Case Study 1

Question 5

Your customer supplied a dataset with production data:

1. `f1`, `f2`: Data measured during production and aggregated into feature-values for each product
2. `error`: continous variable describing the quality for each product / set of feature values: small numbers indicate good quality

The data is stored in the CSV-File S:\Dozenten\jupyterhub\Data02.csv.

Subquestion 5.a **(/ 10)**

Upload the CSV-File to your Jupyter-Hub account for this exam, load the CSV-File and draw a histogram of the values of the `error` column!
In your opinion, do the data reflect a production with a high output of good quality parts? Justify your opinion!

Subquestion 5.b **(/ 5)**

Your customer requests a model to predict whether a part is good or bad from `f1` and `f2`. He specifies a threshold of `error`< 6 to identify good parts.
Apply this threshold to the data and give the number of good and bad parts contained in the data! Justify your answer!

Subquestion 5.c **(/ 5)**

Using the accuracy metric, give a baseline value for a minimum viable predictor! Justify your answer!

Subquestion 5.d **(/ 10)**

Is this dataset balanced? Briefly explain how you would train and evaluate your predictor for this dataset? (No code required) Justify your answers!

Case Study 2

Question 6

Your customer has provided you with the data in the file S:\Dozenten\jupyterhub\Data03.csv containing two feature columns. Each data point corresponds to one produced product. The timestamps reflect the time the production was finished.

Subquestion 6.a **(/ 10)**

Upload the CSV-File to your Jupyter-Hub account for this exam, load the CSV-File and analyze the timestamps! What timespan is covered by this data? Can you identify breaks? Justify your answers!

Subquestion 6.b **(/ 20)**

Perform a cluster analysis on the data!
How many data points did you find in each cluster?
Draw a scatter plot of `f1` and `f2` where the points are coloured according to the cluster they are associated with!
For the smallest cluster, give a list of the timestamps for the data points in the cluster!

Total points (/ 90)