ASSIGNMENT 2

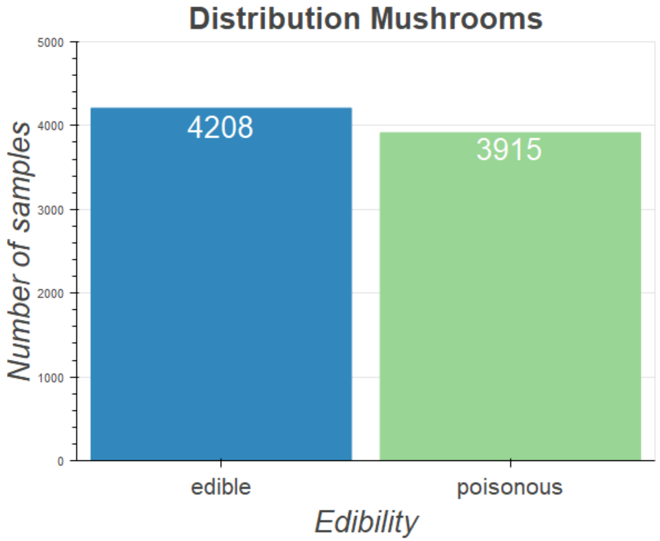
GROUP 20

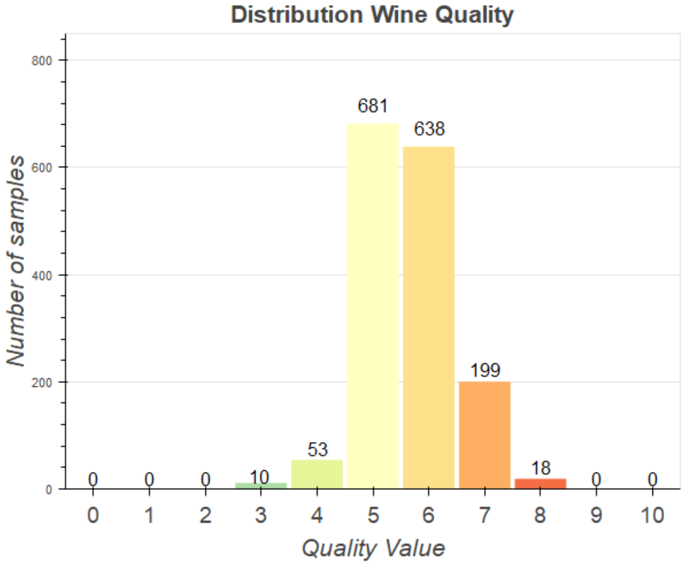
SOPHIE RAIN (01425316),
LUCAS UNTERBERGER (01325438),
PETER STROPPA (01326468).

datasets

- CONGRESS
- AMAZON
- WINE
- MUSHROOMS

distribution of classes





datasets overview

Dataset	# samples	# attributes	Missing values?	# classes	classes eq. important?	Types of attributes
Amazon	750	10 000	no	50	yes	ordinal
Congress	217	18	yes (unknown)	2	yes	categorical (2 values)
Wine	1599	11	no	11	no	numeric
Mushrooms	8124	22	yes	2	no	categorical (2-10)

methods

- K-Nearest-Neighbors
- Support Vector Machines
- Random Forest

pre-processing

scaling and selection

- Min-Max-Scaling
- Z-Score-Scaling
- Feature Selection
- Custom weighted scaling

encoding

- Ordinal Encoding
- One Hot Encoding
- TF-IDF

pre-processing

Weighted scaling

If correlation < 0:

- $$\frac{max(feature)-feature}{max(feature)*|correlation|}$$

If correlation > 0:

- $$\frac{feature}{max(feature)*|correlation|}$$

TF-IDF (text frequency - inverse document frequency)

- Upweighting rare words (important)
- Downweighting frequent words (syntactic)

$$\text{tfidf(w,R)=ln(}\frac{\#reviews}{1+\#reviews\ containing\ w}\text{)*n(w,R)}$$

results

Best result	kNN	Random Forest	SVM
Wine		perfectly	perfectly
Congress			97% accuracy
Mushroom	perfectly	perfectly	perfectly
Amazon		67% accuracy	

conclusion

different datasets

- kNN disappointing
  - high hopes kNN + TF-IDF
- SVM highly dependent on kernel
- Don‘t forget human intuition

Different methods

- SVM & RF outperformed
- kNN troubles with many classes
- only RF can handle heterogeneous data