Cleaning Data

Quacky has gathered a dataset of many ducks for his Machine Learning model. But when he looks closely, the wise owl points out that the data is not always perfect. Real-world data often has mistakes, missing information, or strange values.
Before a machine can learn well, the dataset must be cleaned. Quacky learns that data preprocessing is like preparing clean and polished information for the model.

Problems Quacky Finds in Real Data

lThe dataset looks complete at first, but several issues appear on closer inspection:

- **Missing values**

Some ducks do not have their swimming speed recorded.
Without this information, the model gets confused.

- **Wrong or impossible values**

Quacky finds a duck listed as being negative five years old.
This is clearly incorrect and would mislead the model.

- **Duplicate entries**

The same duck appears twice by accident.
This would give that duck extra importance in the learning process.

- **Outliers**

One duck shows a swimming speed far beyond reality, such as 500 meters per second.
These unusual numbers can distort learning and cause inaccurate predictions.

Quacky realizes that poor data leads to poor learning.

How Quacky Cleans the Data

To create a reliable dataset, Quacky follows some careful steps:

- **Fill or remove missing values**

If speed is missing, he can estimate based on similar ducks, or remove that entry if it provides no useful information.

- **Correct or delete impossible values**

Negative ages or any physically impossible data are fixed or removed to maintain logic.

- **Remove duplicates**

Every duck must be counted once, not multiple times.

- **Check and handle outliers**

Extremely unusual values are reviewed. If there are errors, they are removed; if they are real, they are recorded carefully.

By applying these fixes, Quacky makes sure the data is clean, consistent, and reliable.

How Quacky Cleans the Data

To create a reliable dataset, Quacky follows some careful steps:

- **Fill or remove missing values**

If speed is missing, he can estimate based on similar ducks, or remove that entry if it provides no useful information.

- **Correct or delete impossible values**

Negative ages or any physically impossible data are fixed or removed to maintain logic.

- **Remove duplicates**

Every duck must be counted once, not multiple times.

Quacky's Analogy

Quacky imagines himself checking shiny treasures before selling them at the market:

- Fake coins are thrown away
- Dirty gems are polished
- Real treasure is carefully organized

Just as gems must be prepared before being sold, data must be cleaned before training a model.

Miss Hootsworth

Final Simple Definition

Clean data is the foundation for a successful machine learning model. It removes mistakes and ensures the model learns from correct information only.