CS286 Lab 2: Exploratory data analysis in WEKA

K. A. Tarnowska

31-Aug-2020

Contents

1 Introduction

In this Lab, you apply the concepts of Rough Set Theory to the dataset chosen in Lab 1. In particular, you will describe the dataset as a *decision table*. You will analyze the *decision-making process* that the dataset simulates and *knowledge* represented by the decision table. You will learn how to install and import the dataset into WEKA and perform You will get hands-on experience with Exploratory Data Analysis in WEKA.

Pre-requisites This lab assumes the completion of Lab 1.

2 Rough Set Theory [1]

2.1 Decision Table

Based on the dataset that you chose for Lab 1, **answer the following questions**:

- How many *objects* does the *decision table* describe?
- What does the *object* represent? (i.e. patient, log entry)
- How many *attributes* describe the objects?
- What are the *conditional* and what attribute(s) is the *decision*?

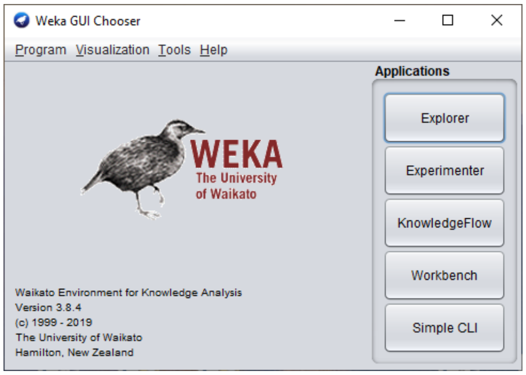
2.2 Knowledge Representation

Based on the dataset that you chose for Lab 1, **answer the following questions**:

- What *knowledge* is represented by the chosen decision table?
- Who is the *expert* involved? (i.e. physician)
- What *decision-making process* does it simulate? (i.e. diagnosis, treatment)

3 Introduction to WEKA [2]-[5]

WEKA is one of several well-known free machine learning choices. Not only is the software widely used, but it is complemented by a textbook [2] an online course [4], and extensive YouTube [5] instructional videos. WEKA requires no programming or math skills and the graphical user interface (GUI) is intuitive, in terms of uploading and navigating the software. It was developed at the University of Waikato in New Zealand. It is free & open source and supports Windows, macOS, and Linux platform. It is Java-based and provides for Java API for embedded machine learning.



Waikato Environment for Knowledge Analysis
Version 3.8.4
(c) 1999 - 2019
The University of Waikato
Hamilton, New Zealand

Figure 1: WEKA GUI chooser.

3.1 Input

The input can be a csv file, an arff file (attribute-relation file format), which is a text file that includes metadata, xlx, xlsx, json and several others. Data can also be uploaded from a URL or a remote database. WEKA comes with multiple sample datasets, so the user can train on validated data.

3.2 User Interface

When the program is opened the choices are (see Figure 1): Explorer (default program), Experimenter (where you can compare model results with t-tests), KnowledgeFlow (Visual workflow), Workbench (combines all of the GUIs into one interface) and SimpleCLI (command-line interface).

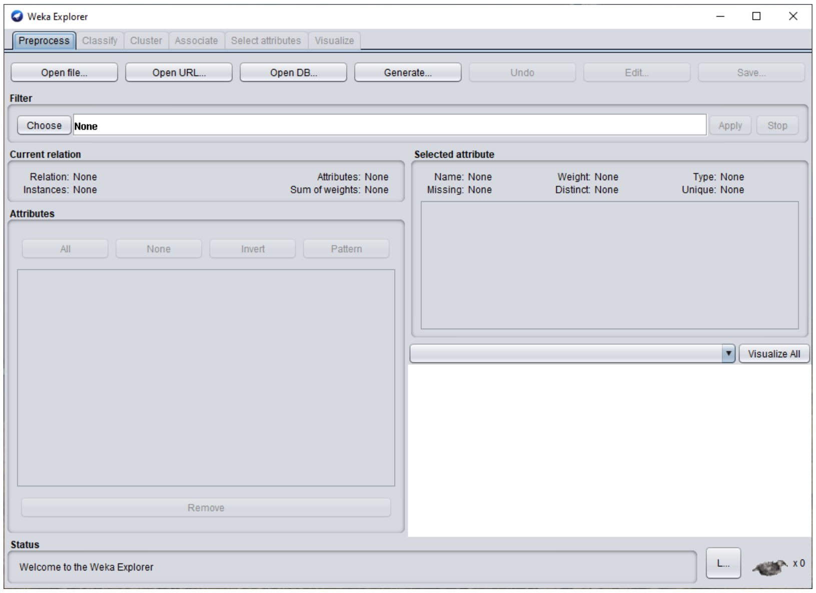
The Explorer The Explorer option is divided into 5 tabs (see Figure 2): Preprocess (upload, visualize, filter, clean, etc.), Classify (classification and regression algorithms), Cluster (unsupervised), Associate (unsupervised), Select Attributes (ranks importance of variables) and Visualize (creates a master scatter plot so you can see if one variable seems to correlate with another).

3.3 Installation

Download WEKA from [3]. **Perform installation** steps as described in [3]. Open WEKA Explorer (see Figures 1 and 2).

3.4 Loading Data

Load the dataset you chose in Lab 1 and described in Section 2. **Include screenshot** showing imported data. See Figure 3 as an example. Explore information displayed in WEKA: relation name, the number of instances, the number of attributes (they all should be visible on the provided screenshot).



Filter: None

Current relation

Relation: None
Instances: None
Attributes: None
Sum of weights: None

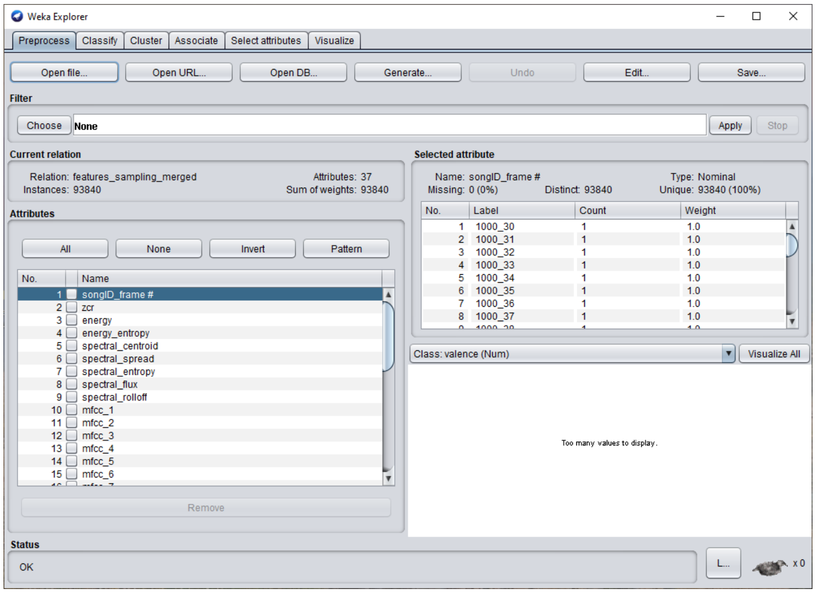
Selected attribute

Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None

Status

Welcome to the Weka Explorer

Figure 2: GUI for WEKA Explorer



Filter: None

Current relation

Relation: features_sampling_merged
Instances: 93840
Attributes: 37
Sum of weights: 93840

Attributes

No.		Name
1	□	songID_frame #
2	□	zcr
3	□	energy
4	□	energy_entropy
5	□	spectral_centroid
6	□	spectral_spread
7	□	spectral_entropy
8	□	spectral_flux
9	□	spectral_rolloff
10	□	mfcc_1
11	□	mfcc_2
12	□	mfcc_3
13	□	mfcc_4
14	□	mfcc_5
15	□	mfcc_6

Selected attribute

Name: songID_frame #
Missing: 0 (0%)
Distinct: 93840
Type: Nominal
Unique: 93840 (100%)

No.	Label	Count	Weight
1	1000_30	1	1.0
2	1000_31	1	1.0
3	1000_32	1	1.0
4	1000_33	1	1.0
5	1000_34	1	1.0
6	1000_35	1	1.0
7	1000_36	1	1.0
8	1000_37	1	1.0

Class: valence (Num)

Too many values to display.

Status

OK

Figure 3: Example of loading data into WEKA (here, audio dataset)

4 Exploratory Data Analysis

Data analysis refers to hands-on data exploration and evaluation. **Exploratory analysis** is an approach to analyzing datasets to find previously unknown relationships. Often such analysis involves using various data visualization approaches.

Statistical vs. Exploratory Data Analysis Statistical approaches have “a priori” (presupposed by experience) hypothesis to test. For example, apriori question: Has increasing fee-structure led to decreasing market share” leads to formulating the hypothesis: “Market share has decreased”. Next, the analyst applies statistical methods to test the hypothesis. However, not always we have a priori notions about data. In this case, use **Exploratory Data Analysis (EDA)**. It is approach useful for:

- Delving into data
- Examining important interrelationships between attributes
- Identifying interesting subsets of the objects
- Develop an initial idea of possible associations amongst the attributes, as well as between the predictors and the decision.

General goals of EDA are:

- Investigate variables
- Examine *distributions* of *categorical* variables
- Look at *histograms* of *numerical* variables
- Explore relationships among sets of variables (i.e. using *correlation*)

Variables Variables (or attributes) describe the data we are working with. An attribute (or a variable) is a label we give to our data. The data can be *numerical* or *categorical*. **Categorical** variables have values that can be separated into distinct categories (i.e. animal taxonomy: mammals, reptiles, etc.). If we use numbers to represent categories, they become **nominal** variables. However, these numbers are only used to represent the categories, but cannot be used for meaningful mathematical or statistical operations. A variable that is thought not to be controlled or affected by other variables is called an **independent** variable. A variable that depends on other variables is called a **dependent** variable.

Frequency Distribution It is useful to plot a graph showing how many times each value occurs. The common way to visualize frequency distribution is to use a **histogram**. Histograms plot values of observations on the horizontal axis, with a bar showing how many times each value occurred in the dataset. The *normal distribution* is characterized by a bell-shaped curve, where data is distributed symmetrically around the center of all values. There are two ways in which a distribution can deviate from normal:

- Lack of symmetry (called **skew**) - a skewed distribution can be either positively skewed ("to the left") or negatively skewed ("to the right")
- Pointiness (called **kurtosis**) - refers to how "pointy" a distribution is.

Measures of Centrality and Distribution Often one number can tell us enough about a distribution. This is typically a number that points to the "center" of a distribution. There are three measures commonly used: **mean**, **median**, and **mode**. *Mean* is commonly known as an **average**. The *median* is the middle value for a dataset. The *mode* is the most frequently occurring value in a dataset. Distributions come in all shapes and sizes. Simply looking at the central point (mean, median, or mode) may not help in understanding the actual shape of a distribution. Therefore, we often look at the *spread*, or the *dispersion*, of a distribution. The common measures are **range**, **variance**, and **standard deviation**. The **range** is calculated by subtracting the smallest value from the largest value. Sometimes we discard extreme values, and calculate range only for the middle 50% of values. This is known as the **interquartile range**. The *variance* and *standard deviation* indicate how spread out the data points are. To measure the variance, the common method is to pick a center of a distribution, typically the mean, and measure how far each data points from the center. The variance gives us the measure of spread in units squared. The standard deviation is the squared root of the variance, which ensures the measure of average spread is in the same units as the original measure.

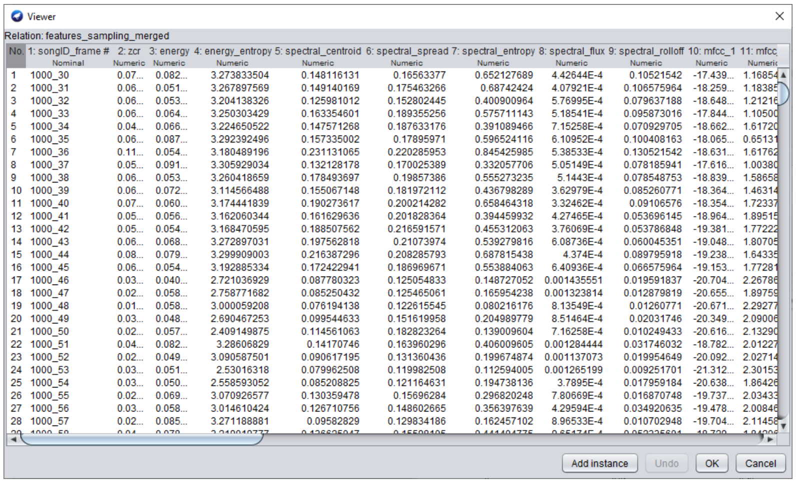
Correlation Correlation is a statistical analysis that is used to measure and describe the *strength* and *direction* of the relationship between two variables. Correlation is a simple statistical measure that examines how two variables change together over time. This indicates how closely two variables are related and ranges from -1 (negatively related) to +1 (positively related.) A correlation of 0 indicates no relation between two variables.

4.1 Get to Know your Data Set

Graphs, plots, and tables often uncover important relationships in data.

4.1.1 Viewing data

Provide a screenshot of the subset of objects from your data set presented in the tabular format. You can use the WEKA Explorer *Edit* button to open your



Relation: features_sampling_merged

No.	1: songID_frame #	2: zcr	3: energy	4: energy_entropy	5: spectral_centroid	6: spectral_spread	7: spectral_entropy	8: spectral_flux	9: spectral_rolloff	10: mfcc_1	11: mfcc
	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	1000_30	0.07...	0.082...	3.273833504	0.148116131	0.16563377	0.652127689	4.42644E-4	0.10521542	-17.439...	1.16854
2	1000_31	0.06...	0.051...	3.267897569	0.149140169	0.175463266	0.68742424	4.07921E-4	0.106575964	-18.259...	1.18385
3	1000_32	0.06...	0.053...	3.204138326	0.125981012	0.152802445	0.400900964	5.76995E-4	0.079637188	-18.648...	1.21216
4	1000_33	0.06...	0.064...	3.250303429	0.163354601	0.189355256	0.575711143	5.18541E-4	0.095873016	-17.844...	1.10500
5	1000_34	0.04...	0.066...	3.224650522	0.147571268	0.187633176	0.391089466	7.15258E-4	0.070929705	-18.662...	1.61720
6	1000_35	0.06...	0.087...	3.292392496	0.157335002	0.17895971	0.596524116	6.10952E-4	0.100408163	-18.065...	0.65131
7	1000_36	0.11...	0.054...	3.180489196	0.231131065	0.220285953	0.845425985	5.38533E-4	0.130521542	-18.631...	1.61762
8	1000_37	0.05...	0.091...	3.305929034	0.132128178	0.170025389	0.332057706	5.05149E-4	0.078185941	-17.616...	1.00380
9	1000_38	0.06...	0.053...	3.260418659	0.178493697	0.19857386	0.555273235	5.1443E-4	0.078548753	-18.839...	1.58658
10	1000_39	0.06...	0.072...	3.114566488	0.155067148	0.181972112	0.436798289	3.62979E-4	0.085260771	-18.364...	1.46314
11	1000_40	0.07...	0.060...	3.174441839	0.190273617	0.200214282	0.658464318	3.32462E-4	0.09106576	-18.354...	1.72337
12	1000_41	0.05...	0.056...	3.162060344	0.161629636	0.201828364	0.394459932	4.27465E-4	0.053696145	-18.964...	1.89515
13	1000_42	0.05...	0.054...	3.168470595	0.188507562	0.216591571	0.455312063	3.76069E-4	0.053786848	-19.381...	1.77222
14	1000_43	0.06...	0.068...	3.272897031	0.197562818	0.21073974	0.539279816	6.08736E-4	0.060045351	-19.048...	1.80705
15	1000_44	0.08...	0.079...	3.299909003	0.216387296	0.208285793	0.687815438	4.374E-4	0.089795918	-19.238...	1.64335
16	1000_45	0.06...	0.054...	3.192885334	0.172422941	0.186969671	0.553884063	6.40936E-4	0.066575964	-19.153...	1.77281
17	1000_46	0.03...	0.040...	2.721036929	0.087780323	0.125054833	0.148727052	0.001435551	0.019591837	-20.704...	2.26786
18	1000_47	0.02...	0.058...	2.758771682	0.085250432	0.125465061	0.165954238	0.001323814	0.012879819	-20.655...	1.89759
19	1000_48	0.01...	0.058...	3.000059208	0.076194138	0.122615545	0.080216176	8.13549E-4	0.01260771	-20.671...	2.29277
20	1000_49	0.03...	0.048...	2.690467253	0.099544633	0.151619958	0.204989779	8.51464E-4	0.02031746	-20.349...	2.09006
21	1000_50	0.02...	0.057...	2.409149875	0.114561063	0.182823264	0.139009604	7.16258E-4	0.010249433	-20.616...	2.13290
22	1000_51	0.04...	0.082...	3.28606829	0.14170746	0.163960296	0.406009605	0.001284444	0.031746032	-18.782...	2.01227
23	1000_52	0.02...	0.049...	3.090587501	0.090617195	0.131360436	0.199674874	0.001137073	0.019954649	-20.092...	2.02714
24	1000_53	0.03...	0.051...	2.53016318	0.079962508	0.119982508	0.112594005	0.001265199	0.009251701	-21.312...	2.30153
25	1000_54	0.03...	0.050...	2.558593052	0.085208825	0.121164631	0.194738136	3.7895E-4	0.017959184	-20.638...	1.86426
26	1000_55	0.02...	0.069...	3.070926577	0.130359478	0.15696284	0.296820248	7.80669E-4	0.016870748	-19.737...	2.03433
27	1000_56	0.03...	0.058...	3.014610424	0.126710756	0.148602665	0.356397639	4.29594E-4	0.034920635	-19.478...	2.00846
28	1000_57	0.02...	0.085...	3.271188881	0.09582829	0.129834186	0.162457102	8.96533E-4	0.010702948	-19.704...	2.11458

Figure 4: Displaying data instances (objects) in the Viewer of WEKA Explorer.

data in the *Viewer* (see Figure 4 for an example).

4.1.2 High-level data analysis

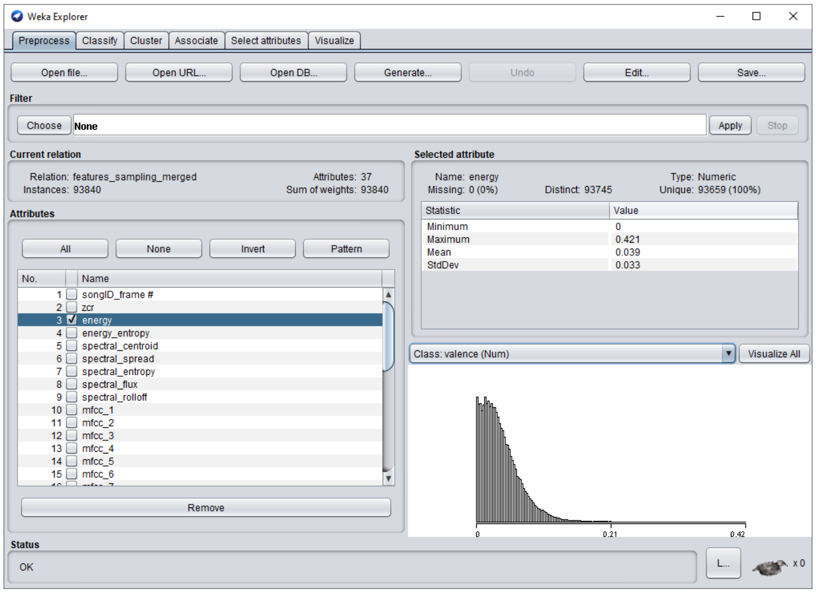
Provide high-level insights from your EDA. Describe attributes, their meanings, their types (i.e numerical, categorical, nominal), the decision attribute, the type of decision attribute (i.e. binary, categorical), data missing.

4.1.3 Fine-grained attribute analysis

Select an attribute and explore the information provided on the *Selected attribute* section, such as minimum, maximum, mean, standard deviation (if numeric) or frequency analysis (if catgeorical). **Provide a screenshot** with the information displayed on the selected attribute. See Figure 5 as an example of exploring numerical attribute and 6 as an example of exploring nominal (categorical) attribute.

4.1.4 Summarization and Visualization

Provide a graph (histogram or distribution chart) or at least one chosen attribute (see Figure 7 for an example). **Provide a comment** about a distribution. Use the *Visualize* tab and analyze scatter plots. **Provide comment** if you have discovered any correlations between variables, if so **provide the scatter plot** that supports this.



Filter: None

Current relation

Relation: features_sampling_merged
Instances: 93840
Attributes: 37
Sum of weights: 93840

Selected attribute

Name: energy
Missing: 0 (0%)
Distinct: 93745
Type: Numeric
Unique: 93659 (100%)

Statistic	Value
Minimum	0
Maximum	0.421
Mean	0.039
StdDev	0.033

Attributes

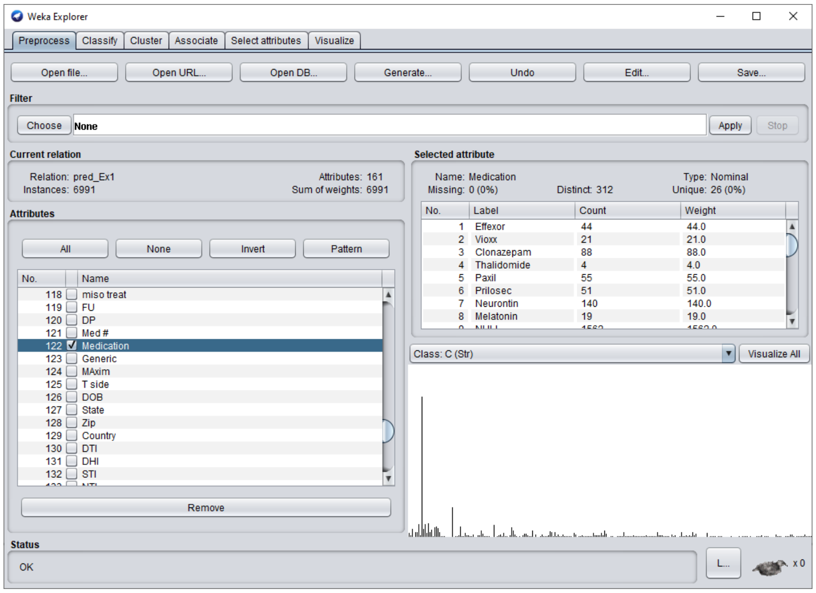
No.		Name
1	□	songID_frame #
2	□	zcr
3	☑	energy
4	□	energy_entropy
5	□	spectral_centroid
6	□	spectral_spread
7	□	spectral_entropy
8	□	spectral_flux
9	□	spectral_rolloff
10	□	mfcc_1
11	□	mfcc_2
12	□	mfcc_3
13	□	mfcc_4
14	□	mfcc_5
15	□	mfcc_6

Class: valence (Num)

Status

OK

Figure 5: Example of exploring a selected attribute (here, numerical attribute for the energy of the sound), showing positively skewed distribution



Filter: None

Current relation

Relation: pred_Ex1
Instances: 6991
Attributes: 161
Sum of weights: 6991

Selected attribute

Name: Medication
Missing: 0 (0%)
Distinct: 312
Type: Nominal
Unique: 26 (0%)

No.	Label	Count	Weight
1	Effexor	44	44.0
2	Vioxx	21	21.0
3	Clonazepam	88	88.0
4	Thalidomide	4	4.0
5	Paxil	55	55.0
6	Prilosec	51	51.0
7	Neurontin	140	140.0
8	Melatonin	19	19.0

Attributes

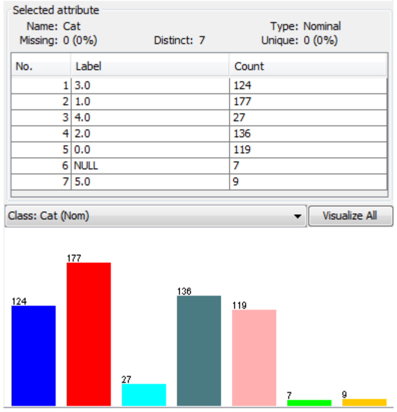
No.		Name
118	□	miso treat
119	□	FU
120	□	DP
121	□	Med #
122	☑	Medication
123	□	Generic
124	□	MAxim
125	□	T side
126	□	DOB
127	□	State
128	□	Zip
129	□	Country
130	□	DTI
131	□	DHI
132	□	STI

Class: C (Str)

Status

OK

Figure 6: Example of exploring a selected attribute (here, a categorical attribute for the medication of the patient)



Selected attribute

Name: Cat

Type: Nominal

Missing: 0 (0%)

Distinct: 7

Unique: 0 (0%)

No.	Label	Count
1	3.0	124
2	1.0	177
3	4.0	27
4	2.0	136
5	0.0	119
6	NULL	7
7	5.0	9

Figure 7: Sample distribution graph of a categorical decision attribute for tinnitus category.

5 Deliverables

Lab report uploaded on Canvas (pdf preferred) with embedded screenshots.

6 Grading

The lab will be graded according to the rubric in Table 1.

Table 1: Grading rubric

Rubric	Criteria	Points
Decision Table	Answered all questions in 2.1 correctly and thoroughly	2
Knolwedge Representation	Answered all questions in 2.2 correctly and thoroughly	1
Data loading	Provided proof of successful data loading as described in 3.4	1
Data exploration	Provided screenshot of data view as described in 4.1.1	1
Data analysis	Provided EDA analysis as described in 4.1.2	2
Attribute analysis	Provided screenshot and analysis of the selected attribute as described in 4.1.3	1
Visualization	Provided a graph with data distribution and commented as described in 4.1.4	1
Correlation analysis	Commented on correlation analysis as described in 4.1.4	1
Total		**10**

7 Self-check

Answer the following questions:

1. What is WEKA and what are its main advantages?
2. Describe the data mining process as visualized by tabs on WEKA Explorer.
3. What is EDA and what is it used for?
4. Describe how exploratory analysis differs from statistical analysis.
5. What is a variable? What are the types of variables?
6. Given one example of a numerical and one example of a categorical variable.
7. What are the common techniques used in the exploratory analysis?
8. What are the measures of centrality?
9. What are the measures of distribution?
10. How can we measure the relationship between variables?

References

[1] Tarnowska KA, Ras ZW, Jastreboff PJ. Knowledge Discovery Approach for Recommendation, Chapter 4 in: Decision Support System for Diagnosis and Treatment of Hearing Disorders. The Case of Tinnitus. Studies in Computational Intelligence. Springer, 2017

[2] Witten IH, Frank E and Hall MA. Introduction to WEKA/The Explorer, Chapter 10 & 11.1, in: Data Mining: Practical Machine Learning Tools and Techniques. Fourth Edition. Morgan Kaufmann. 2017

[3] WEKA 3: Data Mining Software in Java. https://www.cs.waikato.ac.nz/ml/weka/ Accessed April 2, 2019

[4] Future Learn. Data Mining with WEKA. https://www.futurelearn.com/courses/data-mining-withweka Accessed April 3, 2019

[5] YouTube WEKA videos. https://www.youtube.com/results?search_query=WEKA Accessed April 3, 2019