Practical Application 2

Machine Learning

Yuxiao Xiong

y.xiong@alumnos.upm.es - a18m033

1 Introduction

In this assignment, a Dataset about voice[1] will be applied to the three probabilistic supervised classification algorithms (including Logistic Regression, Naive Bayes and Discriminant Analysis) and some Metaclassifiers in order to identify a voice as either male or female. The objective is achieved by using four kinds of analyses (with all original variables, with a univariate filter feature subset selection, with a multivariate filter feature subset selection and with a multivariate Wrapper feature subset selection).

Consequently, we will extract some results and discuss them to analyze the performance of these algorithms if it is possible.

2 Problem Description

The Dataset used in this work consists of 3,168 recorded voice samples, collected from male and female speakers. The information of these samples are represented by 20 acoustic properties and 1 target variable. The 20 properties are numeric and continuous attributes. More specifically, these variables are:

- `meanfreq`: mean frequency (in kHz).
- `sd`: standard deviation of frequency.
- `median`: median frequency (in kHz).
- `Q25`: first quantile (in kHz).
- `Q75`: third quantile (in kHz).
- `IQR`: interquantile range (in kHz).
- `skew`: skewness.
- `kurt`: kurtosis.
- `sp. ent`: spectral entropy.
- `sfm`: spectral flatness.
- `mode`: mode frequency.
- `centroid`: frequency centroid.
- `peakf`: peak frequency (frequency with highest energy).
- `meanfun`: average of fundamental frequency measured across acoustic signal.
- `minfun`: minimum fundamental frequency measured across acoustic signal.
- `maxfun`: maximum fundamental frequency measured across acoustic signal.
- `meandom`: average of dominant frequency measured across acoustic signal.
- `mindom`: minimum of dominant frequency measured across acoustic signal.
- `maxdom`: maximum of dominant frequency measured across acoustic signal
- `dfrange`: range of dominant frequency measured across acoustic signal.
- `modindx`: adjacent measurements of fundamental frqeuncies divided by the frequency range.
- `label`: target variable, its value is either male or female.

3 Methodology

Before we move further on, it is necessary to pre-process the input data. While the data is very clean and does not contain any empty values, some features have bigger values than others, for instance, `kurt` has a mean value of 36.6 while the mean value of `IQR` is only 0.08. Therefore, we need to normalize numeric variables, using Python, Pandas and Sklearn, to avoid problems in future analysis.

3.1 Classifiers & Metaclassifiers

As we mentioned in the introduction, three classifiers and some Metaclassifiers will be used in this assignment.

First of all, the **Logistic Regression** model adapts the linear regression formula to act as a classifier which predicts a dependent data variable by analyzing the relationship between the existing independent variables.[2] The classifier is specified as **RLog** in Weka.

Secondly, the **Bayesian** classifier chosen in this assignment is called "Tree Augmented Naive Bayes", also known as the **TAN** model. It relaxes the naive Bayes attribute independence assumption by employing a tree structure, in which each predictor variable only depends on the root node and one other attribute.[3] The classification is performed by building a maximum weighted spanning tree. The TAN model is used for discrete variables while all the variables of the dataset are numeric and continuous, therefore, the variables need a discretization with the filter *Discretize* of Weka. In order to use the TAN model in Weka, we choose **TAN** as the search algorithm in the **BayesNet** classifier configuration.

In addition, the **Linear Discriminant Analysis**, also known as the Fisher's Linear Discriminant Analysis, is a dimensionality reduction technique that can be used as a classifier and it attempts to maximize the seperation between classes of the dataset. In Weka, this classifier is shown as **FLDA**.[4]

Finally, Metaclassifiers such as, **Bagging with Multilayer Perceptron**, **Boosting with RLog**, **Random Forest** and **Fusion with Majority vote** (including k-NN, RIPPER, Support Vector Machine, C4.5 tree and Logistic Regression) will involve in this project.

While applying all these algorithms, Cross-Validation with 10 folds is used as testing. With this kind of testing, Weka first takes 100 labeled data and produces 10 equal sized sets. Each set is divided into two groups: 90 labeled data are used for training and 10 labeled data are used for testing. Then Weka produces a classifier with an algorithm from 90 labeled data and applies that on the 10 testing data for set 1. Weka repeats these steps for the rest 9 sets and produce 9 more classifies. Finally, Weka averages the performance of the 10 classifiers produced from 10 equal sized sets. [4]

3.2 Feature Subset Selection

Among the four analyses, despite of the first one with all original variables, the rest need to go through the Feature Subset Selection process where irrelevant and redundant variables will be removed as many as possible.

For univariate filtering, **Gain Ratio** is used to evaluate the attributes by measuring the gain ratio respecting to the class.[4] The threshold assigned is 0.1, so the variables with a ratio smaller than 0.1 will not be considered. The multivariate filtering consists of the Correlation-based feature selection (**CFS**) that evaluates by considering the individual predictive ability of each feature along with the degree of redundancy between them. [4]

As for Wrapper approaches, **Best First** is the method that helps the selection process. The classifiers for each Wrapper approach have the same parameters as the ones used for all variables analysis, so for the TAN model, the data also needs to be discretized before the Wrapper process. Moreover, the evaluation metrics is accuracy for discrete classes and RMSE (Root Mean Square Error)) for numeric classes.

However, in this practical application, the Metaclassifiers will be applied only to the univariate subset.

4 Results

After executing the classifiers menstioned above, we have obtained the following results, where Table 1 shows the results of applying Feature Subset Selection, and Table 2 describes the accuracy of three classifiers on different subsets respectively.

	Univariate	Multivariate	Wrapper		
			RLog	TAN	FLDA
meanfreq					
sd	✓		✓	✓	✓
median			✓		
Q25	✓				
Q75					
IQR	✓	✓	✓	✓	✓
skew					
kurt			✓		✓
sp.ent	✓		✓		
sfm	✓		✓	✓	✓
mode	✓		✓		✓
centroid					
meanfun	✓	✓	✓	✓	✓
minfun			✓		✓
maxfun				✓	✓
meandom					✓
mindom			✓	✓	
maxdom					✓
dfrange				✓	
modindx			✓		

Table 1: Attributes selected by univariate and multivariate filters and 3 Wrappers

	All variables	Univariate	Multivariate	Wrapper
RLog	97.096 %	97.033%	96.559%	97.285%
TAN	97.222%	97.790%	96.654%	98.043%
FLDA	96.843%	96.181%	96.117%	96.970%

Table 2: Accuracy of the three classifiers on each subset

The logistic regression model returns the coefficients of each attribute which is illustrated in Figure 1.

Variable	Class male
meanfreq	3.678
sd	28.631
median	-8.4765
Q25	-20.9992
Q75	18.9092
IQR	32.9556
skew	0.1274
kurt	-0.0073
sp.ent	41.4255
sfm	-12.0302
mode	3.2437
centroid	3.678
meanfun	-166.1935
minfun	37.5739
maxfun	-1.3221
meandom	0.0701
mindom	-0.5341
maxdom	-0.0024
dfrange	-0.0022
modindx	-3.2592
Intercept	-16.0721

(a) β - All variables

Coefficients...

Variable	Class male
meanfun	-29.3152
IQR	14.1977
Q25	-0.6112
sp.ent	11.1021
sd	-1.0666
sfm	-9.8514
mode	1.7779
Intercept	6.032

(b) β - Univariate

Variable	Class male
IQR	36.4373
meanfun	-169.5078
Intercept	20.7174

(c) β - Multivariate

Variable	Class male
sd	29.2995
median	-5.5751
IQR	52.1173
kurt	-0.0039
sp.ent	38.2678
sfm	-11.8108
mode	2.978
meanfun	-166.4702
minfun	37.0013
mindom	-0.1822
modindx	-3.1219
Intercept	-12.8512

(d) β - Wrapper

Figure 1: Coefficients of β in Logistic Regression

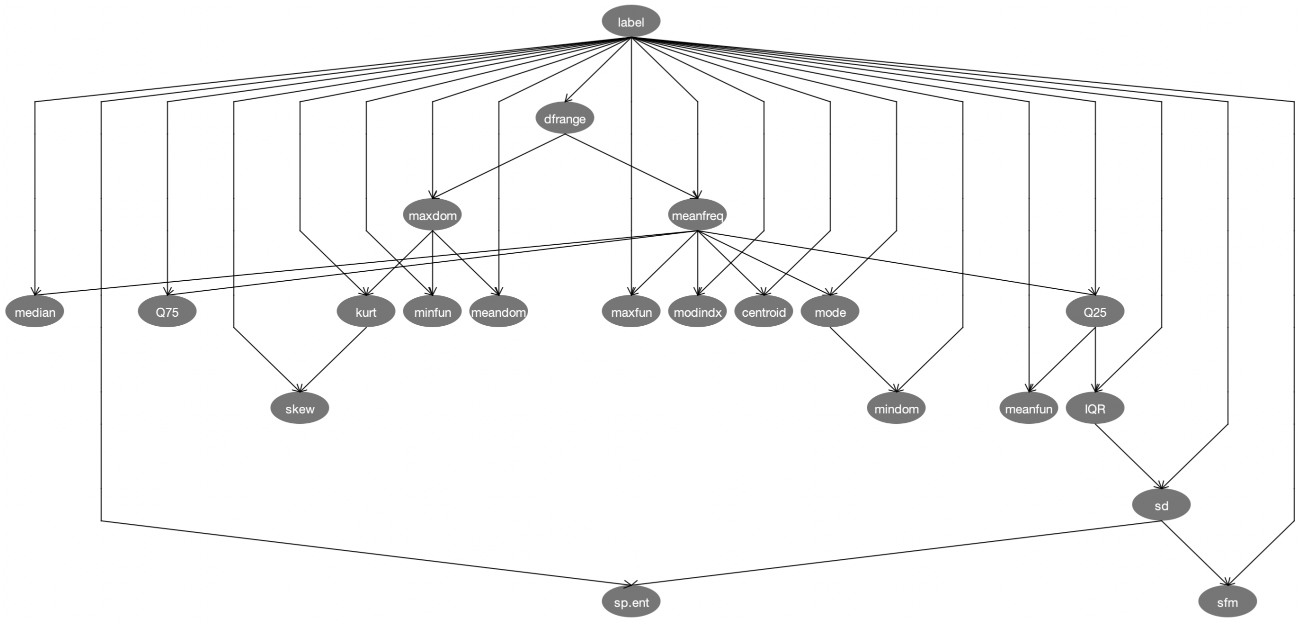
Regarding to the TAN model, we can see the graphs of the TAN structure in Figure 2, where every attribute is a node and except `label` which is the parent node, every node could depend only on the parent node, or also on another node. As for the linear discriminant analysis, Figure 3 shows the weights of each attribute to determine the final class.

Apart from the three probabilistic classifiers, we also obtained some results of the Metaclassifiers mentioned in the previous section. First of all, we have the accuracy of these Metaclassifiers and the auxiliary classifiers involved which is presented in Table 3. Additionally, Figure 4 illustrates the attribute importance based on average impurity decrease and the number of nodes using that attribute in Random Forest.

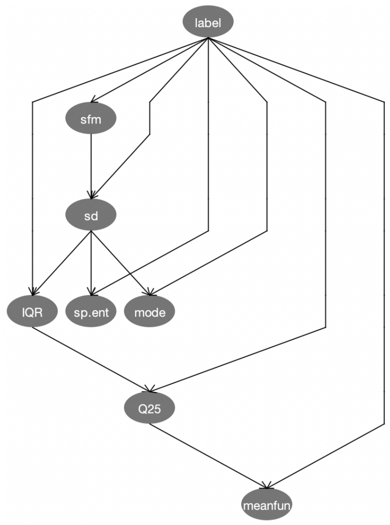
The confusion matrices of the classifiers on wrapper subsets are shown in Figure 5.

5 Discussion

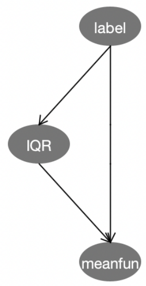
Observing Table 1, some ideas about the importance of variables can be deduced:



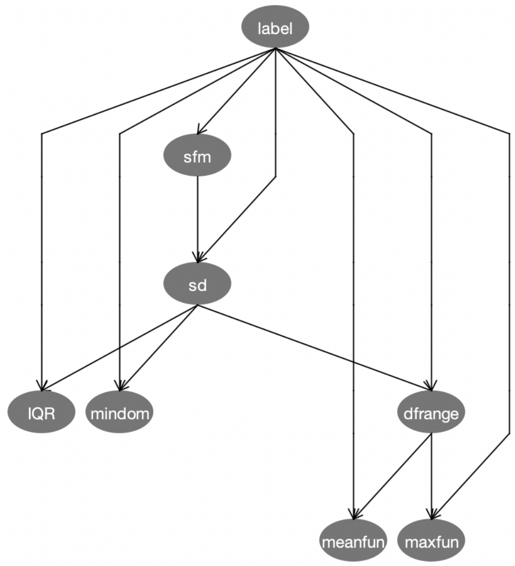
(a) TAN - All variables



(b) TAN - Univariate



(c) TAN - Multivariate



(d) TAN - Wrapper

Figure 2: TAN structures

1. `IQR` and `meanfun` are the most relevant attributes since they are chosen by all filters or Wrappers.
2. `sd`, `sfm` and `mode` are less relevant but they are still selected by most of the filters or wrappers.
3. `meanfreq`, `Q75`, `skew`, `centroid` are the most redundant variables as none of the filters of wrappers choose them.

From Table 2 and 3 we can conclude that the classifiers and Metaclassifiers used for the Dataset have an outstanding performance, with an overall accuracy more than 96%. In general, wrapper subsets have better results in accuracy which may suggest that, in some way, variables chosen by wrappers are more relevant than others. Moreover, Table 2 shows that the linear discriminant analysis performs worse, with a slight difference, than other classifiers.

```
Threshold: -0.08785505180791896

Weights:

meanfreq:       -0.0037473995992913876
sd:      0.23910104018631942
median:         -0.07368148904683382
Q25:     0.02962482392271831
Q75:     0.1327895256989009
IQR:     0.1996395650583844
skew:   -0.0020070645619731774
kurt:    2.8795717451011457E-5
sp.ent:         -0.03651594742276666
sfm:    -0.024217095295127753
mode:    0.030897229671869076
centroid:       -0.003747399547950062
meanfun:        -0.7765744572985018
minfun:          0.18528984816988855
maxfun:          0.04521267454759539
meandom:        -0.0041544975598889365
mindom:          0.28634701829331305
maxdom:         -0.2780398540192442
dfrange:         0.2781433304757384
modindx:         0.0015170606202304468
```

(a) FLDA - All variables

```
Threshold: -0.09800704021516991

Weights:

meanfun:        -0.8131826856596647
IQR:     0.514006256479134
Q25:     0.16366315221002983
sp.ent:         0.12939340252022644
sd:     -0.014369557324120878
sfm:    -0.1590754578264192
mode:    0.07410615747818013
```

(b) FLDA - Univariate

```
Threshold: -0.11700157823314802

Weights:

IQR:     0.2515917943197355
meanfun:        -0.9678334407484461
```

(c) FLDA - Multivariate

```
Threshold: -0.0804493211800813

Weights:

sd:      0.23638337956404912
IQR:     0.27662078286116015
kurt:   -3.6410261160389544E-5
sfm:    -0.03805776906220941
mode:    0.04560814364693591
meanfun:        -0.9008294786424694
minfun:          0.2217927459996069
maxfun:          0.05810055629900191
meandom:        -0.003725926543678035
maxdom:          1.1634589184096956E-4
```

(d) FLDA - Wrapper

Figure 3: The weights of attributes in Linear Discriminant Analysis

```
0.42 (  1364)  meanfun
0.34 (   925)  Q25
0.34 (  1160)  IQR
0.33 (   675)  sp.ent
0.29 (   535)  mode
0.27 (   765)  sd
0.27 (   732)  sfm
```

Figure 4: Attribute importance of Random Forest on Univariate subset

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1546   38 |    a = male
   48 1536 |    b = female
```

(a) RLog

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1560   24 |    a = male
   38 1546 |    b = female
```

(b) TAN

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1550   34 |    a = male
   62 1522 |    b = female
```

(c) FLDA

Figure 5: The confusion matrix of the three classifiers on wrapper subsets

		Bagging with NN	Boosting with RLog	Random Forest	Vote
Auxiliary classifiers	KNN	-	-	-	98.296%
	RIPPER	-	-	-	97.096%
	SVM	-	-	-	97.191%
	NN	98.854%	-	-	-
	C4.5	-	-	-	97.191%
	RLog	-	97.033%	-	97.033%
Metaclassifier		**98.854%**	**97.032%**	**97.822%**	**97.790%**

Table 3: The Accuracy of the Metaclassifiers and auxiliary classifiers involved

5.1 Logistic Regression

Giving $P(C=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+...+\beta_k x_k)}}$, in this project, ***C=1*** is the `male` class and $\boldsymbol{x}$ is the vector of variables. If $P(C=1|x) > threshold$, the prediction will be `male` and if $P(C=1|x) < threshold$, it will be `female`. From the *logit* form of the Logistic model, $logit(P(C=\texttt{male}|x)) = \beta_0+\beta_1 x_1+...+\beta_k x_k$, we can refer that if β_k is positive, the bigger x_j is, the bigger $P(C=\texttt{male}|x_j)$ will be, thus x_j is a predictor for `male` class. Similarly, if if β_k is negative, x_j will be the predictor for `female` class. In this case, as we can observe from Figure 1a that, great values of variables like `sp.ent`, `minfun`, `IQR`, `sd`, `Q75` and `meanfreq` are contributors to predict the `male` class, and on the contrary, variables like `meanfun`, `Q25`, `sfm` and `median` contribute to predict the `female` class.

Regarding to different filter or wrapper subsets, we can see that the two attributes `IQR` and `meanfun`, that appear in all subsets, maintain their feature of predicting `male` and `female` class, respectively.

However, the value of these coefficients changes significantly comparing the univariate subset with others. The reason for this is that the some attributes are highly correlated and we may need to remove them. Figure 6 draws the heat map of the correlation of the variables, which means we need to remove the predictors with correlation value that are 1 or -1 to get more stable coefficients. Concretely, we will remove the predictors with absolute correlation value bigger than 0.95, and they are `kurt`, `centroid`, `dfrange`.

After executing the logistic model for the datasets without these 3 variables, we obtain the following results:

1. For the entire dataset without the 3 correlated attributes, the accuracy increases slightly from 97.096% to 97.123%.
2. The univariate and multivariate subsets are still the same and so does the accuracy.
3. The wrapper of RLog now selects `skew` and does not chooses `kurt` or `sp.ent`. The accuracy increases slightly from 97.285% to 97.317%.
4. The value of the coefficients now has smaller changes between filter and wrapper subsets (Figure 7).

5.2 Tree-Augmented Naive Bayes

This model generates the graph of dependencies for each subset (Figure 2). As we can see that all nodes (except the parent) have one directed edge coming from `label` and the other coming from another attribute. These edges are added according to the conditional mutual information quantities arranged

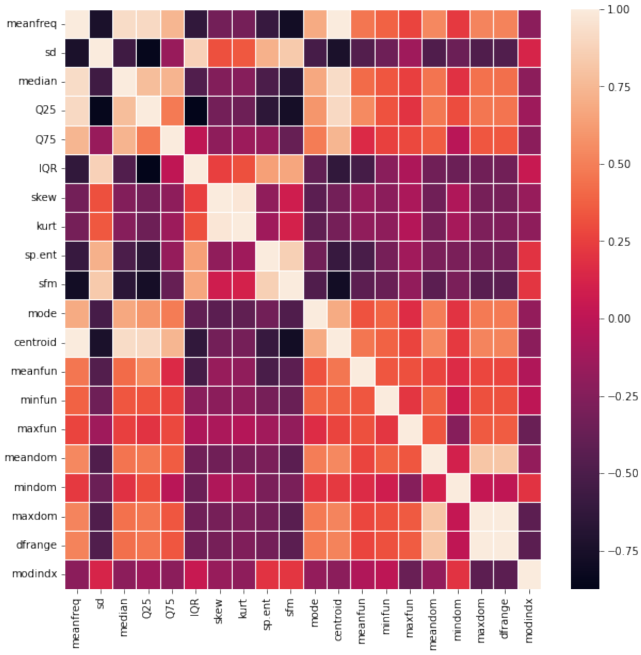


Figure 6: The heat map of the correlation of the variables

Coefficients...

Variable	Class male
meanfreq	11.0659
sd	36.4529
median	-9.0799
Q25	-20.6678
Q75	16.2128
IQR	31.6966
skew	-0.1304
sp.ent	35.9921
sfm	-11.4029
mode	2.6652
meanfun	-165.529
minfun	36.5214
maxfun	-0.1875
meandom	-0.0211
mindom	-0.0418
maxdom	-0.0015
modindx	-3.0592
Intercept	-11.5077

(a) β - All variables

Coefficients...

Variable	Class male
meanfun	-161.0093
IQR	59.7376
Q25	-2.4732
sp.ent	45.6223
sd	-11.0066
sfm	-12.2216
mode	6.3495
Intercept	-18.9366

(b) β - Univariate

Coefficients...

Variable	Class male
sd	3.0846
median	-3.4756
IQR	57.3557
skew	-0.2314
sfm	-4.4542
mode	1.9
meanfun	-171.7703
minfun	32.3452
mindom	-0.2074
modindx	-3.0193
Intercept	21.1044

(c) β - Wrapper

Figure 7: The coefficients of β in Logistic Regression after removing correlated predictors. (Multivariate does not change, see Figure 1c)

in ascending order. Therefore, comparing the four graphs, we can deduced some strong relationships between certain attributes from Figure 2. For instance, `sd` is strongly related with `sfm` and `IQR` because they are connected in all subsets except the multivariate one.

5.3 Linear Discriminant Analysis

As we use a binary-classification dataset, the decision boundary is a hyperplane $w^T(x - x_0) = 0$. In this case, Weka chooses `male` class as the prediction label, so if $w^T(x - x_0) > 0$, the output class will be `male`. w^T represents the vector of weights of attributes (Figure 3), $w^T * x_0$ is the threshold, thus the LDA classifier is actually evaluating if $w^T * x > threshold$ or not. Since the decision boundary is a hyperplane in 2D and not necessarily perpendicular to the line separating the means, some values of the vector weights w^T alter, such as `IQR` that only takes half of the value in multivariate subset than in the univariate.

5.4 Metaclassifiers

From Table 3 we can see that Bagging and Boosting do not improve the performance of the classifiers comparing to the original ones because they combine models of same type in all iterations.

Furthermore, Figure 4 demonstrates the importance of each attribute based on the average impurity decrease, `meanfun`, `Q25` and `IQR` have more predictive power than others.

Finally, for the Majority Vote classifier, according to Table 3 it improves the accuracy comparing to four of the five auxiliary classifiers (the exception is KNN).

5.5 Confusion Matrices

An interesting discovery about the confusion matrices is that, although the classifiers and meta-classifiers performs similarly on predicting both classes[1], generally they get better results on `male` class than `female` (Figure 5).

6 Conclusion

In this practical assignment, we have tried to predict if a voice belongs to a male or a female, based on a Dataset of 20 attributes and around 3,000 observations. We have accomplished the goal by using three probabilistic supervised classification algorithms in case of four analyses about feture subset selection and some Metaclassifiers.

They all have an excellent performance regarding to the accuracy but meanwhile the dataset generates some problems. On the one hand, the variables are numeric and continuous. This feature produces problem for Tree-Augmented Naive Bayes because the model is used to identify discrete variables. Although we pre-process the original dataset with discretization, the values of one bin could belong to the consecutive bin due to the similarity. On the other hand, as we comment in the previous section, various attributes are strongly correlated and cause the problem of instability of the coefficients in Logisitic Regression model. Despite of the fact that we remove the predictors with correlation value bigger than 0.95, the work can be improved if we try more numbers for the correlation value to find the balance point between stable β values and correlation.

[1]The classes are balanced and this is why we do not use F-score to evaluate the performance of the classifiers, the F-score takes exactly the same value as accuracy.

References

[1] Kory Becker. *Gender Recognition by Voice*. data retrieved from Kaggle, https://www.kaggle.com/datasets/primaryobjects/voicegender. 2016.

[2] Pedro Larrañaga and Concha Bielza. *Logistic Regression*. 2022.

[3] Fei Zheng and Geoffrey I. Webb. "Tree Augmented Naive Bayes". In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2010, pp. 990–991. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_850. URL: https://doi.org/10.1007/978-0-387-30164-8_850.

[4] Eibe Frank, Mark A. Hall, and Ian H. Witten. *Data Mining: Practical Machine Learning Tools and Techniques*. Online Appendix - The Weka Workbench. 2016. URL: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf.