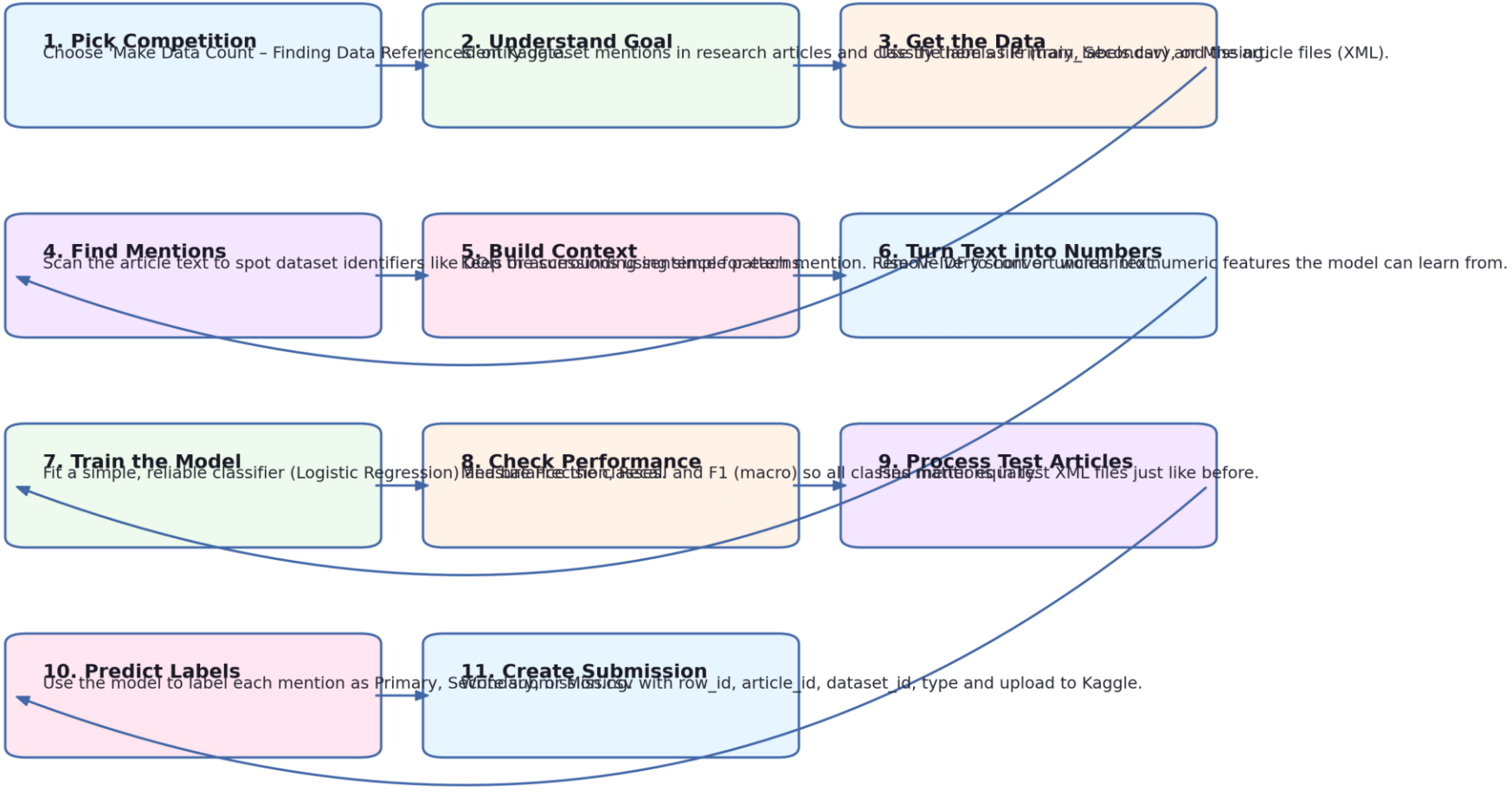
1. Pick Competition
Choose 'Make Data Count – Finding Data References' on Kaggle.
2. Understand Goal
Classify dataset mentions in research articles and identify their source.
3. Get the Data
Load the labels file (train_labels.csv) and the article files (XML).
4. Find Mentions
Scan the article text to spot dataset identifiers like DOIs or accession numbers.
5. Build Context
Keep the sentences surrounding each mention.
6. Turn Text into Numbers
Convert the context words into numeric features the model can learn from.
7. Train the Model
Fit a simple, reliable classifier (Logistic Regression) that balances precision and recall.
8. Check Performance
Measure accuracy, precision, recall, and F1 (macro) so all classes matter equally.
9. Process Test Articles
Apply the same steps to the 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 a CSV with row_id, article_id, dataset_id, type and upload to Kaggle.

Flowchart Legend & Steps

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.