1 Disclaimer

1.1 This document contains privileged information as property of GfK SE. Note that unauthorized copying, disclosure or distribution of the material in this document is not permitted.

2 Python / Distributed systems

Dataset N: Dataset_N.csv
Dataset Description:
Dataset_N is supposed to contain records with article texts and the belonging product group. The following information is known about the columns:

Column	Info
id	A unique record identifier
product group	Product category
main_text	a describing text about the article
add_text	an additional describing text about the article
manufacturer	the manufacturer belonging to the article

2.1 Tasks:

Data Prep:
Unfortunately, it happened that during the data generation process the column names have been mixed up 🙂

1. Use PySpark to import and modify the data accordingly to a schema as described above.

Modeling:

1. Create a machine learning model in order to predict the product category based on appropriate columns. Use scikit-learn with one or more machine learning algorithms.
2. Present the result in a vivid way (e.g. Jupyter) and explain your model from a statistical PoV.

Productizing:

1. Create a pod (Application & Webserver) in Kubernetes or Minikube. If you are not familiar with K8s, create isolated containers
2. Create a ML module in Python with the ability to predict the product category based on appropriate columns. (Train your model based on "Dataset_N")
3. Create a simple HTTP REST-API on top of your ML module that takes "X" as parameter for the request and responds with prediction "Y"
4. Augment your containers and serve the applications with HTTPS via Nginx
5. Commit all code & results to a local Git repository. (Note: Only the git repository will be considered as valid submission!)