we’re building a machine learning model to **predict colorectal cancer (CRC)** based on the **gut microbiome composition** of patients, using **16S rRNA sequencing data** collapsed at the genus level.

1st step : Load the data:

* metadata.csv: Includes SampleID and whether the patient has CRC.
* seqtab\_nochim.csv: Each row = a sample; each column = an ASV (amplicon sequence variant); values = counts.
* taxa.csv: ASVs matched to their taxonomic classification (e.g., Genus, Family, etc.).

2nd step: Map ASVs to Genera

* Maps each ASV to its genus using the taxa table.
* If a genus is missing, it's labeled "Unassigned".
* The ASV table’s columns are renamed from ASV IDs to genus names.

"Instead of calling this column ASV\_1234, let’s call it *Bacteroides*, because that’s the genus it belongs to."

3rd Step: Collapse to Genus level :

Some ASVs belong to the same genus, so this combines all columns with the same genus name. we’re summing their counts per sample

4th step : Join with metadata

Combines CRC status with genus abundance for each sample in a file named seqtab\_genus

A screenshot of a computer

AI-generated content may be incorrect.

**Each row** = a single **sample** (a patient's stool sample).  
**Each column** = a **bacterial genus** (like *Bacteroides*, *Prevotella*, etc.).  
**Each value** = how many DNA sequences (reads) were assigned to that genus.

1. We start with seqtab, which has:
   * Rows = samples
   * Columns = ASVs (DNA sequences)
   * Values = counts of each ASV per sample
2. We map each ASV to its **Genus name** (from taxa.csv)
3. Then we **group** all ASVs belonging to the same genus and **sum** their counts.

So seqtab\_genus **=** Our **ASV table**, converted to the **genus level**, with counts summed across all ASVs that belong to each genus.

5th step: Normalize the Microbiome Data

* Converts raw read counts to **relative abundances** (percentage per sample).
* Prepares X (features) and y (labels) for training.

6 Split the dataset : Splits into training and test sets (80/20 split).

7th step : Train a random forest classifier :

* A random forest learns how microbial genera relate to CRC status.
* It builds many decision trees and combines their outputs.

8 Evaluate model performance:

* Outputs precision, recall, F1-score, and accuracy.
* Tells you how well your model is classifying CRC vs non-CRC.

9 identify important genera :

Shows which **genera are most predictive** of CRC status according to the model.