# Make Data Count – Flowchart (Non-Technical)

A simple, colored flow for non-technical readers. The diagram below is followed by an editable table version.

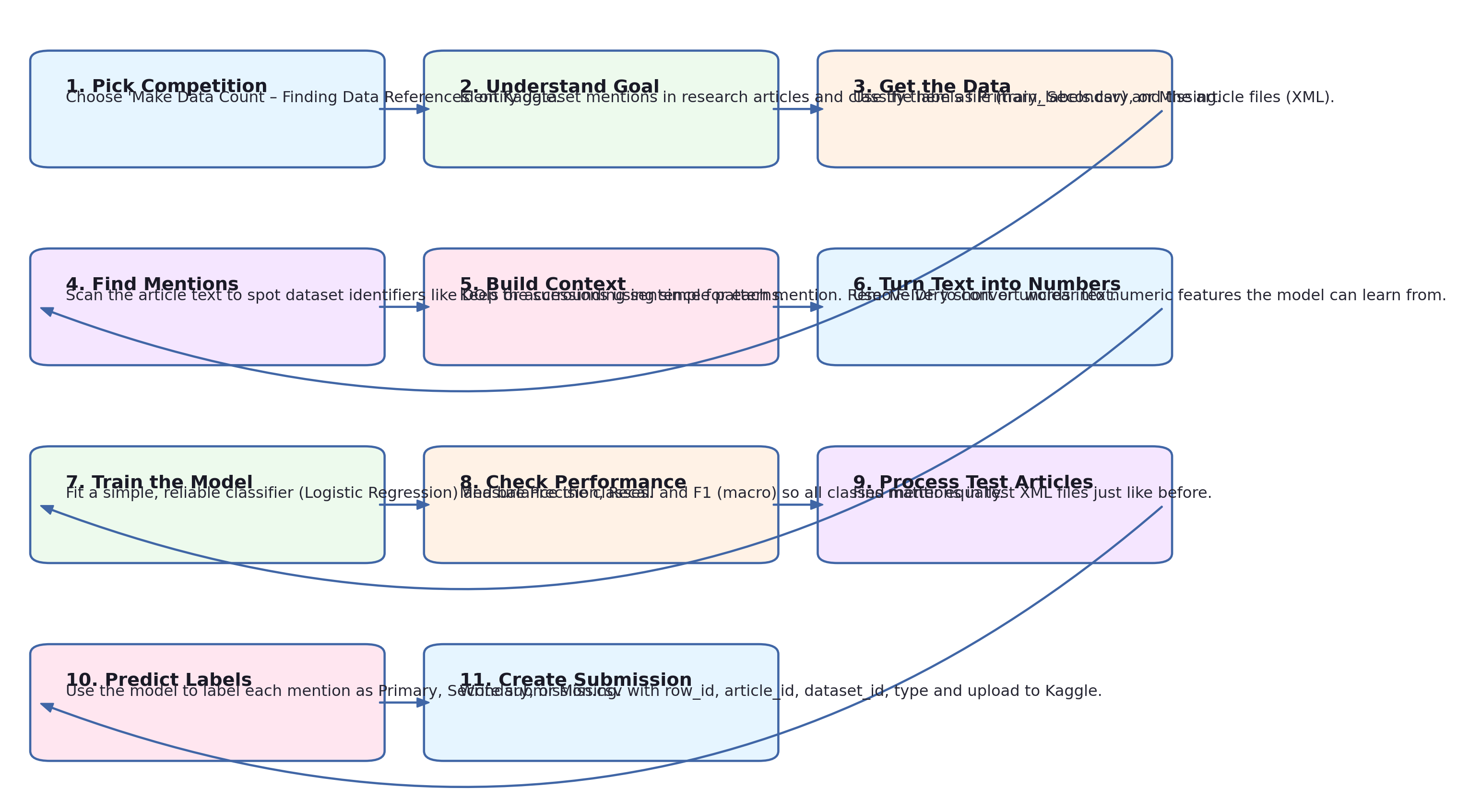


Figure 1: End-to-end process at a glance.

## Editable Flow (Table Version)

|  |  |  |
| --- | --- | --- |
| **Step** | **What Happens** | **Why It Matters (Plain English)** |
| 1 | Pick Competition | Choose 'Make Data Count – Finding Data References' on Kaggle. |
| 2 | Understand Goal | Identify dataset mentions in research articles and classify them as Primary, Secondary, or Missing. |
| 3 | Get the Data | Use the labels file (train\_labels.csv) and the article files (XML). |
| 4 | Find Mentions | Scan the article text to spot dataset identifiers like DOIs or accessions using simple patterns. |
| 5 | Build Context | Keep the surrounding sentence for each mention. Remove very short or unclear text. |
| 6 | Turn Text into Numbers | Use TF‑IDF to convert words into numeric features the model can learn from. |
| 7 | Train the Model | Fit a simple, reliable classifier (Logistic Regression) and balance the classes. |
| 8 | Check Performance | Measure Precision, Recall and F1 (macro) so all classes matter equally. |
| 9 | Process Test Articles | Find mentions in test XML files just like before. |
| 10 | Predict Labels | Use the model to label each mention as Primary, Secondary, or Missing. |
| 11 | Create Submission | Write submission.csv with row\_id, article\_id, dataset\_id, type and upload to Kaggle. |

Tip: You can edit this table directly in Word or copy it to slides.