.7109%
Correctly Classified Instances: 770
Incorrectly Classified Instances: 43
Kappa Statistic: 0.8941
Mean Absolute Error (MAE): 0.0529
Root Mean Squared Error (RMSE): 0.2300
Relative Absolute Error (RAE): 0.1059
Root Relative Squared Error (RRSE): 0.4603
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             0.952494  0.054245   0.945755  0.952494   0.949112  0.894081  0.957814  0.947454
1             0.941327  0.051414   0.948586  0.941327   0.944942  0.894081  0.957814  0.947454
Weighted Avg  0.946910  0.052830   0.947120  0.947109   0.947102  0.894081  0.957814  0.947454

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         401           20
Actual 1          23          369
```

Gain-Ratio with Support Vector Classifier (SVC)



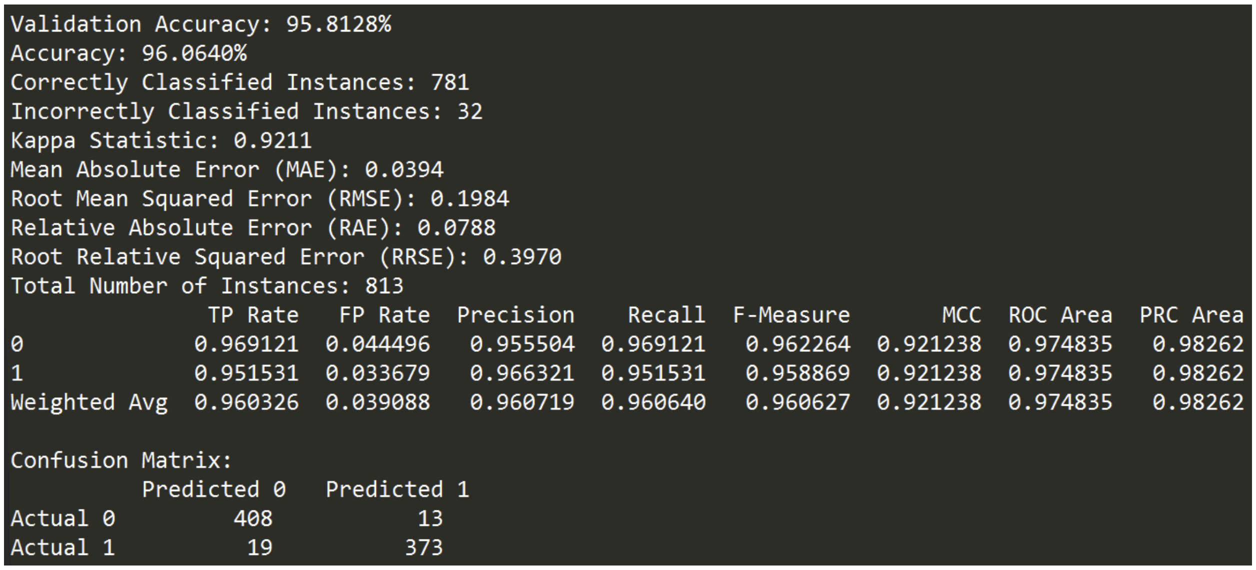
```
Validation Accuracy: 99.0148%
Accuracy: 98.8930%
Correctly Classified Instances: 804
Incorrectly Classified Instances: 9
Kappa Statistic: 0.9778
Mean Absolute Error (MAE): 0.0111
Root Mean Squared Error (RMSE): 0.1052
Relative Absolute Error (RAE): 0.0222
Root Relative Squared Error (RRSE): 0.2106
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             1.000000  0.020930   0.979070  1.000000   0.989424  0.978055  0.999952  0.999949
1             0.977041  0.000000   1.000000  0.977041   0.988387  0.978055  0.999952  0.999949
Weighted Avg  0.988520  0.010465   0.989162  0.988930   0.988924  0.978055  0.999952  0.999949

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         421            0
Actual 1           9          383
```

Info-Gain with J48 Decision Tree



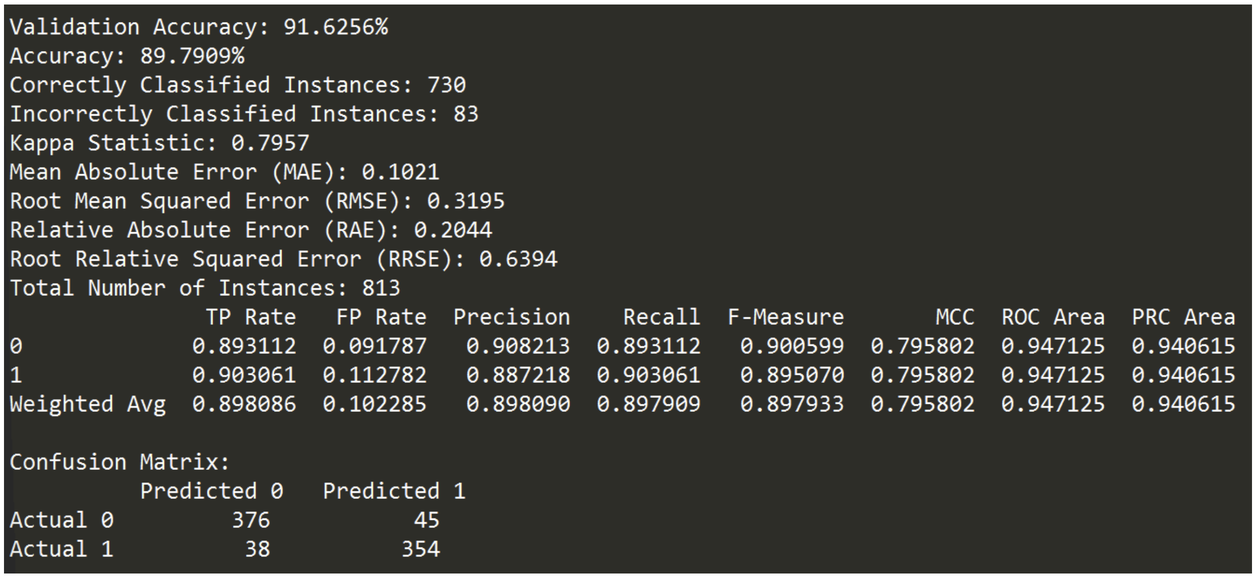
```
Validation Accuracy: 96.3054%
Accuracy: 95.4490%
Correctly Classified Instances: 776
Incorrectly Classified Instances: 37
Kappa Statistic: 0.9090
Mean Absolute Error (MAE): 0.0455
Root Mean Squared Error (RMSE): 0.2133
Relative Absolute Error (RAE): 0.0911
Root Relative Squared Error (RRSE): 0.4269
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             0.931116  0.020000   0.980000  0.931116   0.954933  0.910245  0.979798  0.972987
1             0.979592  0.070218   0.929782  0.979592   0.954037  0.910245  0.979798  0.972987
Weighted Avg  0.955354  0.045109   0.955787  0.954490   0.954501  0.910245  0.979798  0.972987

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         392           29
Actual 1           8          384
```

Info-Gain with Quadratic Discriminant Analysis (QDA)



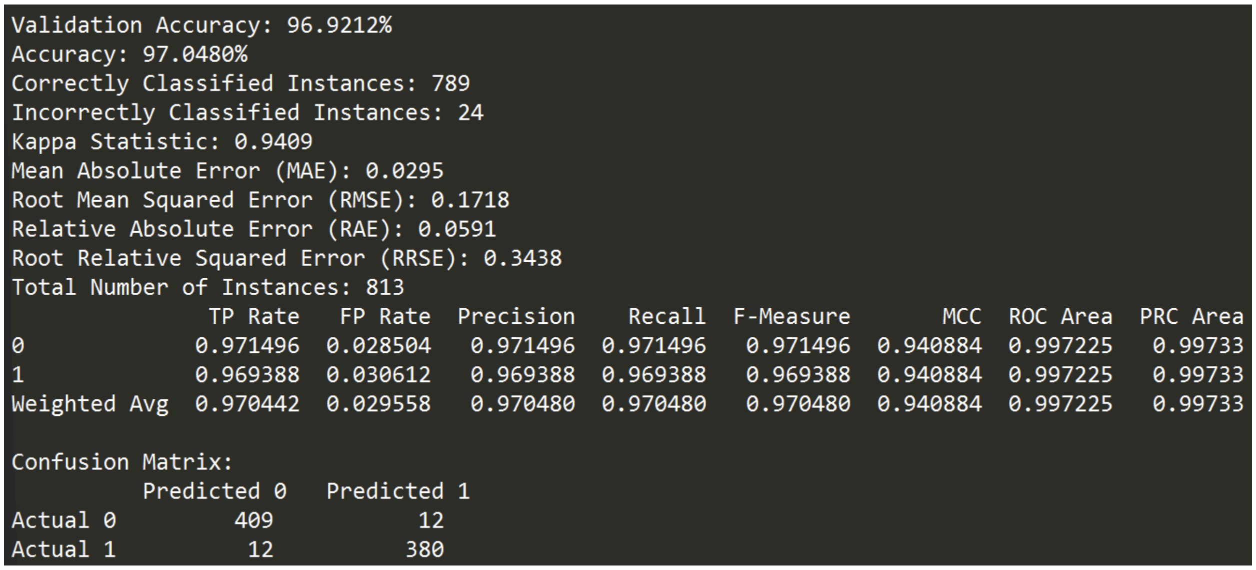
```
Validation Accuracy: 95.8128%
Accuracy: 96.0640%
Correctly Classified Instances: 781
Incorrectly Classified Instances: 32
Kappa Statistic: 0.9211
Mean Absolute Error (MAE): 0.0394
Root Mean Squared Error (RMSE): 0.1984
Relative Absolute Error (RAE): 0.0788
Root Relative Squared Error (RRSE): 0.3970
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             0.969121  0.044496   0.955504  0.969121   0.962264  0.921238  0.974835   0.98262
1             0.951531  0.033679   0.966321  0.951531   0.958869  0.921238  0.974835   0.98262
Weighted Avg  0.960326  0.039088   0.960719  0.960640   0.960627  0.921238  0.974835   0.98262

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         408           13
Actual 1          19          373
```

Info-Gain with Logistic Regression



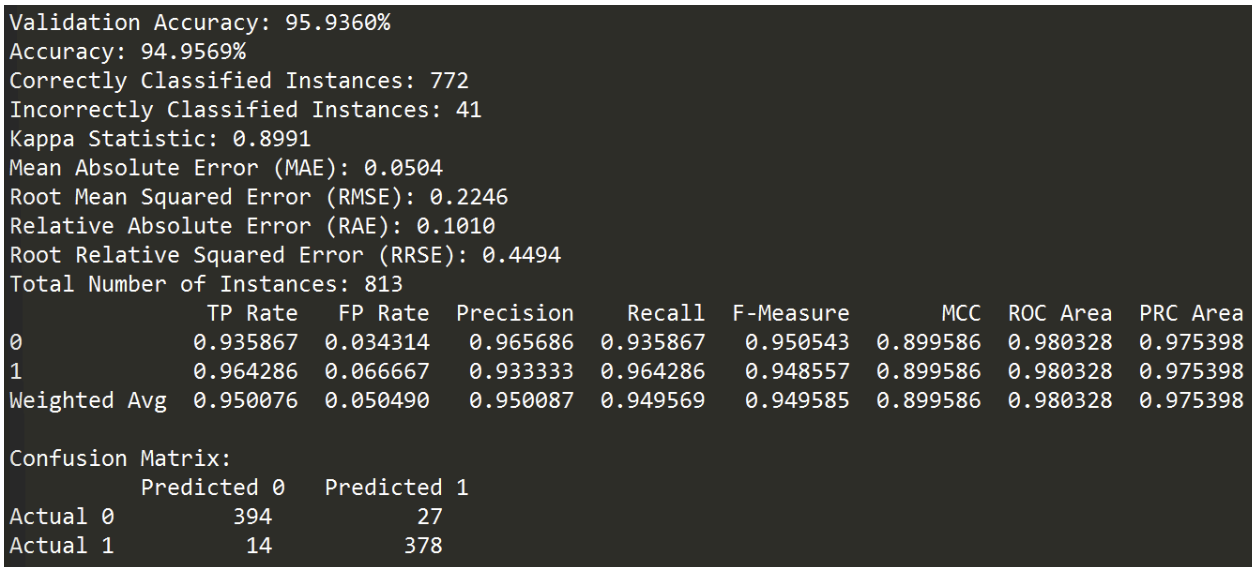
```
Validation Accuracy: 91.6256%
Accuracy: 89.7909%
Correctly Classified Instances: 730
Incorrectly Classified Instances: 83
Kappa Statistic: 0.7957
Mean Absolute Error (MAE): 0.1021
Root Mean Squared Error (RMSE): 0.3195
Relative Absolute Error (RAE): 0.2044
Root Relative Squared Error (RRSE): 0.6394
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             0.893112  0.091787   0.908213  0.893112   0.900599  0.795802  0.947125  0.940615
1             0.903061  0.112782   0.887218  0.903061   0.895070  0.795802  0.947125  0.940615
Weighted Avg  0.898086  0.102285   0.898090  0.897909   0.897933  0.795802  0.947125  0.940615

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         376            45
Actual 1          38           354
```

Info-Gain with Support Vector Classifier (SVC)



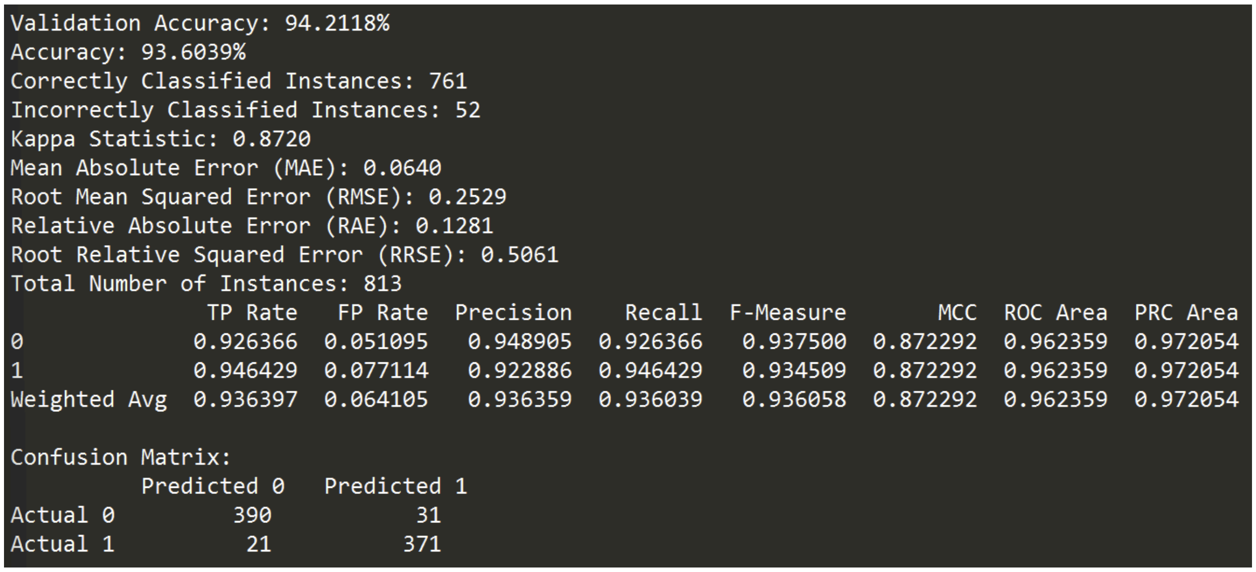
```
Validation Accuracy: 96.9212%
Accuracy: 97.0480%
Correctly Classified Instances: 789
Incorrectly Classified Instances: 24
Kappa Statistic: 0.9409
Mean Absolute Error (MAE): 0.0295
Root Mean Squared Error (RMSE): 0.1718
Relative Absolute Error (RAE): 0.0591
Root Relative Squared Error (RRSE): 0.3438
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             0.971496  0.028504   0.971496  0.971496   0.971496  0.940884  0.997225   0.99733
1             0.969388  0.030612   0.969388  0.969388   0.969388  0.940884  0.997225   0.99733
Weighted Avg  0.970442  0.029558   0.970480  0.970480   0.970480  0.940884  0.997225   0.99733

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         409            12
Actual 1          12           380
```

Intuition (Non-Weka Attribute Selection) with J48 Decision Tree



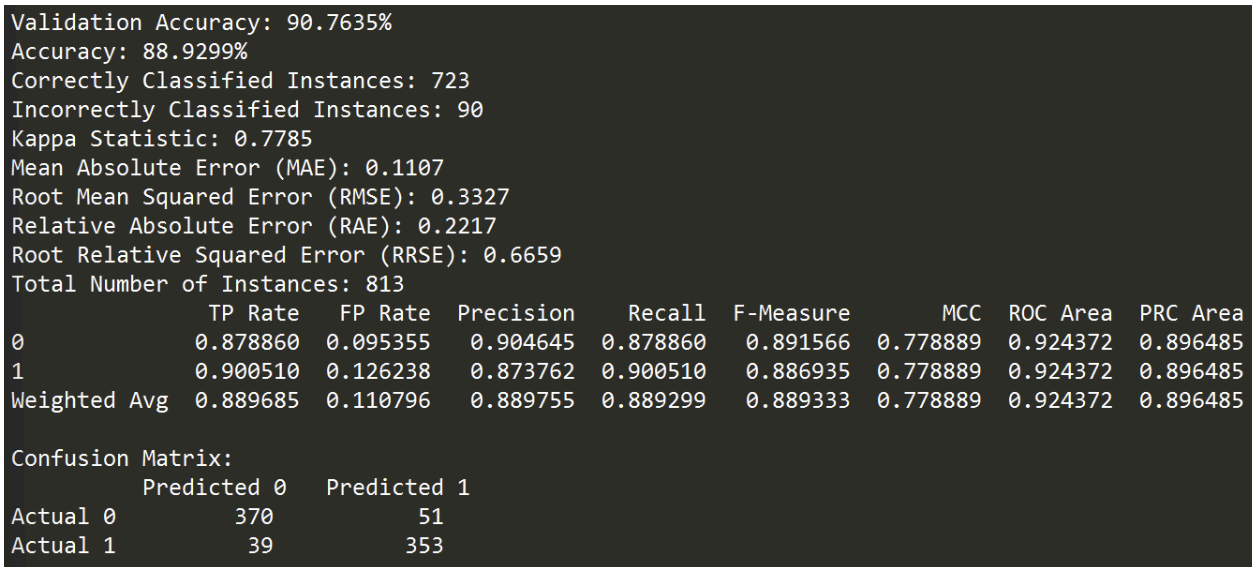
```
Validation Accuracy: 95.9360%
Accuracy: 94.9569%
Correctly Classified Instances: 772
Incorrectly Classified Instances: 41
Kappa Statistic: 0.8991
Mean Absolute Error (MAE): 0.0504
Root Mean Squared Error (RMSE): 0.2246
Relative Absolute Error (RAE): 0.1010
Root Relative Squared Error (RRSE): 0.4494
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             0.935867  0.034314   0.965686  0.935867   0.950543  0.899586  0.980328  0.975398
1             0.964286  0.066667   0.933333  0.964286   0.948557  0.899586  0.980328  0.975398
Weighted Avg  0.950076  0.050490   0.950087  0.949569   0.949585  0.899586  0.980328  0.975398

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         394           27
Actual 1          14          378
```

Intuition (Non-Weka Attribute Selection) with Quadratic Discriminant Analysis (QDA)



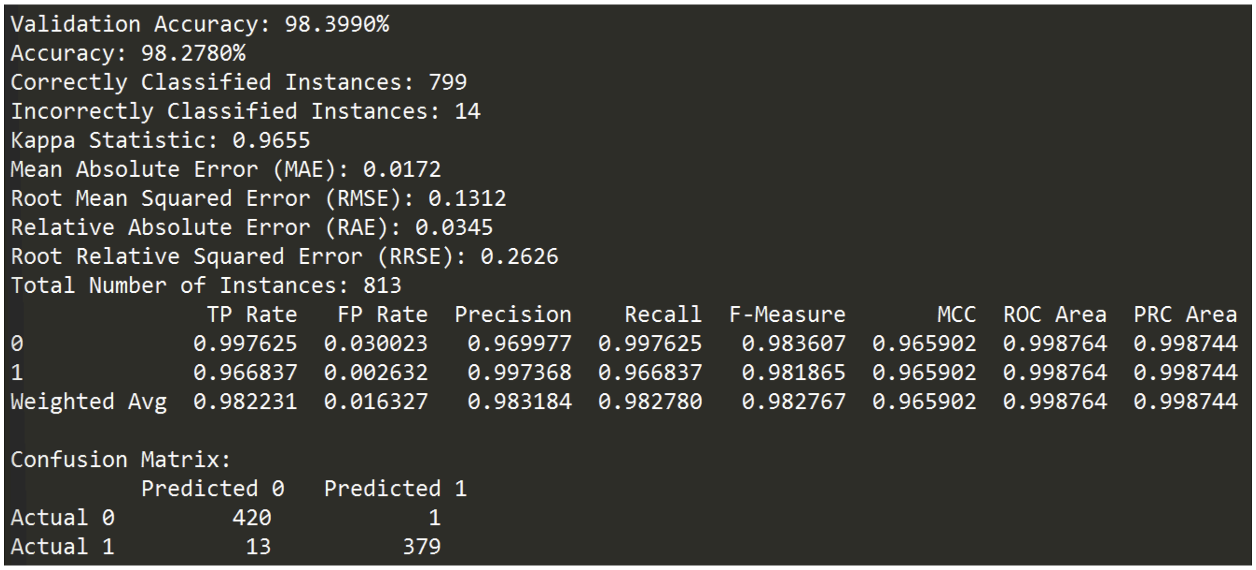
```
Validation Accuracy: 94.2118%
Accuracy: 93.6039%
Correctly Classified Instances: 761
Incorrectly Classified Instances: 52
Kappa Statistic: 0.8720
Mean Absolute Error (MAE): 0.0640
Root Mean Squared Error (RMSE): 0.2529
Relative Absolute Error (RAE): 0.1281
Root Relative Squared Error (RRSE): 0.5061
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             0.926366  0.051095   0.948905  0.926366   0.937500  0.872292  0.962359  0.972054
1             0.946429  0.077114   0.922886  0.946429   0.934509  0.872292  0.962359  0.972054
Weighted Avg  0.936397  0.064105   0.936359  0.936039   0.936058  0.872292  0.962359  0.972054

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         390           31
Actual 1          21          371
```

Intuition (Non-Weka Attribute Selection) with Logistic Regression



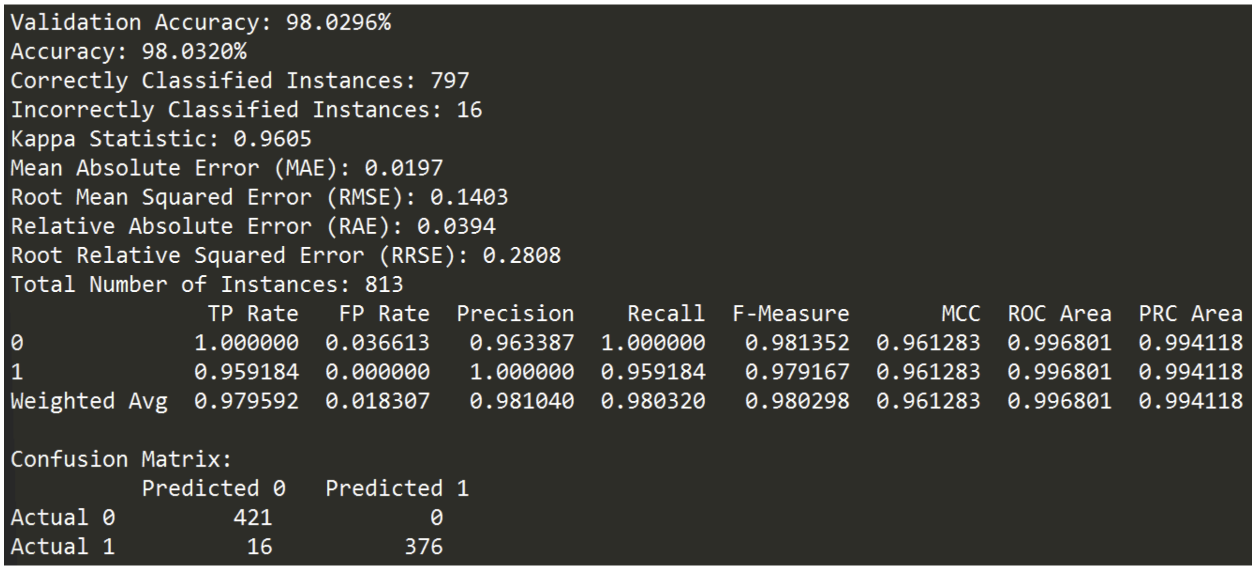
```
Validation Accuracy: 90.7635%
Accuracy: 88.9299%
Correctly Classified Instances: 723
Incorrectly Classified Instances: 90
Kappa Statistic: 0.7785
Mean Absolute Error (MAE): 0.1107
Root Mean Squared Error (RMSE): 0.3327
Relative Absolute Error (RAE): 0.2217
Root Relative Squared Error (RRSE): 0.6659
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             0.878860  0.095355   0.904645  0.878860   0.891566  0.778889  0.924372  0.896485
1             0.900510  0.126238   0.873762  0.900510   0.886935  0.778889  0.924372  0.896485
Weighted Avg  0.889685  0.110796   0.889755  0.889299   0.889333  0.778889  0.924372  0.896485

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0        370           51
Actual 1         39          353
```

Intuition (Non-Weka Attribute Selection) with Support Vector Classifier (SVC)



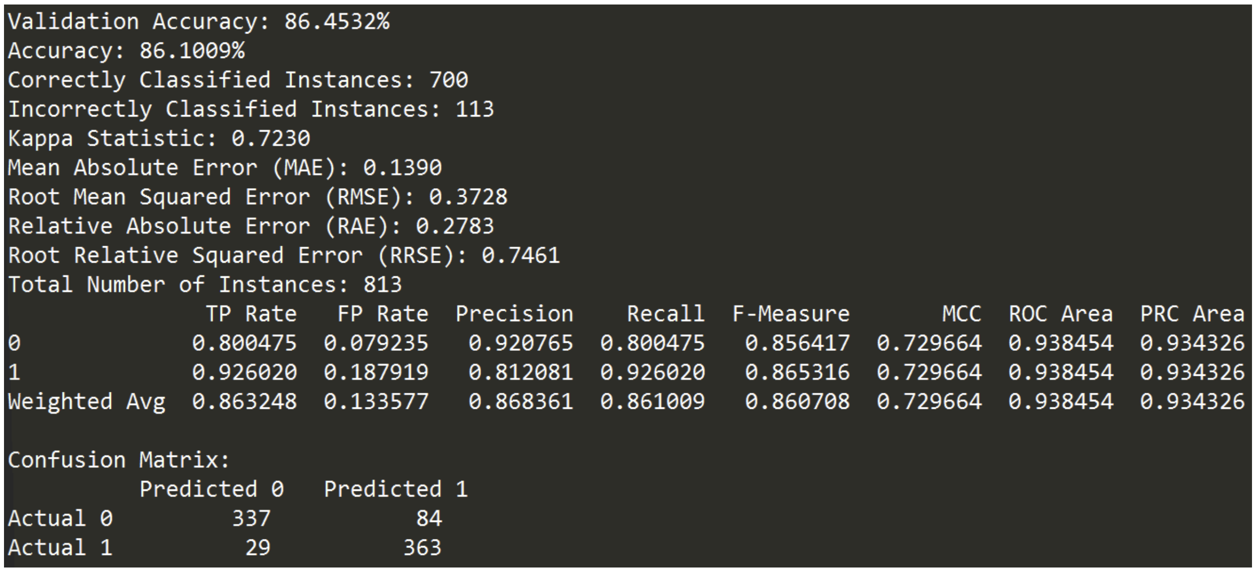
```
Validation Accuracy: 98.3990%
Accuracy: 98.2780%
Correctly Classified Instances: 799
Incorrectly Classified Instances: 14
Kappa Statistic: 0.9655
Mean Absolute Error (MAE): 0.0172
Root Mean Squared Error (RMSE): 0.1312
Relative Absolute Error (RAE): 0.0345
Root Relative Squared Error (RRSE): 0.2626
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             0.997625  0.030023   0.969977  0.997625   0.983607  0.965902  0.998764  0.998744
1             0.966837  0.002632   0.997368  0.966837   0.981865  0.965902  0.998764  0.998744
Weighted Avg  0.982231  0.016327   0.983184  0.982780   0.982767  0.965902  0.998764  0.998744

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0        420            1
Actual 1         13          379
```

Wrapper-Subset with J48 Decision Tree



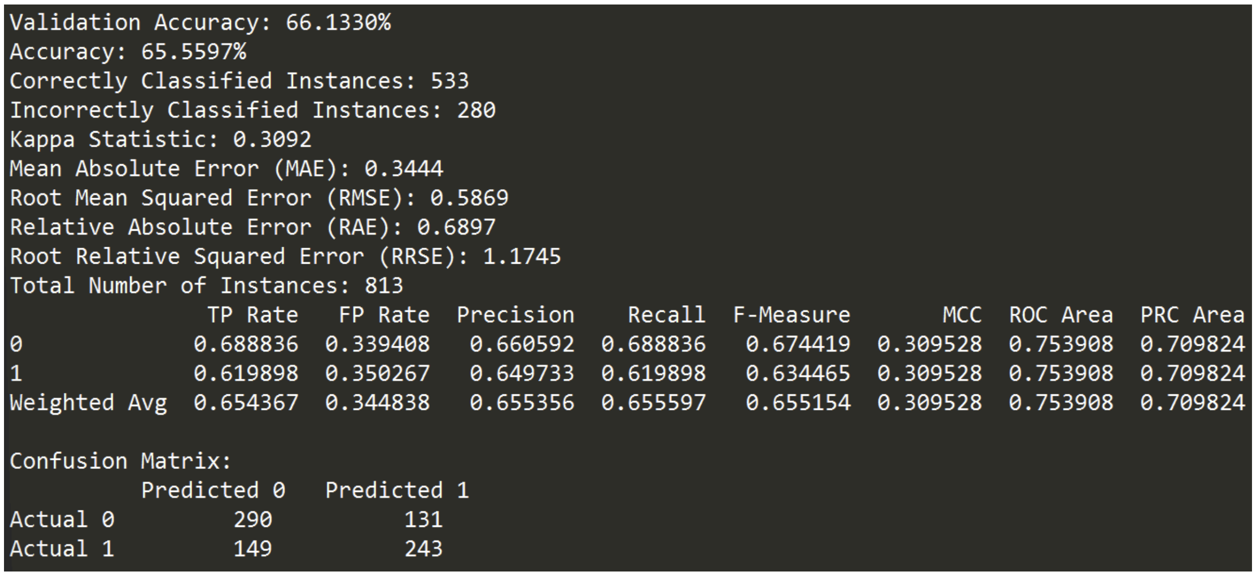
```
Validation Accuracy: 98.0296%
Accuracy: 98.0320%
Correctly Classified Instances: 797
Incorrectly Classified Instances: 16
Kappa Statistic: 0.9605
Mean Absolute Error (MAE): 0.0197
Root Mean Squared Error (RMSE): 0.1403
Relative Absolute Error (RAE): 0.0394
Root Relative Squared Error (RRSE): 0.2808
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             1.000000  0.036613   0.963387  1.000000   0.981352  0.961283  0.996801  0.994118
1             0.959184  0.000000   1.000000  0.959184   0.979167  0.961283  0.996801  0.994118
Weighted Avg  0.979592  0.018307   0.981040  0.980320   0.980298  0.961283  0.996801  0.994118

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0          421             0
Actual 1           16           376
```

Wrapper-Subset with Quadratic Discriminant Analysis (QDA)



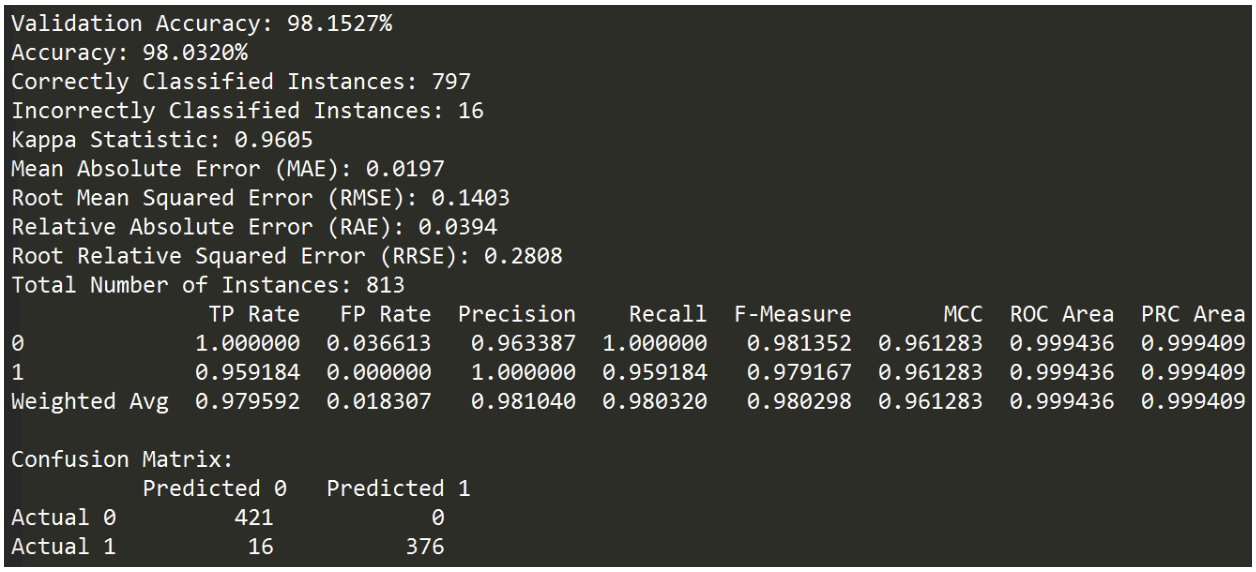
```
Validation Accuracy: 86.4532%
Accuracy: 86.1009%
Correctly Classified Instances: 700
Incorrectly Classified Instances: 113
Kappa Statistic: 0.7230
Mean Absolute Error (MAE): 0.1390
Root Mean Squared Error (RMSE): 0.3728
Relative Absolute Error (RAE): 0.2783
Root Relative Squared Error (RRSE): 0.7461
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             0.800475  0.079235   0.920765  0.800475   0.856417  0.729664  0.938454  0.934326
1             0.926020  0.187919   0.812081  0.926020   0.865316  0.729664  0.938454  0.934326
Weighted Avg  0.863248  0.133577   0.868361  0.861009   0.860708  0.729664  0.938454  0.934326

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0          337            84
Actual 1           29           363
```

Wrapper-Subset with Logistic Regression



```
Validation Accuracy: 66.1330%
Accuracy: 65.5597%
Correctly Classified Instances: 533
Incorrectly Classified Instances: 280
Kappa Statistic: 0.3092
Mean Absolute Error (MAE): 0.3444
Root Mean Squared Error (RMSE): 0.5869
Relative Absolute Error (RAE): 0.6897
Root Relative Squared Error (RRSE): 1.1745
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             0.688836  0.339408   0.660592  0.688836   0.674419  0.309528  0.753908  0.709824
1             0.619898  0.350267   0.649733  0.619898   0.634465  0.309528  0.753908  0.709824
Weighted Avg  0.654367  0.344838   0.655356  0.655597   0.655154  0.309528  0.753908  0.709824

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         290           131
Actual 1         149           243
```

Wrapper-Subset with Support Vector Classifier (SVC)



```
Validation Accuracy: 98.1527%
Accuracy: 98.0320%
Correctly Classified Instances: 797
Incorrectly Classified Instances: 16
Kappa Statistic: 0.9605
Mean Absolute Error (MAE): 0.0197
Root Mean Squared Error (RMSE): 0.1403
Relative Absolute Error (RAE): 0.0394
Root Relative Squared Error (RRSE): 0.2808
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure       MCC  ROC Area  PRC Area
0             1.000000  0.036613   0.963387  1.000000   0.981352  0.961283  0.999436  0.999409
1             0.959184  0.000000   1.000000  0.959184   0.979167  0.961283  0.999436  0.999409
Weighted Avg  0.979592  0.018307   0.981040  0.980320   0.980298  0.961283  0.999436  0.999409

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0         421             0
Actual 1          16           376
```

Part 5.2 – Analysis

We tested four chosen classifier models on the five datasets created by our selected attribute selection algorithms. The following table summarizes the accuracies achieved by each model across our attribute selection techniques:

	Decision Tree	*QDA*	*Logistic Regression*	*SVC*
Intuition Based Selection	94.9569%	93.6039%	88.9299%	98.2780%
CorrelationAttributeEval	98.0320%	97.1710%	95.2030%	99.0160%
GainRatioAttributeEval	98.0320%	94.7109%	94.7109%	98.8930%
InfoGainAttributeEval	95.4490%	96.0640%	89.7909%	97.0480%
WrapperSubsetEval	98.0320%	86.1009%	65.5597%	98.0320%

As can be seen above, the **SVC model with Correlation selection** performs the best, as it has the highest accuracy. Even though no models had an accuracy of 100%, this model still has an accuracy above 99%, with only 8 out of 813 total instances being classified incorrectly.

The area under the receiver operating characteristic (ROC) curve (AUC) looks at the model's ability to obtain a high true positive rate while maintaining a low false positive rate and is another measure that can be used to evaluate the performance of the model. The following table summarizes the ROC areas for each model across our attribute selection techniques:

	Decision Tree	*QDA*	*Logistic Regression*	*SVC*
Intuition Based Selection	0.980328	0.962359	0.924372	0.998764
CorrelationAttributeEval	0.990184	0.979822	0.980125	0.999727
GainRatioAttributeEval	0.990184	0.969218	0.957814	0.999952
InfoGainAttributeEval	0.979798	0.974835	0.947125	0.997225
WrapperSubsetEval	0.996801	0.938454	0.753908	0.999436

Unlike with accuracy, the **SVC model with Gain Ratio selection** performs the best, beating SVC with Correlation by 0.0002.

Part 6 – Conclusion/How to Reproduce Our Model

By testing multiple classifier and attribute selection technique pairs, we ensured a comprehensive and thorough evaluation of the models in varying configurations. We conducted our evaluation on several key metrics of the model, including precision, recall, ROC–AUC, and confusion matrices. As stated above, the Support Vector Classifier (SVC) with correlation attribute selection yielded the best results across multiple runs for this project. We successfully developed and evaluated multiple predictive classification models to determine the edibility of mushrooms, and the SVC classifier demonstrated strong performance in conjunction with our attribute selection strategy. However, after analyzing the attributes chosen by our attribute selection, it is clear that there is room for improvement. Specifically, techniques like correlation-based selection could overlook relevant attributes with lower individual correlations but important combined correlations with other attributes. Thus, exploring alternative attribute selection strategies, such as Random Forest or Gradient Boosting, could further improve model performance.

Steps to Reproduce Our Most Effective Model: SVC with Correlation Attribute Selection

1. Download the mushrooms.csv from the official UCI repository
2. Place the mushrooms.csv file in a new directory named "Q1 Project"
3. Open Weka and load the mushrooms.csv file
4. Go to the "Select Attributes" tab and ensure the class variable is set to **class**
5. Select CorrelationAttributeEval as Attribute Evaluator and Ranker as Search Method
6. Begin the attribute selection and take inventory of features in the output that have a ratio not greater than or equal to 0.25 (our cutoff)
7. Open a Python IDE (Jetbrains DataSpell was used for this paper) with CWD as "Q1 Project"
   7a. If the Python environment is not in the "Q1 Project" directory, type
   ```
   cd [Path of Q1 Project directory]
   ```
8. Copy the *split.py*, *model.py*, and *load.py* programs from the paper's source[2] to the directory
9. Install required dependencies using the shell terminal:
   ```
   pip install numpy pandas scikit-learn
   ```
10. Open the *split.py* program in the IDE
11. Configure the `attributes_to_remove` list in the *split.py* program to contain the names of the features in the earlier inventory as strings
12. Change the contents of the `folder_name` string to "correlation"
13. Run the *split.py* program to remove irrelevant attributes, encode data, perform KNN imputation, and test-train-split into separate directories

[2] https://drive.google.com/drive/folders/1hnUSwRu4W7yZf5P-Dw0NbJQ8kPvf-cbH

This results in a file tree structure similar to the following:

```
Q1 Project/

|-- split.py
|-- model.py
|-- load.py
|-- dataset/
    |-- correlation/
        |-- x_train.csv
        |-- x_val.csv
        |-- x_test.csv
        |-- y_train.csv
        |-- y_val.csv
        |-- y_test.csv
```

14. Open the *model.py* program in the IDE
15. Run the *model.py* program to derive metrics for Correlation Attribute Selection with SVC and save the model
16. Model can be found here: **Q1 Project/models/svc_correlation.pkl**
17. Verify the stored model as functional by running *load.py* and observing prediction accuracy

Part 7 – Team Members and Tasks Performed

Finding the Dataset: Anieesh
Building the Proposal: Anieesh and Pranav
Preprocessing Initial Attempt: Anieesh
Preprocessing & Project Update: Anieesh and Pranav
Non-Weka Attribute Selection Algorithm: Pranav
Attribute Selection Algorithms and Classifiers: Anieesh and Pranav
Results Output: Anieesh
Results Analysis: Pranav
Building Final Report: Anieesh and Pranav

Part 8 – Sources

References

DecisionTreeClassifier. (n.d.). scikit-learn. Retrieved October 22, 2024, from

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

1.4. Support Vector Machines. (n.d.). scikit-learn. Retrieved October 22, 2024, from

https://scikit-learn.org/1.5/modules/svm.html

LogisticRegression. (n.d.). scikit-learn. Retrieved October 22, 2024, from https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html

Package weka.attributeSelection. (n.d.). Weka Documentation. Retrieved October 22, 2024, from https://weka.sourceforge.io/doc.dev/weka/attributeSelection/package-summary.html

QuadraticDiscriminantAnalysis. (n.d.). scikit-learn. Retrieved October 22, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis.html