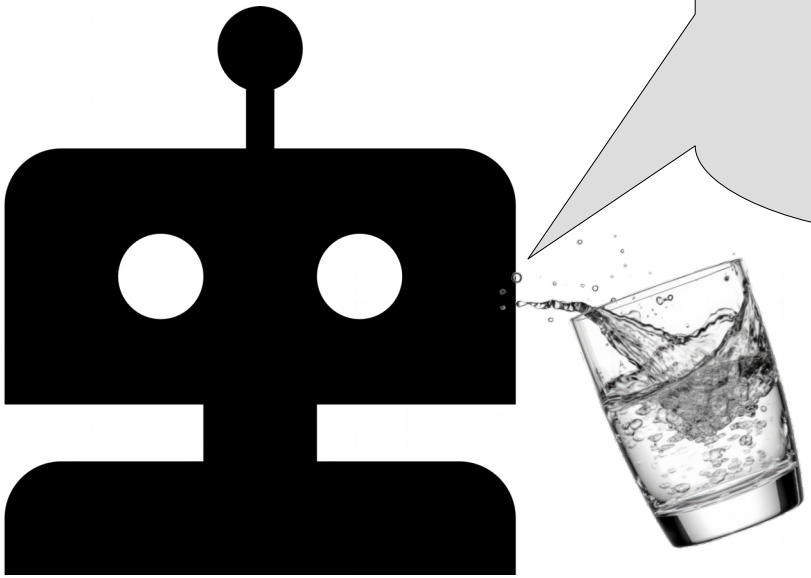
Water Conneisseur



Mmhh… **tasty…**

but besides the fact of the water's **hardness** being able to **reduce the lifespan of your brand-new coffee machine**, the amount of **trihalomethanes** could cause little side-effects inside your delicate human flesh such as **bladder cancer**.

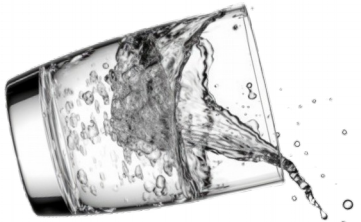
I'd **pour** this liquid **away**...

Water Conneisseur

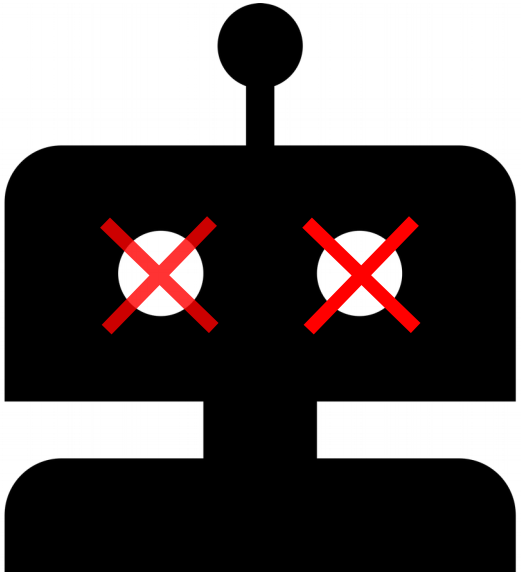
Mmhh… **tasty…**

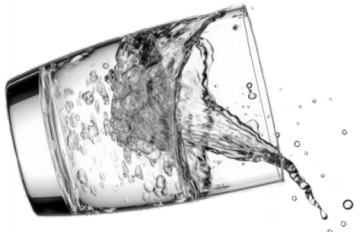
but besides the fact of the water's **hardness** being able to **reduce the lifespan of your brand-new coffee machine**, the amount of **trihalomethanes** could cause little side-effects inside your delicate human flesh such as **bladder cancer**.

I'd **pour** this liquid **away**...



Water Conneisseur





Water Quality

1516 | New Notebook

Data Card | Code (552) | Discussion (23) | Suggestions (0)

water_potability.csv (525.19 kB)

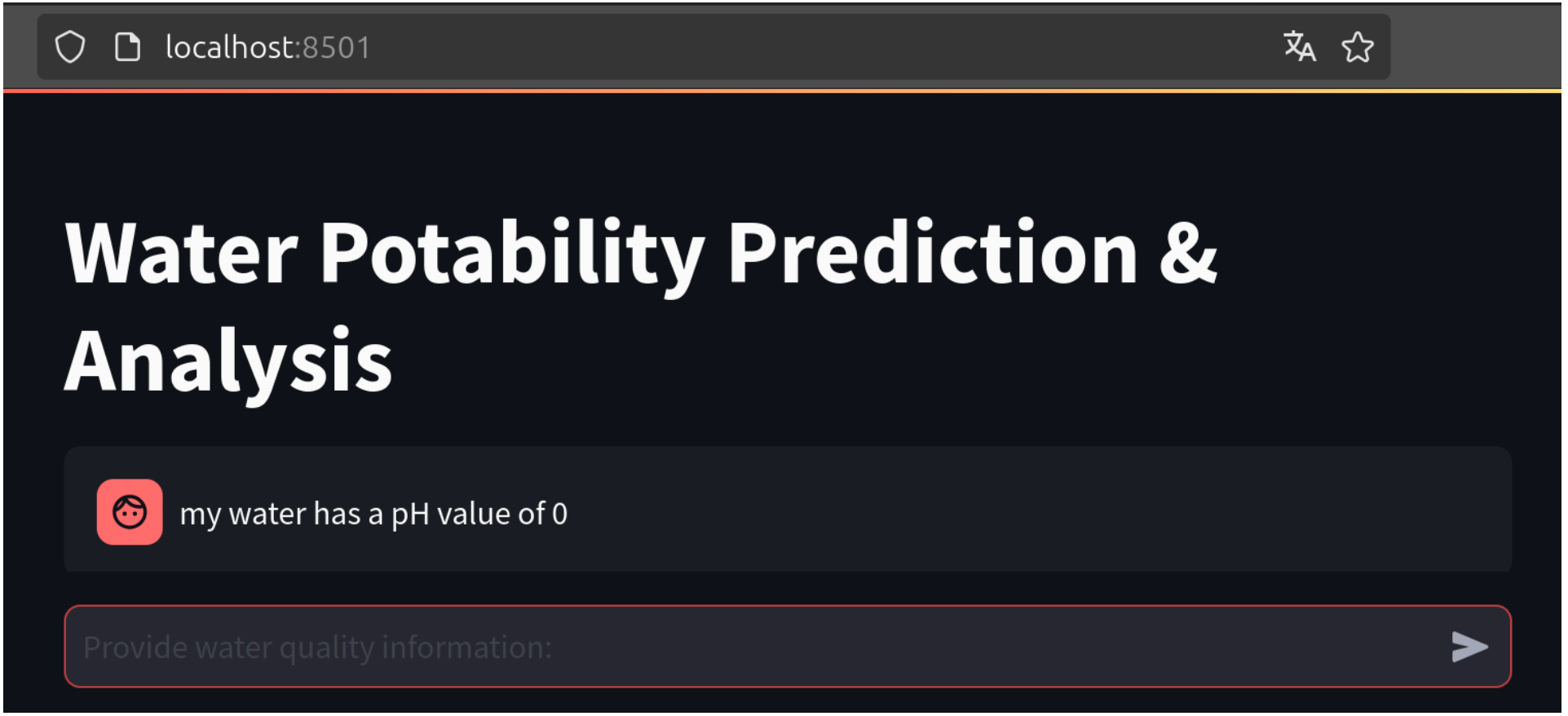
Detail | Compact | Column | 10 of 10 columns

~~µg/L: microgram per litre~~
mg/L: milligram per litre

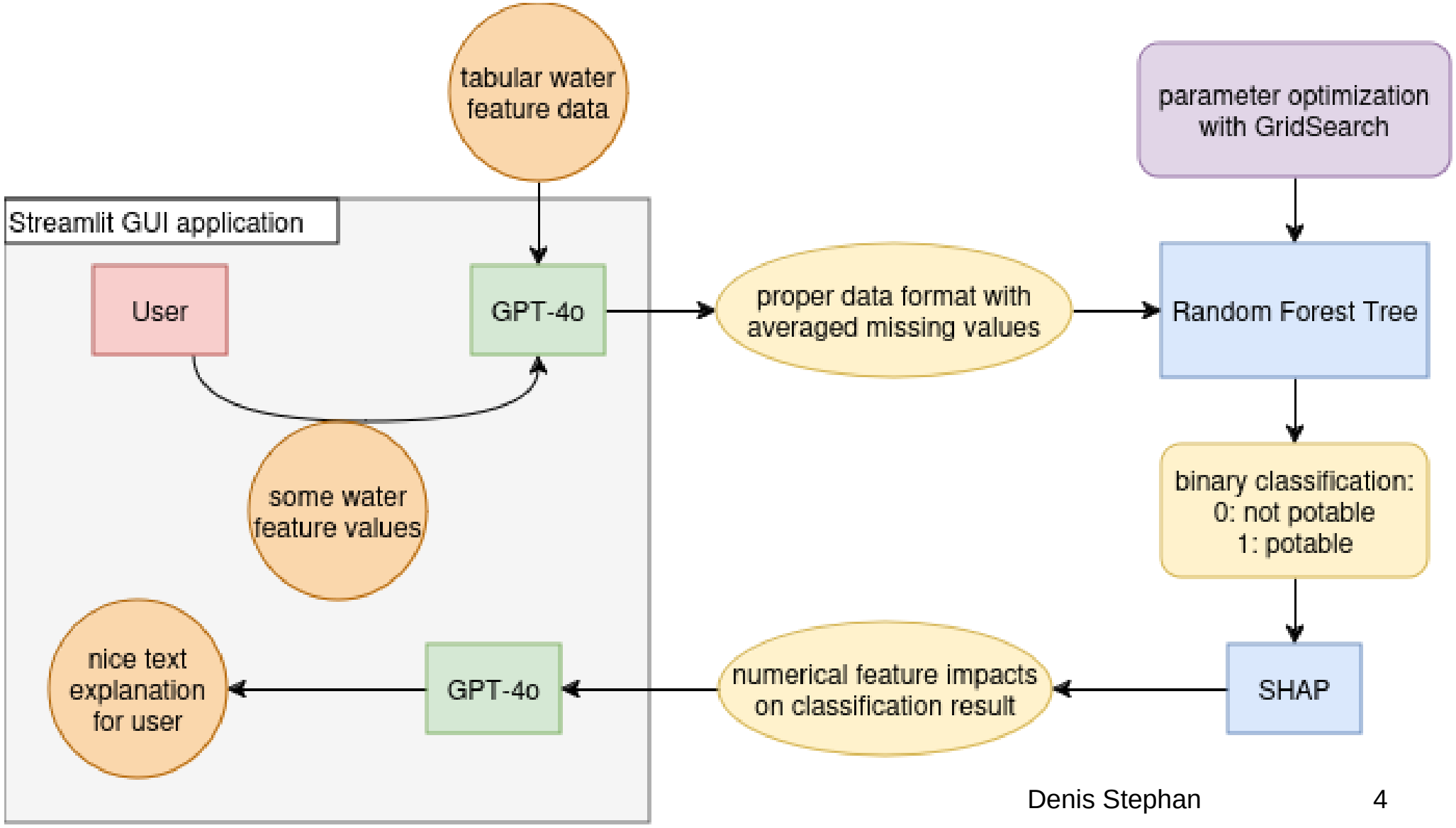
Column description:

1. ph: pH of 1. water (0 to 14).
2. Hardness: Capacity of water to precipitate soap in mg/L.
3. Solids: Total dissolved solids in ppm.
4. Chloramines: Amount of Chloramines in ppm.
5. Sulfate: Amount of Sulfates dissolved in mg/L.
6. Conductivity: Electrical conductivity of water in µS/cm.
7. Organic_carbon: Amount of organic carbon in ppm.
8. Trihalomethanes: Amount of Trihalomethanes in µg/L.
9. Turbidity: Measure of light emiting property of water in NTU.
10. Potability: Indicates if water is safe for human consumption. Potable -1 and Not potable -0

# ph pH of water	# Hardness Capacity of water to precipitate soap in mg/L	# Solids Total dissolved solids in ppm	# Chloramines Amount of Chloramines in ppm	# Sulfate Amount of Sulfates dissolved in mg/L	# Conduc Electrical c water in µS
0 – 14	47.4 – 323	321 – 61.2k	0.35 – 13.1	129 – 481	181
	204.8904554713363	20791.318980747026	7.300211873184757	368.51644134980336	564.30865
3.71608007538699	129.42292051494425	18630.057857970347	6.635245883862		592.88535
8.099124189298397	224.23625939355776	19909.541732292393	9.275883602694089		418.60621
8.316765884214679	214.37339408562252	22018.417440775294	8.05933237743854	356.88613564305666	363.26651
9.092223456290965	181.10150923612525	17978.98633892625	6.546599974207941	310.13573752420444	398.41081



→ LLM should format user data
to be used as input for model prediction


tabular water feature data
parameter optimization with GridSearch
Streamlit GUI application
User
GPT-4o
proper data format with averaged missing values
Random Forest Tree
some water feature values
binary classification: 0: not potable 1: potable
nice text explanation for user
GPT-4o
numerical feature impacts on classification result
SHAP

Constraints:

- Averaged missing values **too pessimistic in Germany**
  (→ not to be taken seriously, just as **Proof of Concept**)
- **Minimize false positives**
- **SHAP** calculation not required in case of result == potable
- GPT-4o should distinguish in explanation of final decision:
  user provided values vs kaggle database
  (positive / negative influences each)
- **Feedback** questions of users are **out of scope**
  → next message = new request