----NEW POSSIBLE PROBLEMES

Sure, here are examples of regression, classification, and clustering problems using the ChEMBL dataset:

1. Regression Problem:

Goal: Predict the bioactivity value (e.g., IC50, EC50, Ki) of compounds for a specific target protein.

Features: Molecular descriptors, chemical fingerprints, physicochemical properties (e.g., molecular weight, LogP).

Target: Continuous bioactivity value (IC50, EC50, etc.).

Approach:

Use regression algorithms like Random Forest Regressor, XGBoost, or Neural Networks.

Evaluate using metrics like RMSE or R².

Example: Predict the IC50 value of kinase inhibitors from their molecular fingerprints.

| **Type** | **Columns** | **Notes** |
| --- | --- | --- |
| **Target** | standard\_value  pchembl\_value | Numeric bioactivity values; choose one as target |
| **Numeric Features** | mw\_freebase  alogp  hba  hbd | Molecular descriptors |
| **Categorical Features** | standard\_type  standard\_units  assay\_type  target\_name  target\_type  mesh\_heading efo\_term | May require encoding (e.g., one-hot) |
| **Text / Sequence** | assay\_description canonical\_smiles | Usually need special processing / embeddings |
| **Quality / Confidence** | confidence\_score | Useful for filtering or weighting |
| **Identifiers (not features)** | activity\_id  assay\_id  molregno  component\_id | For joining or grouping, usually excluded from model |

2. Classification Problem:

Goal: Classify compounds as "active" or "inactive" against a specific biological target.

Features: Molecular structure, chemical fingerprints, bioactivity data.

Target: Binary label (1 for active, 0 for inactive), based on a threshold (e.g., IC50 < 10 µM is active).

Approach:

Use classification algorithms like Logistic Regression, SVM, or GNN.

Evaluate with metrics like accuracy, F1-score, and AUC-ROC.

Example: Classify kinase inhibitors as active or inactive based on their structural features.

-----------------------------

3. Clustering Problem:

Goal: Group compounds based on chemical similarity or bioactivity profiles.

Features: Molecular fingerprints (e.g., ECFP, MACCS), physicochemical properties.

Approach:

Use clustering algorithms like K-Means, DBSCAN, or hierarchical clustering.

Visualize clusters using dimensionality reduction (e.g., PCA or t-SNE).

Example: Cluster small molecules to identify chemical scaffolds or group compounds with similar bioactivity profiles.