1. GMM (Gaussian Mixture Model)

Most suitable.

- This is because the data *exactly follows* the assumption GMM makes: each class is generated from a Gaussian distribution.
- GMM models data as a mixture of Gaussians, estimating the mean and variance of each.
- Since the diabetic and non-diabetic groups are separated (80 vs 220 mean), GMM should **easily identify and separate** the two distributions.
- It even works better than K-means in 1D when variances differ (20 vs 50 here).

2. K-means

- **Works decently but not ideal.**
- It minimizes **Euclidean distance to centroids**, so it assumes equal spherical variance — which isn't true here (20 vs 50).
- It still classify reasonably well due to the large separation, but **boundary is suboptimal**.

3. DBSCAN

- **Not ideal here.**
- DBSCAN is designed for **density-based clustering**. In 1D, and especially with different variances, it's sensitive to parameters like `eps` and `min_samples`.
- It potentially **misclassify data at the boundary or treat sparse regions as noise**.
- It also assumes clusters have high density regions — which can be skewed with different standard deviations.

Conclusion:

- **GMM > K-means >> DBSCAN**

Variation of threshold with number of samples?
--> Yes. The threshold varies with the number of samples. But the variation is higher for DBSCAN and K-Means (DBSCAN is highly unstable with the sample size) while it is under control in the case of GMM.