Benchmark Datasets for Courseworks 1—3

Datasets requiring **time series prediction (e.g. House Price, Interest Rates)** must use **past data only for training** to avoid leakage.

The datasets paired with a task are as follows:

The datasets are as follows:

- **Bank-full**
  https://www.kaggle.com/datasets/nimishsawant/bankfull/data
  Predict whether a person will default.
- **Predict Age Abalone**
  https://www.kaggle.com/datasets/farkhod77/abalone-age-predict/data
  Predict the age of abalones.
- **Telco Customer Churn**
  https://www.kaggle.com/datasets/blastchar/telco-customer-churn/data
  Predict whether a customer will churn.
- **Stroke Prediction Dataset**
  https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/code
  Predict the stroke occurrence given a person's profile.
- **Synthetic Financial Datasets for Fraud Detection**
  https://www.kaggle.com/datasets/ealaxi/paysim1/code?datasetId=1069&sortBy=voteCount
  Predict whether a payment is fraudulent.
- **Water Quality**
  https://www.kaggle.com/datasets/adityakadiwal/water-potability/data
  Predict whether a water body is safe for drinking (labelled potability in the dataset).
- **World Happiness Report 2021**
  https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021/data
  Predict healthy life expectancy given other geographical and socioeconomic variables.
- **USA House Price Prediction ML Analysis**
  https://www.kaggle.com/code/sahilislam007/usa-house-price-prediction-ml-analysis
  Predict the fall/rise of housing prices one month in advance.

  ⚠️ This is a **binary classification task** requiring you to convert price data into rise/fall labels.

  ⚠️ For true predictive analysis, you must **not use data from the month being predicted or any future months**. Only past data should be used for training and validation.

- **Federal Reserve Interest Rates, 1954–Present**
  https://www.kaggle.com/datasets/federalreserve/interest-rates
  Predict real GDP change rate for the next quarter.
  ⚠️ For true predictive analysis, you must **not use data from the quarter being predicted or any future quarters**. Only past data should be used for training and validation.