COMP30120 - Assignment 2: Feature Selection in Weka

Felipe Guth **Student Id: 14210231**

1. Apply one filter and one wrapper feature selection technique from those available in Weka and report the feature subsets that they select.

Dataset 1: Bikes

Information Gain Filter:

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 12 Level):
        Information Gain Ranking Filter

Ranked attributes:
 0.40388   8 temp
 0.40192   9 atemp
 0.38533   3 mnth
 0.25917   1 season
 0.07112   7 weathersit
 0.05061  10 hum
 0.04754  11 windspeed
 0.03367   5 weekday
 0.03362   6 workingday
 0.01469   2 yr
 0.0044    4 holiday

Selected attributes: 8,9,3,1,7,10,11,5,6,2,4 : 11
```

Backward Selection Wrapper

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: 1,2,3,4,5,6,7,8,9,10,11,
        Search direction: backward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 75
        Merit of best subset found:    0.904

Attribute Subset Evaluator (supervised, Class (nominal): 12 Level):
        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: 2,6,7,9,11 : 5
                     yr
                     workingday
                     weathersit
                     atemp
                     windspeed
```

Dataset 2: Basketball

Information Gain Filter

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 30 Result):
        Information Gain Ranking Filter

Ranked attributes:
 0.97095    1 Opponent
 0.41997   23 Field_Goals_Made
 0.38534    7 Opp_Free_Throws_Made
 0.00485   11 Home
 0          9 Free_Throw_Pct
 0         10 Fouls
 0         12 Opp_Free_Throw_Pct
 0          8 Opp_Field_Goal_Pct
 0          6 Opp_Steals
 0         14 Turnovers
 0          5 Def_Rebounds
 0          2 Opp_Free_Throws_Att
 0          3 Opp_Fouls
 0          4 Opp_Field_Goals_Made
 0         13 Free_Throws_Made
 0         29 Opp_Def_Rebounds
 0         28 Opp_Total_Rebounds
 0         22 3Pt_Field_Goals_Att
 0         25 Opp_Turnovers
 0         26 Opp_3Pt_Field_Goals_Att
 0         27 Opp_3Pt_Field_Goal_Pct
 0         24 Opp_3Pt_Field_Goals_Made
 0         21 Off_Rebounds
 0         16 Opp_Off_Rebounds
 0         20 Total_Rebounds
 0         17 Field_Goals_Att
 0         18 3Pt_Field_Goal_Pct
 0         19 Steals
 0         15 Opp_Blocks

Selected attributes: 1,23,7,11,9,10,12,8,6,14,5,2,3,4,13,29,28,22,25,26,27,24,21,16,20,17,18,19,15 : 29
```

Backward Selection Wrapper

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,
        Search direction: backward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 857
        Merit of best subset found:    0.9

Attribute Subset Evaluator (supervised, Class (nominal): 30 Result):
        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: 23,28 : 2
                     Field_Goals_Made
                     Opp_Total_Rebounds
```

2. Quantify and comment on the overlap between these alternative feature subsets.

In the bikes dataset, the information gain filter ranked the following variables, given a threshold of 0.05, temp (0.40), atemp (0.40), mnth (0.38), season (0.25), weathersit (0.07) and hum (0.05). The backward selection wrapper has selected the variables yr, workingday, weathersit, atemp and windspeed. The overlapped variables in the methods were atemp and weathersit. This two variables have the potential to obtain good results in conjunction with classifiers, given that they were selected in two feature selection methods.

In the basketball dataset, the information gain filter ranked the following variables with rank > 0, Opponent (0.97) Field_Goals_Made (0.41), Opp_Free_Throws_Made (0.38) and Home (0.004). The backward selection wrapper has selected the variables Field_Goals_Made and Opp_Total_Rebounds. The overlapping variable was Field_Goals_Made. This shows how the number of variables can be effectively reduced in this dataset using just the more representative ones.

3. Discuss the performance of these feature selection techniques when combined with two different classifiers of your choice available in Weka (i.e. there will be four experimental combinations for each dataset).

In the bikes dataset the Information gain filter selection variables (temp, atemp, mnth, season, weathersit and hum) were combined with the KNN and Naïve Bayes classifiers. In the KNN, the correct number of classified instances was equal to 356 (81.27%) and incorrect classified instances were equal to 82 (18.72%). The mean absolute error was equal to 0.18 and the relative absolute error equal to 41.26%. In the Naïve Bayes classifier, the correct number of classified instances was equal to 370 (84.47%) and incorrect classified instances were equal to 68 (15.52%). The mean absolute error was equal to 0.16 and the relative absolute error equal to 35.56%. Using the the backward selection wrapper variables (yr, workingday, weathersit, atemp and windspeed) also in conjunction to the KNN and Naïve Bayes classifiers the results were improved. In the KNN the number of correctly classified instances was equal to 383 (87.44%) and incorrect classified instances were equal to 55 (12.55%). The mean absolute error was equal to 0.12 and the relative absolute error equal to 27.85%. In the Naïve Bayes classifier, the correct number of classified instances was equal to 395 (90.18%) and incorrect classified instances were equal

to 43 (9.81%). The mean absolute error was equal to 0.18 and the relative absolute error equal to 39.73%.

In the basketball dataset the Information Gain variables (Opponent, Field_Goals_Made, Opp_Free_Throws_Made and Home were combined with Decision Trees (trees.J48) and Naïve Bayes Classifiers. Both, Decision Tree and Naïve Bayes classifiers, correctly classified 14 instances (70%), and incorrectly classified 6 instances (30%). The mean absolute error in the Decision Tree was equal to 0.34 whether in the Naïve Bayes was equal to 0.28. The relative absolute error was equal to 70.72% in the Decision Tree and 58.05% in the Naïve Bayes classifier. Using the the Backward Selection Wrapper variables (Field_Goals_Made and Opp_Total_Rebounds), the Decision Tree classifier has classified correctly 18 instances (90%) and incorrectly 2 instances (10%), whether the Naïve Bayes has classified correctly 13 instances (65%) and incorrectly 7 instances (35%). The mean absolute error of the Decision Tree was equal to 0.13 and relative absolute error equal to 28.06 %. The mean absolute error of the Naïve Bayes classifier was equal to 0.36 and the relative absolute error equal to 74.76%.

4. When Information Gain is used as a filter in feature selection some features will score 0, suggesting that they have no predictive power and can be ignored. Yet in the datasets for this assignment some features that have an Information Gain score of 0 are selected by other feature selection methods. See if you can find examples of this and discuss why this might occur.

In the bike dataset the feature yr was ranked as one of the least discriminative variables by the information gain method, however, it was selected by the wrapper. Likewise, in the basketball dataset the Opp_Total_Rebounds had a rank score equal to 0 but was selected by the wrapper method. The information gain consider a variable independently, it does not takes into account relationships between variables, measuring just the value of discrimination of a given variable against the dataset. The wrapper method, evaluate subsets of variables which allows to identify relationship between variables. This is the reason why some variables are selected in the wrapper method and are not selected in the information gain method, given to the strong relationship with another variable(s) in the discrimination of the dataset.