🧠 THE MACHINE LEARNING MINDSET

1. Problem First, Not Algorithm First

“What question am I trying to answer?”

- **Frame the problem**: Is it classification or regression?
- **Understand the domain**: Here, you're dealing with **health data** → think ethically, consider sensitivity, false positives/negatives.
- **Target column**: What are you trying to predict? (`Class/ASD` in this case)

---

2. Data Understanding (EDA) Is Critical

“What does the data say? What’s weird or interesting?”

- Check:
  - **Missing values**
  - **Imbalanced classes**
  - **Distribution of values**
  - **Data types**
  - **Outliers or noise**
- Look for **signal vs. noise**. In this autism dataset, the questionnaire scores may carry strong signal.

💡 Mindset tip: *Let data exploration drive your decisions.*

---

3. Ask: What Needs to Be Cleaned or Converted?

“Can this be fed to a model as is?” → Usually no.

- Label encode categories (`yes`/`no`, `gender`, `ethnicity`)
- Normalize or scale numeric data (e.g., results, age)
- Impute or drop missing values
- Decide: do I balance the dataset? Do I need synthetic sampling?

💡 Think of this as preparing raw ingredients before cooking.

---

4. Feature Thinking

> *"What are the most relevant pieces of data? Can I create better ones?"*

- Use domain knowledge to engineer new features.
- Use correlation plots, domain logic, or feature importance to reduce or prioritize features.
- Be skeptical of features like `ID` — they might leak or confuse.

💡 Rule of thumb: **better features beat fancier algorithms.**

---

5. Modeling With Intention

> *"What model fits this data + problem + resources?"*

- For **tabular classification** (like here):
  - Start with Logistic Regression (baseline)
  - Try tree-based models (e.g. XGBoost)
  - Consider SVMs if data is small and clean
- Compare models fairly using the **same train-test split** and metrics.

💡 Don't just chase accuracy — also check:

- **Precision** (if false positives matter),
- **Recall** (if false negatives are worse),
- **F1-score** (balance of both)

---

6. Validate Like a Scientist

> *"How do I know my model generalizes?"*

- Always split your data: **train/test** (maybe cross-validation too)
- Avoid data leakage
- Use **stratified splits** if the target class is imbalanced

💡 Mindset: *Don't trust your model until it's passed a blind test.*

---

7. Explainability & Trust

> *"Why did the model make this decision?"*

- Use:
  - **Feature importance** (tree models)
  - **SHAP / LIME** for local explanations
  - Clear visualizations
- Especially important in health/personal domains like autism diagnosis

---

8. What Next? Deployment or Iteration

> *"Is this useful to someone?"*

- Save models
- Create a dashboard or report
- Set up retraining pipeline (if data will change)

💡 Always circle back: **Did this help answer the original question?**

⚙️ Summary Cheat Sheet

Phase	Key Questions
Understand	What's the goal? Who benefits? What is success?
Explore	What does the data look like? Any patterns or issues?
Clean	What's missing, dirty, or needs converting?
Feature	What makes a good input? Anything new I can create?
Model	Which model makes sense? What tradeoffs am I accepting?
Validate	Am I overfitting? Does this generalize?
Explain	Can I justify my predictions to others?
Deliver	Is this usable by someone else? Can it be improved later?