Inter-IIT Round 2: Synthetic Dataset Generation

A Simple Approach Using Generator Fusion

Team_94 Solution

1 What We Did (Summary)

Problem: We have 5 generators (G1-G5). Each gives us only SOME columns. We need ALL 12 columns in one dataset.

Our Solution:

1. Pick G3 as the "base" (it has the most connections to other generators)
2. Find matching rows from other generators using shared columns
3. Combine everything into one final dataset with 1000 rows

Result: 1000 rows × 12 columns (exactly what was required)

2 How We Connected the Generators

Each generator shares some columns with others. We use these shared columns to find matching rows.

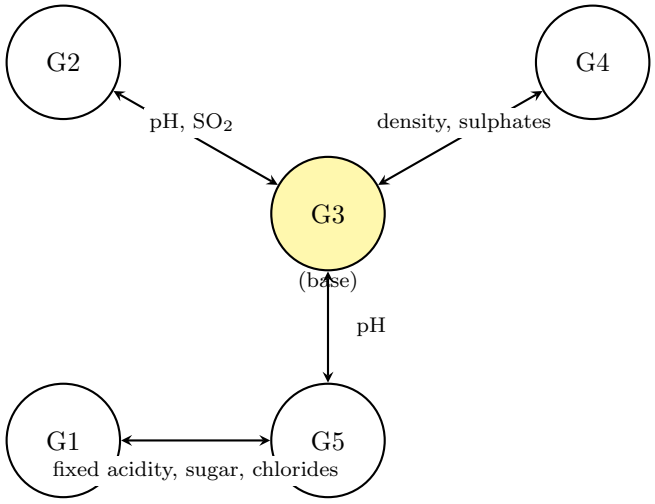

Figure 1: How generators are connected through shared columns

3 Step-by-Step Process

3.1 Step 1: Start with G3

We generate 1000 rows from G3. This gives us 5 columns:

free SO_2, total SO_2, density, pH, sulphates

3.2 Step 2: Get alcohol and quality from G4

G3 and G4 both have "density" and "sulphates". So:

- For each G3 row, find the G4 row with the closest density and sulphates
- Take "alcohol" and "quality" from that G4 row

3.3 Step 3: Get fixed acidity, sugar, chlorides from G5

G3 and G5 both have "pH". So:

- For each G3 row, find the G5 row with the closest pH
- Take "fixed acidity", "residual sugar", "chlorides" from that G5 row

3.4 Step 4: Get volatile acidity and citric acid from G1 and G2

- G2 shares columns with G3 → match G2 to G3
- G1 shares columns with G5 → match G1 to G5 (which is already matched to G3)
- Take volatile acidity and citric acid from both, average them

4 Making the Data More Realistic

We added some wine chemistry rules:

Rule 1: pH depends on acidity

- More acid → lower pH
- We trained a small model to predict pH from acidity values
- Final pH = mix of predicted pH and matched pH

Rule 2: Density depends on sugar and alcohol

- More sugar → higher density
- More alcohol → lower density
- We trained a model for this too

Rule 3: Free SO_2 must be less than Total SO_2

- This is always true in real wine
- If any row violates this, we fix it

5 Better Matching (Not Just 1 Neighbor)

Instead of finding just 1 closest match, we:

Find 5 closest rows
Average their values
(This is smoother and more reliable)

6 Where Each Column Comes From

Column	Source
free sulfur dioxide	G3 (base)
total sulfur dioxide	G3 (base)
density	G3 + chemistry model
pH	G3 + chemistry model
sulphates	G3 (base)
alcohol	G4 (matched)
quality	G4 (matched)
fixed acidity	G5 (matched)
residual sugar	G5 (matched)
chlorides	G5 (matched)
volatile acidity	G1 + G2 (averaged)
citric acid	G1 + G2 (averaged)

Table 1: Final column sources

7 Note

For reproducibility, please keep the file structure as:
`/kaggle/input/round-2-inter-iit/round_2_inter_iit/Round_2_Inter_IIT`