Exercise 2 – Report

Notebook can also be found on Kaggle: https://www.kaggle.com/code/se21m003goll/excercise3-se21m003-goll-michael/edit/run/97638711

Description of Datasets

1. Stroke Dataset

The dataset contains data regarding the physical condition (gender, age, heart disease, bmi,...) of people, data about the context in which the people are living in (work type, residence) and a classification if the person was victim of a stroke or not (class-column). There are 24 features in total. The “stroke” column represents the classification. Possible values for the classification are true and false.

The bmi-feature contains N/A values. These values have been replaced by 0.

The gender, marital status (ever_married), work type and residency features contain categorical data in string-format. These values have been encoded (label-encoded or one-hot encoded in case of non-binary categories). All other data is in numerical format.

2. Breast Cancer

The dataset contains data regarding properties of a growth suspected to be cancer (interpretation of the features without knowing anything about the values and names of the features, since I am not a doctor!!!). There are 64 features in total. All of them are numerical. The “class” column represents the classification. Possible values for the classification are true and false.

In both datasets, the class-column has been label-encoded, because “true” and “false” values could not be processed by the classification algorithms. Also, the ID column has been excluded from processing in both datasets, since is does not represent a processable feature.

Software used for creating the solution

I used python as the programming language / environment. The Notebook has been created with Visual Studio Code. The Classifier-Implementations have been taken from the sklearn library.

Algorithms and Classifiers & Results

Classifier	Parameters	Dataset	Training time (seconds)	Test time (seconds)	Accuracy	Weighted F1
Decision Tree	min. samples/split: 100 min. samples/leaf: 1	Stroke	0,003	0,001	0,95498	0,93931
K-NN	neighbors: 3	Stroke	0,00102	0,03899	0,94905	0,93745
Decision Tree	min. samples/split: 50 min. samples/leaf: 1	Stroke	0,00399	0	0,94905	0,93745
K-NN	neighbors: 10	Stroke	0,001	0,05699	0,95735	0,93648
Perceptron	alpha: 0.1 penalty: l2	Stroke	0,001	0,001	0,95735	0,93648
Perceptron	alpha: 0.31622776601683794 penalty: l2	Stroke	0,002	0,001	0,95735	0,93648
Perceptron	alpha: 0.31622776601683794 penalty: l1	Stroke	0,003	0,001	0,95735	0,93648
Perceptron	alpha: 1.0 penalty: l2	Stroke	0,001	0,001	0,95735	0,93648
Perceptron	alpha: 1.0 penalty: l1	Stroke	0,00199	0,001	0,95735	0,93648
Perceptron	alpha: 3.1622776601683795 penalty: l2	Stroke	0,002	0,001	0,95735	0,93648
Perceptron	alpha: 3.1622776601683795 penalty: l1	Stroke	0,002	0	0,95735	0,93648
Perceptron	alpha: 10.0 penalty: l2	Stroke	0,002	0,001	0,95735	0,93648
Perceptron	alpha: 10.0 penalty: l1	Stroke	0,002	0,001	0,95735	0,93648
Decision Tree	min. samples/split: 2 min. samples/leaf: 50	Stroke	0,002	0,001	0,95735	0,93648
Decision Tree	min. samples/split: 2 min. samples/leaf: 100	Stroke	0,002	0,00101	0,95735	0,93648
Decision Tree	min. samples/split: 2 min. samples/leaf: 1000	Stroke	0,00099	0,001	0,95735	0,93648
Decision Tree	min. samples/split: 50 min. samples/leaf: 50	Stroke	0,003	0,001	0,95735	0,93648
Decision Tree	min. samples/split: 50 min. samples/leaf: 100	Stroke	0,002	0,001	0,95735	0,93648
Decision Tree	min. samples/split: 50 min. samples/leaf: 1000	Stroke	0,001	0,001	0,95735	0,93648
Decision Tree	min. samples/split: 100 min. samples/leaf: 50	Stroke	0,00301	0,00101	0,95735	0,93648
Decision Tree	min. samples/split: 100 min. samples/leaf: 100	Stroke	0,002	0,001	0,95735	0,93648

Decision Tree	min. samples/split: 100 min. samples/leaf: 1000	Stroke	0,00101	0	0,95735	0,93648
Decision Tree	min. samples/split: 1000 min. samples/leaf: 1	Stroke	0,002	0,00099	0,95735	0,93648
Decision Tree	min. samples/split: 1000 min. samples/leaf: 50	Stroke	0,002	0,001	0,95735	0,93648
Decision Tree	min. samples/split: 1000 min. samples/leaf: 100	Stroke	0,00201	0,001	0,95735	0,93648
Decision Tree	min. samples/split: 1000 min. samples/leaf: 1000	Stroke	0,00099	0,00101	0,95735	0,93648
SVC	SVC default	Stroke	0,01507	0,01393	0,95735	0,93648
Random Forest	n estimators: 100 max. feaatures: sqrt	Stroke	0,119	0,012	0,95616	0,93589
Random Forest	n estimators: 100 max. feaatures: log2	Stroke	0,12212	0,01212	0,95616	0,93589
K-NN	neighbors: 5	Stroke	0,001	0,047	0,95379	0,9347
Random Forest	n estimators: 20 max. feaatures: sqrt	Stroke	0,02524	0,003	0,95142	0,93351
Random Forest	n estimators: 20 max. feaatures: log2	Stroke	0,02561	0,00301	0,95142	0,93351
Decision Tree	min. samples/split: 2 min. samples/leaf: 1	Stroke	0,004	0,001	0,91706	0,92268
Random Forest	n estimators: 20 max. feaatures: log2	Breast Cancer	0,016	0,00202	0,91579	0,91522
Random Forest	n estimators: 100 max. feaatures: sqrt	Breast Cancer	0,07129	0,006	0,90526	0,9044
K-NN	neighbors: 3	Breast Cancer	0,00107	0,00317	0,89474	0,89443
Random Forest	n estimators: 100 max. feaatures: log2	Breast Cancer	0,06759	0,006	0,89474	0,89443
Random Forest	n estimators: 20 max. feaatures: sqrt	Breast Cancer	0,017	0,00306	0,89474	0,89402
Perceptron	alpha: 3.1622776601683795 penalty: l1	Breast Cancer	0,001	0	0,87368	0,87394

Perceptron	alpha: 0.31622776601683794 penalty: l1	Breast Cancer	0,001	0,001	0,87368	0,87368
Perceptron	alpha: 1.0 penalty: l1	Breast Cancer	0,001	0,001	0,87368	0,87331
K-NN	neighbors: 10	Breast Cancer	0,00102	0,00199	0,87368	0,87221
Decision Tree	min. samples/split: 2 min. samples/leaf: 1	Breast Cancer	0,00276	0,00102	0,86316	0,86297
SVC	SVC default	Breast Cancer	0,002	0,001	0,86316	0,8603
Decision Tree	min. samples/split: 100 min. samples/leaf: 1	Breast Cancer	0,002	0,001	0,85263	0,85309
Decision Tree	min. samples/split: 2 min. samples/leaf: 50	Breast Cancer	0,001	0,001	0,85263	0,85293
Decision Tree	min. samples/split: 50 min. samples/leaf: 50	Breast Cancer	0,002	0	0,85263	0,85293
Decision Tree	min. samples/split: 100 min. samples/leaf: 50	Breast Cancer	0,001	0,001	0,85263	0,85293
K-NN	neighbors: 5	Breast Cancer	0,001	0,002	0,85263	0,8522
Decision Tree	min. samples/split: 50 min. samples/leaf: 1	Breast Cancer	0,002	0,001	0,82105	0,81846
Perceptron	alpha: 0.1 penalty: l1	Breast Cancer	0,001	0,001	0,8	0,78597
Perceptron	alpha: 1.0 penalty: l2	Breast Cancer	0,002	0	0,55789	0,39957
Perceptron	alpha: 3.1622776601683795 penalty: l2	Breast Cancer	0,001	0	0,55789	0,39957
Perceptron	alpha: 10.0 penalty: l2	Breast Cancer	0,001	0,001	0,55789	0,39957
Perceptron	alpha: 10.0 penalty: l1	Breast Cancer	0,001	0,001	0,55789	0,39957

Decision Tree	min. samples/split: 2 min. samples/leaf: 100	Breast Cancer	0,001	0	0,55789	0,39957
Decision Tree	min. samples/split: 2 min. samples/leaf: 1000	Breast Cancer	0,001	0,001	0,55789	0,39957
Decision Tree	min. samples/split: 50 min. samples/leaf: 100	Breast Cancer	0,001	0	0,55789	0,39957
Decision Tree	min. samples/split: 50 min. samples/leaf: 1000	Breast Cancer	0,00199	0	0,55789	0,39957
Decision Tree	min. samples/split: 100 min. samples/leaf: 100	Breast Cancer	0,00101	0,00099	0,55789	0,39957
Decision Tree	min. samples/split: 100 min. samples/leaf: 1000	Breast Cancer	0,001	0	0,55789	0,39957
Decision Tree	min. samples/split: 1000 min. samples/leaf: 1	Breast Cancer	0,001	0,001	0,55789	0,39957
Decision Tree	min. samples/split: 1000 min. samples/leaf: 50	Breast Cancer	0,001	0,001	0,55789	0,39957
Decision Tree	min. samples/split: 1000 min. samples/leaf: 100	Breast Cancer	0,001	0	0,55789	0,39957
Decision Tree	min. samples/split: 1000 min. samples/leaf: 1000	Breast Cancer	0,001	0,001	0,55789	0,39957
Perceptron	alpha: 0.1 penalty: l2	Breast Cancer	0,001	0,001	0,44211	0,27107
Perceptron	alpha: 0.31622776601683794 penalty: l2	Breast Cancer	0,001	0	0,44211	0,27107
Perceptron	alpha: 0.1 penalty: l1	Stroke	0,003	0	0,12441	0,15617

Locally (see table above – orange marked rows), the best result has been achieved as follows

- **Stroke Dataset**: Decision Tree with parameters: min. samples/split: 100 min. samples/leaf: 1
- **Breast Cancer Dataset**: Random Forest Classifier with n-estimators: 20 and max features log2.

Comparison between local evaluation and Kaggle:

Regarding the stroke dataset, I only received a score of 0. I could not find the cause for this in time. The cancer dataset received a score of 0.96 which is higher than the locally achieved measures.