|  |
| --- |
| **T.C.**  **ÇUKUROVA ÜNİVERSİTESİ**  **MÜHENDİSLİK-MİMARLIK FAKÜLTESİ**  **Data Mining Report**  **2018556059**  **Cahit Barkın Özer**    Açıklama: http://muh.zirve.edu.tr/pictures/CS_002.jpg  **ADANA** |

# Abstract

In this article, two distinct datasets were chosen based on data size, number of features, and missing values. These datasets were trained using four distinct types of algorithms, and the accuracy, model build times, and causes of the results were examined.

# TABLE OF CONTENTS

Chapter Page

[**ABSTRACT 2**](file:///C:\Users\barkin\Downloads\Table%20of%20Contents%20Template%20Word%2004.doc#__RefHeading___Toc208803836)

[**TABLE OF CONTENTS 3**](file:///C:\Users\barkin\Downloads\Table%20of%20Contents%20Template%20Word%2004.doc#__RefHeading___Toc208803838)

[**INTRODUCTION 4**](file:///C:\Users\barkin\Downloads\Table%20of%20Contents%20Template%20Word%2004.doc#__RefHeading___Toc208803841)

[**CHAPTER I: Dataset Descriptions .5**](file:///C:\Users\barkin\Downloads\Table%20of%20Contents%20Template%20Word%2004.doc#__RefHeading___Toc208803842)

[Mushroom Dataset Description 5](file:///C:\Users\barkin\Downloads\Table%20of%20Contents%20Template%20Word%2004.doc#__RefHeading___Toc208803843)

[Students Performance Dataset Description 8](file:///C:\Users\barkin\Downloads\Table%20of%20Contents%20Template%20Word%2004.doc#__RefHeading___Toc208803843)

[**CHAPTER II: Analysis Results 11**](file:///C:\Users\barkin\Downloads\Table%20of%20Contents%20Template%20Word%2004.doc#__RefHeading___Toc208803849)

[Mushroom Dataset Analysis Results 11](file:///C:\Users\barkin\Downloads\Table%20of%20Contents%20Template%20Word%2004.doc#__RefHeading___Toc208803843)

[Students Performance Dataset Analysis Results 15](file:///C:\Users\barkin\Downloads\Table%20of%20Contents%20Template%20Word%2004.doc#__RefHeading___Toc208803843)

[**CONCLUSION 20**](file:///C:\Users\barkin\Downloads\Table%20of%20Contents%20Template%20Word%2004.doc#__RefHeading___Toc208803850)

[**REFERENCES 23**](file:///C:\Users\barkin\Downloads\Table%20of%20Contents%20Template%20Word%2004.doc#__RefHeading___Toc208803850)

**INTRODUCTION**

Data mining is extracting information from data by finding patterns. In this Project, data mining techniques were used for 2 different data sets. To obtain the datasets, mushroom.arff from Github [3] (there is also a Kaggle page [1] of the mushroom dataset in CSV format) and Student\_por.csv, taken from the Kaggle students' performance analysis dataset webpage[2].

4 classification algorithms are used for comparement. These are:

Bayes->NaiveBayes (Naive Bayes classifier)

Tree->J48 (Decision tree classifier)

Rules->OneR (Rule based classifier)

Lazy->IBk(k-Nearest neighbor classifier)

To apply these algorithms Weka application is used. [4]

Weka is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms.Weka uses ARFF file formatted datasets that consist of independent, unordered instances and do not involve relationships among instances.[4]

**CHAPTER I: Dataset Descriptions**

In this chapter, datasets context, content, inspiration, acknowledgments, and attributes will be explained for both of the datasets.

**Mushroom Dataset Description[1]**

**Context**

Although this dataset was originally contributed to the UCI Machine Learning repository nearly 30 years ago, mushroom hunting (otherwise known as "shrooming") is enjoying new peaks in popularity. Learn which features spell certain death and which are most palatable in this dataset of mushroom characteristics. And how certain can your model be?

**Content**

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be'' for Poisonous Oak and Ivy.

Time period: Donated to UCI ML 27 April 1987

**Inspiration**

What types of machine learning models perform best on this dataset?

Which features are most indicative of a poisonous mushroom?

**Acknowledgements**

This dataset was originally donated to the UCI Machine Learning repository.

Number of Instances: 8124

**Attributes**

Number of Attributes: 22 (all nominally valued)

Attribute Information: (classes: edible=e, poisonous=p)

1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s

2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s

3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r,

pink=p,purple=u,red=e,white=w,yellow=y

4. bruises?:bruises=t,no=f

5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f,

musty=m,none=n,pungent=p,spicy=s

6. gill-attachment: attached=a,descending=d,free=f,notched=n

7. gill-spacing: close=c,crowded=w,distant=d

8. gill-size: broad=b,narrow=n

9. 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

10. stalk-shape: enlarging=e,tapering=t

11. stalk-root: bulbous=b,club=c,cup=u,equal=e,

rhizomorphs=z,rooted=r,missing=?

12. stalk-surface-above-ring: ibrous=f,scaly=y,silky=k,smooth=s

13. stalk-surface-below-ring: ibrous=f,scaly=y,silky=k,smooth=s

14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y

15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,

pink=p,red=e,white=w,yellow=y

16. veil-type: partial=p,universal=u

17. veil-color: brown=n,orange=o,white=w,yellow=y

18. ring-number: none=n,one=o,two=t

19. ring-type: cobwebby=c,evanescent=e,flaring=f,large=l,

none=n,pendant=p,sheathing=s,zone=z

20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r,

orange=o,purple=u,white=w,yellow=y

21. population: abundant=a,clustered=c,numerous=n,

scattered=s,several=v,solitary=y

22. habitat: grasses=g,leaves=l,meadows=m,paths=p,

urban=u,waste=w,woods=d

23.mushroom\_class: edible, poisonous

Missing Attribute Values: 2480 of them (denoted by "?"), all for attribute #11.

**Class Distribution**:

edible: 4208 (51.8%)

poisonous: 3916 (48.2%)

total: 8124 instances

### Students Performance Dataset Description[2]

### Context

This Dataset contains the information about the students and their Performance try to predict using the simple machine Learning Algorithm.

### Content

It contains the Specific information about the Schooling,Family issues,personal Relationship of the students,Internet facility and so on.

### Acknowledgements

This data are gathered from numerous sources thanks a lot [@uci](https://www.kaggle.com/uci)\_repository.

### Inspiration

Try to find out the issues with the Students as they are future Personalities, to save their Schooling and Youth life.

**Attributes**

1 school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)

2 sex - student's sex (binary: "F" - female or "M" - male)

3 age - student's age (numeric: from 15 to 22)

4 address - student's home address type (binary: "U" - urban or "R" - rural)

5 famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)

6 Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)

7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

9 Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at\_home" or "other")

10 Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at\_home" or "other")

11 reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")

12 guardian - student's guardian (nominal: "mother", "father" or "other")

13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16 schoolsup - extra educational support (binary: yes or no)

17 famsup - family educational support (binary: yes or no)

18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19 activities - extra-curricular activities (binary: yes or no)

20 nursery - attended nursery school (binary: yes or no)

21 higher - wants to take higher education (binary: yes or no)

22 internet - Internet access at home (binary: yes or no)

23 romantic - with a romantic relationship (binary: yes or no)

24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29 health - current health status (numeric: from 1 - very bad to 5 - very good)

30 absences - number of school absences (numeric: from 0 to 93)

These grades are related with the course subject,Porteguese:

31 G1 - first period grade (numeric: from 0 to 20)

32 G2 - second period grade (numeric: from 0 to 20)

33 G3 - final grade (numeric: from 0 to 20, output target)

31 pass-student’s passing the class(binary: yes or no)

The 31,32,33th attributes got deleted and a new 31th attribute (pass) is created by using the 33th attribute G3(final score). The new attribute checks if the student’s score is over the character D (%60) in this situation over 12/20.

**CHAPTER II: ANALYSIS RESULTS**

In this chapter you can see Weka’s output include build time, validation method, accuracy, evaluation scores, and confusion matrix.

**Mushroom Dataset Analysis Results**

**Naive Bayes**

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 7785 95.8272 %

Kappa statistic 0.9162

Mean absolute error 0.0419

Root mean squared error 0.1757

Relative absolute error 8.397 %

Root relative squared error 35.1617 %

Total Number of Instances 8124

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0,992 0,078 0,932 0,992 0,961 0,918 0,998 0,998 e

0,922 0,008 0,991 0,922 0,955 0,918 0,998 0,998 p

Weighted Avg. 0,958 0,044 0,960 0,958 0,958 0,918 0,998 0,998

=== Confusion Matrix ===

a b <-- classified as

4176 32 | a = e

307 3609 | b = p

**J48**

Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 8124 100 %

Kappa statistic 1

Mean absolute error 0

Root mean squared error 0

Relative absolute error 0 %

Root relative squared error 0 %

Total Number of Instances 8124

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1,000 0,000 1,000 1,000 1,000 1,000 1,000 1,000 e

1,000 0,000 1,000 1,000 1,000 1,000 1,000 1,000 p

Weighted Avg. 1,000 0,000 1,000 1,000 1,000 1,000 1,000 1,000

=== Confusion Matrix ===

a b <-- classified as

4208 0 | a = e

0 3916 | b = p

**IBK**

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 8124 100 %

Kappa statistic 1

Mean absolute error 0

Root mean squared error 0

Relative absolute error 0.0029 %

Root relative squared error 0.003 %

Total Number of Instances 8124

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1,000 0,000 1,000 1,000 1,000 1,000 1,000 1,000 e

1,000 0,000 1,000 1,000 1,000 1,000 1,000 1,000 p

Weighted Avg. 1,000 0,000 1,000 1,000 1,000 1,000 1,000 1,000

=== Confusion Matrix ===

a b <-- classified as

4208 0 | a = e

0 3916 | b = p

**OneR**

Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 8004 98.5229 %

Kappa statistic 0.9704

Mean absolute error 0.0148

Root mean squared error 0.1215

Relative absolute error 2.958 %

Root relative squared error 24.323 %

Total Number of Instances 8124

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1,000 0,031 0,972 1,000 0,986 0,971 0,985 0,972 e

0,969 0,000 1,000 0,969 0,984 0,971 0,985 0,984 p

Weighted Avg. 0,985 0,016 0,986 0,985 0,985 0,971 0,985 0,978

=== Confusion Matrix ===

a b <-- classified as

4208 0 | a = e

120 3796 | b = p

**Student Performance Analysis Dataset Analysis Results**

**Naive Bayes**

Time taken to build model: 0.1 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 483 74.4222 %

Kappa statistic 0.4811

Mean absolute error 0.2745

Root mean squared error 0.4527

Relative absolute error 55.1864 %

Root relative squared error 90.7779 %

Total Number of Instances 649

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0,658 0,181 0,759 0,658 0,705 0,485 0,801 0,787 no

0,819 0,342 0,735 0,819 0,774 0,485 0,801 0,786 yes

Weighted Avg. 0,744 0,267 0,746 0,744 0,742 0,485 0,801 0,786

=== Confusion Matrix ===

a b <-- classified as

198 103 | a = no

63 285 | b = yes

**J48**

Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 432 66.5639 %

Kappa statistic 0.3227

Mean absolute error 0.367

Root mean squared error 0.5367

Relative absolute error 73.7808 %

Root relative squared error 107.6279 %

Total Number of Instances 649

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0,585 0,264 0,657 0,585 0,619 0,324 0,640 0,572 no

0,736 0,415 0,672 0,736 0,702 0,324 0,640 0,619 yes

Weighted Avg. 0,666 0,345 0,665 0,666 0,664 0,324 0,640 0,597

=== Confusion Matrix ===

a b <-- classified as

176 125 | a = no

92 256 | b = yes

**IBK**

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 424 65.3313 %

Kappa statistic 0.3

Mean absolute error 0.3472

Root mean squared error 0.5878

Relative absolute error 69.8062 %

Root relative squared error 117.8679 %

Total Number of Instances 649

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0,595 0,296 0,635 0,595 0,614 0,301 0,653 0,572 no

0,704 0,405 0,668 0,704 0,685 0,301 0,653 0,633 yes

Weighted Avg. 0,653 0,355 0,652 0,653 0,652 0,301 0,653 0,605

=== Confusion Matrix ===

a b <-- classified as

179 122 | a = no

103 245 | b = yes

**OneR**

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 411 63.3282 %

Kappa statistic 0.258

Mean absolute error 0.3667

Root mean squared error 0.6056

Relative absolute error 73.7282 %

Root relative squared error 121.432 %

Total Number of Instances 649

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0,558 0,302 0,615 0,558 0,585 0,259 0,628 0,548 no

0,698 0,442 0,646 0,698 0,671 0,259 0,628 0,613 yes

Weighted Avg. 0,633 0,377 0,632 0,633 0,631 0,259 0,628 0,583

=== Confusion Matrix ===

a b <-- classified as

168 133 | a = no

105 243 | b = yes

**CONCLUSION**

**Comparing Datasets**

|  |  |  |
| --- | --- | --- |
|  | **Mushroom Dataset** | **Students Perf. Dataset** |
| **Data Size** | Larger (8123 rows) | Smaller (649 rows) |
| **Attribute Number** | Smaller (23 columns) | Larger (31 columns) |
| **Missing Values** | No | Yes |
| **Balance of the Dataset** | Balanced with 51% for edible (4208 edible, 3916 posionous) | Balanced with 53% for yes  (301 no, 348 yes) |

By data size, the Mushroom Dataset is a lot larger than Student Performance Dataset. Student Performance Dataset, on the other hand, has 8 more attributes than Mushroom Dataset. In contrast to the Mushroom Dataset, the Student Performance Dataset contains missing values; nonetheless, these missing values are insignificant in terms of the dataset. Both datasets are balanced in terms of class types.

**Comparing Algorithms Accuracy**

|  |  |  |
| --- | --- | --- |
|  | **Mushroom Dataset** | **Students Perf. Dataset** |
| **Naive Bayes** | 95.8272% | 74.4222% |
| **J48** | 100% | 66.5639% |
| **IBK** | 100% | 65.3313% |
| **OneR** | 98.5229% | 63.3228% |

**Since both datasets are balanced, we can be sure of their accuracy.** If they weren't balanced, we could look at the F-measure, which is the harmonic mean of Precision and Recall. Alternatively, you can use a weighted f-measure to balance the weight between Precision and Recall.

J48 and IBK are the two highest scoring algorithms in the Mushroom Dataset, both with 100 percent accuracy. **Normally, 100 % accuracy indicates overfitting (the model memorizes values), which is undesirable, but when stratified cross-validation is applied, it is fair to believe that these methods work perfectly and do not overfit.**

The best algorithm for the Students Performance Dataset is Nave Bayes. J48, IBK, and OneR follow Naïve Bayes with similar scores among themselves.

**The reason why Nave Bayes scored best on the Student's Performance Dataset and lowest on the Mushroom Dataset is that** Bayesian classifiers perform better on larger datasets. However, naive Bayes assumes that attributes are independent, and if they are dependent, the model's accuracy drops. There are more attributes in the Student's Performance Dataset, and these attributes are mainly independent (except income-related attributes, they are dependent). However, it is safe to presume that the features of the Mushroom Dataset are much more closely related (gill related and stalk related multiple attributes exist).

**Because J48 is a decision tree classifier, it performed second best on the Students Performance Dataset.** With smaller datasets, decision trees perform better. Decision trees also outperform other models when it comes to categorical values. Whereas other models cannot detect patterns between categorical values, decision trees can.

**Because it is a rule-based classifier, OneR performs similarly but slightly worse than the J48.** Decision trees are quite similar to rule-based classifiers (you can also convert trees to rules). Rule-based classifiers provide rules that classify data, and these rules are much easier for people to understand but perform worse (larger the tree, it gets harder to understand by humans).

**Comparing Model Build Time**

|  |  |  |
| --- | --- | --- |
|  | **Mushroom Dataset** | **Students Perf. Dataset** |
| **Naive Bayes** | 0.03 s | 0.10 s |
| **J48** | 0.04 s | 0.05 s |
| **IBK** | 0 s | 0 s |
| **OneR** | 0.04 s | 0.01 s |

The IBK algorithm is a lazy learner algorithm that does not require model development (it stores training data), **which explains why creating the model for both datasets takes zero seconds.**

Bayesian classifiers have high speed on huge datasets. **As a result, Nave Bayes performs second fastest on the mushroom dataset.**

**However, Nave Bayes performs significantly worse on the Students Performance Dataset** because, while being smaller, it contains more attributes.

The Mushroom Dataset is large,and J48 and OneR both have the same build time, but for students performance dataset (which is a smaller dataset), OneR is clearly faster. Which makes sense because we know it is less accurate in comparison to J48 from the "Comparing Accuracy Algorithms Accuracy" table.

**REFERENCES**

**[1]UCI Machine Learning,(2017), Mushroom Classification**

**[https://www.kaggle.com/uciml/mushroom-classification]**

**[2]Bala Vashan,(Last Update 31,12,2021),Students Performance Dataset**

**[https://www.kaggle.com/balavashan/students-performance-dataset?select=student-por.csv]**

**[3]renatopp,(20 December 2012), arff-datasets**

**[https://github.com/renatopp/arff-datasets/tree/master/classification]**

**[4]Weka wiki,(view 2 January 2022), weka-wiki**

**[https://waikato.github.io/weka-wiki]**