EXPERIMENT DESIGN FOR DATA SCIENCE

EXERCISE: **PAPER REPRODUCTION**

Konstantin Damnjanovic, 01151948
Moritz Leidinger, 11722966
Dzan Operta, 11935976

DATA PROCESSING

USER RATING

parsed XML

fill missing values with mean

VISUAL

avg. of both rows per column

or

append 2nd row to 1st

METADATA

parsed XML

binarized categorical columns

AUDIO

avg. per row

transpose

TEXTUAL

transpose values but keep column labels

add movie names

TABLE 1: **WEKA**
RESULTS

	Precision	Recall	F1
User rating	very low (0.371)	0.610 (0.609)	very low (0.461)
Visual	0.545 (0.447)	0.520 (0.476)	0.526 (0.458)
Metadata	0.462 (0.524)	0.448 (0.516)	0.454 (0.519)
Metadata + user rating	0.478 (0.581)	0.462 (0.6)	0.468 (0.583)
Metadata + visual	0.528 (0.584)	0.511 (0.6)	0.517 (0.586)

TABLE 2: **CLASSIFIER SELECTION**
+ LAS VEGAS WRAPPER

STEPS

- load data
- select classifiers
- use LVW script for feature selection
  - combination of GitHub script and pseudocode

RESULTS

- 29 classifiers with none of the scores below 0.5
- most overlap with table 2 from the paper (18/21)

k-Nearest neighbor	**metadata**
Nearest mean classifier	**metadata**
Decision tree	**metadata**
Logistic regression	**metadata**
SVM (Gaussian Kernel)	**metadata**
Gradient Boosting Tree	**metadata**
k-Nearest neighbor	**textual**
Nearest mean classifier	textual
Decision tree	textual
Logistic regression	textual
SVM (Gaussian Kernel)	**textual**
Bagging	textual
Random Forest	textual
Gradient Boosting Tree	textual
Naive Bayes	**textual**
k-Nearest neighbor	**visual**
Nearest mean classifier	visual
Decision tree	**visual**
Logistic regression	**visual**
SVM (Gaussian Kernel)	**visual**
Bagging	visual
Random Forest	**visual**
AdaBoost	**visual**
Gradient Boosting Tree	**visual**
Naive Bayes	visual
Logistic regression	**audio**
SVM (Gaussian Kernel)	audio
AdaBoost	audio
Gradient Boosting Tree	**audio**

OUTLOOK

Metadata

- try other methods for encoding categorical values
- try better handling of missing values

Table 3

- implement voting, label stacking and label attribute stacking

Misc.

- test for significance of differences in results
- write report