Machine Learning Assignment 1

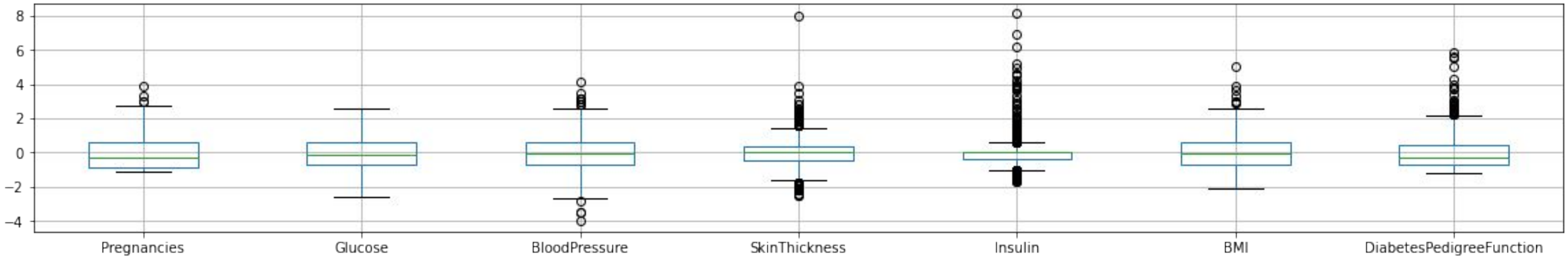
Group 1
Jonas Kompauer, 11776872
Lukasz Sobocinski, 12123563
Florian Lackner, 11704916

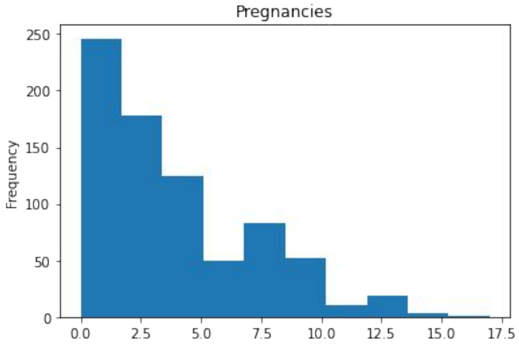
Dataset - Diabetes

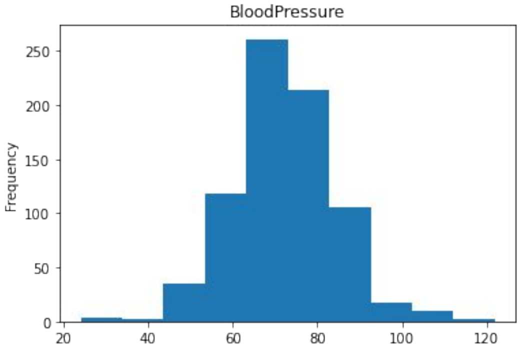
- Small Dataset, ~770 rows
- 8 numeric features
- Missing values

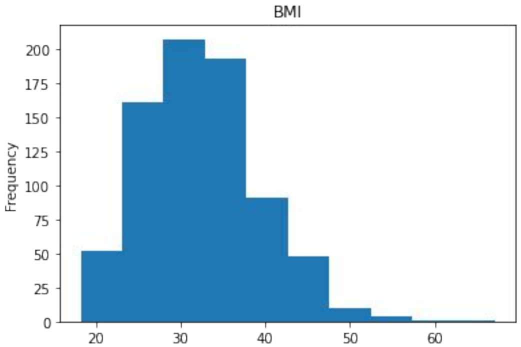
Index	Pregnancies	Glucose	Blood Presure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0

Dataset - Diabetes









Dataset - Breast Cancer

- Small Dataset, ~280 rows
- 30 numeric features
- No missing values

ID	class	radiusMean	textureMean	perimeter Mean	smoothnes Mean	compactnes Mean	...	symmetry Worst	fractal Dimension Worst
913102	false	14.64	16.85	94.21	0.08641	0.06698	...	0.2455	0.06596
89511501	false	12.2	15.21	78.01	0.08673	0.06545	...	0.2661	0.07961
87163	true	13.43	19.63	85.84	0.09048	0.06288	...	0.2884	0.07371
894047	false	8.597	18.6	54.09	0.1074	0.05847	...	0.3142	0.08116

Dataset - Purchase

- Large Dataset, 10k rows
- 600 binary attributes
- Target attribute consists of 100 classes
- No missing values

ID	0	1	2	3	4	5	...	599	class
0	0	0	0	1	1	0	...	0	86
1	0	0	0	1	1	0	...	0	81
2	0	1	1	1	1	1	...	0	3
3	0,	1	0	1	1	1	...	0	19

Dataset - Speeddating ❤️

- Data about 2 Persons -> find out if they match
- Large Dataset, ~8k rows
- Mixture of numerical and nominal data
- 121 features
- Missing Values

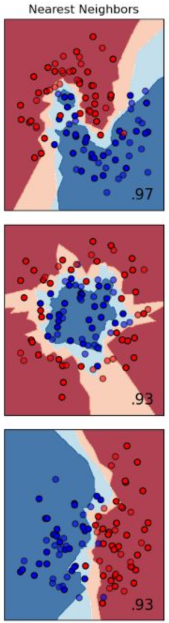
wave	gender	age	race	importance_same_race	attractive	funny	...	met	match
1	female	21	Asian/Pacific Islander/Asian-American	2	6	8	...	0	0
2	male	24	European/Caucasian-American	1	3	7	...	1	1
3	female	26	'Latino/Hispanic American'	1	9	9	...	0	0

Classifier

	kNN	Gaussian Naive Bayes	Decision Tree
Parameters	k, distance function, weighting	smoothing	depth, sample split, split selection, criterion
Results	mostly good results	best results for Purchase Dataset	less accurate, but still mostly good results
Time	fast fitting, slow predicting	more or less fast in both fitting and predicting	slow fitting, fast predicting

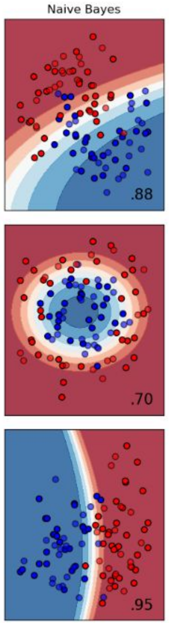
Usefulness kNN

- Feature Scaling needed
- Not Useful for large number of features
  - “Purchase” with 600 features no good results
  - “Speed Dating” with ~120 feature still good results
  - Otherwise very good results



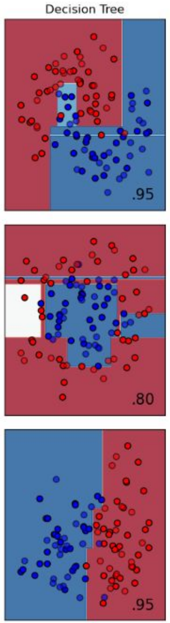
Usefulness Gaussian Naive Bayes

- Assumes independence of variables and their normal distribution
  - Not 100% given in our datasets
- Still has the best results for “Purchase” Dataset
  - Accuracy of ~0.5
- For other Datasets also good results
  - Diabetes with accuracy of 0.73
  - Speed Dating with accuracy of 0.859
  - Breast Cancer with accuracy of 0.947



Usefulness Decision Tree

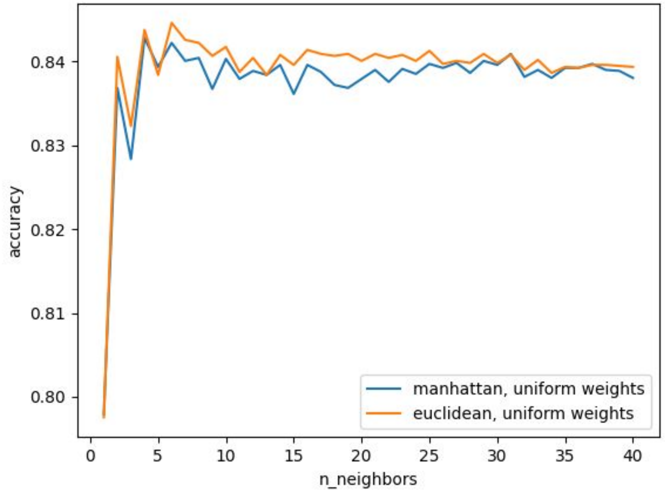
- Short classification time, when the model is trained
- Too complex with large number of features/data
  - “Purchase” with 600 features bad results
  - “Speed Dating” with ~120 feature still good results, accuracy of ~0.86



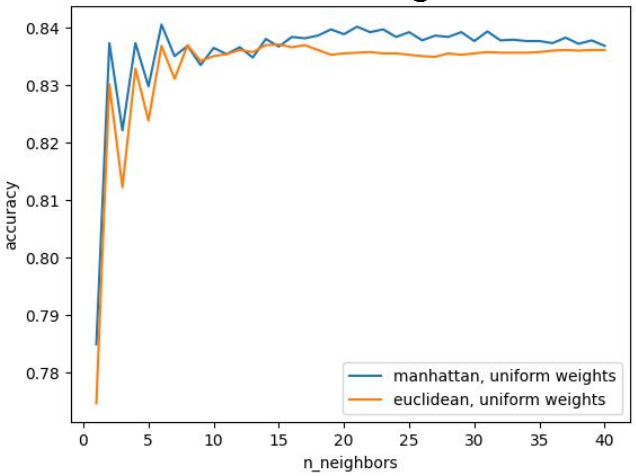
Findings

- Scaling had close to no effect for kNN on accuracy

Scaling



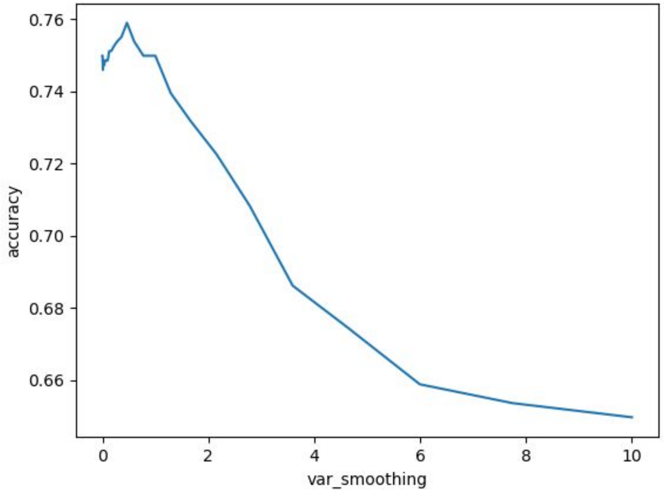
No Scaling



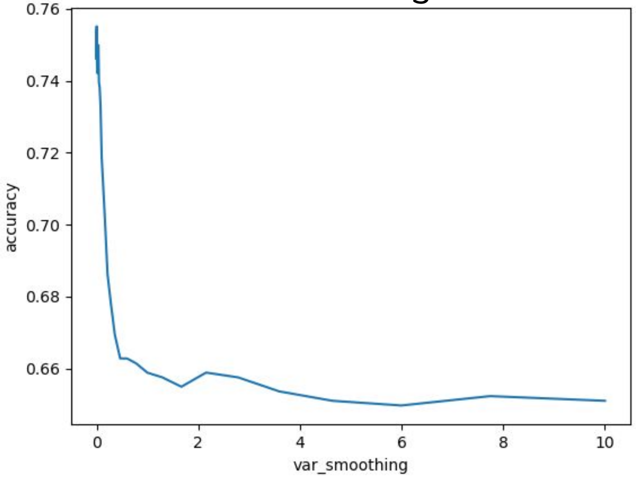
Findings

- except...

Scaling



No Scaling



Issues

- Hard to know on what to focus with experiments
- Find important parameters to compare
- 4 datasets for 3 group members - difficult to split
- Some classifiers were unstable - difficult to get the reliable performance measures

Summary

- Decent results for the datasets
- All Algorithms performed equally good on the datasets except for “Purchase”
- CV is much more reliable than holdout for smaller datasets
- There are no “universal” optimal tuning parameters, they depend on dataset and preprocessing

Thank you for your attention!