Machine Learning – Day 4 Notes

Why Data Quality Matters More Than Data Quantity

---

1. Initial Assumption

I believed that **better Machine Learning results require more data**.

This assumption turned out to be incomplete and often misleading.

2. Key Realization

More data **does not automatically improve model performance**.

What matters far more is:

- How accurate the data is
- How relevant it is to the problem
- Whether it represents reality correctly

Data quality has a greater impact than data quantity.

3. Understanding Data Quality in Machine Learning

A. Noisy Data

Noisy data contains:

- Errors
- Random variations
- Inconsistent or incorrect values

Impact:

- True patterns become harder to detect
- Models learn randomness instead of signal

Real-life example:

- Sensor data with faulty readings
- User-entered data with spelling mistakes or incorrect values

Insight:
Noise hides meaningful patterns, regardless of dataset size.

B. Biased Data

Biased data represents the world **unevenly or unfairly**.

Impact:

- Models learn incorrect relationships
- Predictions become unreliable or unfair

Real-life example:

- Hiring models trained on historically biased recruitment data
- Credit risk models trained on limited demographic groups

Insight:
Models only learn what data shows — not what is true.

C. Irrelevant Data

Irrelevant data does not contribute to solving the target problem.

Impact:

- Adds complexity without improving understanding
- Can increase confidence while reducing real accuracy

Real-life example:

- Including a user's device color when predicting loan approval
- Adding unrelated website metrics to sales prediction

Insight:
More features do not mean more information.

4. Small Clean Data vs Large Messy Data

A **small, well-curated dataset** can outperform a large dataset when:

- Labels are accurate
- Features are meaningful
- Noise and bias are controlled

Why this happens:

- Clean data provides clearer learning signals
- Models generalize better with trustworthy inputs

5. Shift in Perspective

Earlier Thinking

- More data → better model
- Focus on collecting as much data as possible

Updated Thinking

- Better data → better learning
- Focus on:
    - Data cleaning
    - Data validation
    - Data relevance

6. Practical ML Mindset Change

Before asking:

“Which model should I use?”

It is more important to ask:

“Can this data be trusted?”

Data understanding comes **before** model selection.

7. Final Takeaway

Machine Learning performance depends heavily on data quality.

Clean data → Clear patterns → Reliable models

Large datasets cannot compensate for:

- Noise
- Bias
- Irrelevance

In Machine Learning,
quality creates accuracy — not quantity.