Project Deliverable 1: Dataset Selection and Justification

Project Title:
AI-Powered Chemical Compound Discovery Recommender using HSNW

Dataset Links

- **Hugging Face:** https://huggingface.co/datasets/kashaf-nadeem/SMILES_Big_Data_Set
- **GitHub Repository:** https://github.com/nashrah692/AI-Powered-Chemical-Compound-Discovery-Recommender-using-HSNW

Team Contribution

- **Nashrah (22F-3156):**
  - Researched and selected the dataset relevant to chemical compound recommendation.
  - Uploaded the dataset to GitHub.
- **Kashaf (22F-3095):**
  - Analyzed the structure of the dataset and identified key features and labels.
  - Uploaded the dataset to Hugging Face.
  - Wrote the dataset description and justification.
- **Abeera (22F-3158):**
  - Assisted in preparing this deliverable report.

Dataset Description

- **Source:**
  The dataset was originally obtained from Kaggle (https://www.kaggle.com/datasets/yanmaksi/big-molecules-smiles-dataset).

- **Name:**
  SMILES_Big_Data_Set.csv

- **Size:**
  The dataset contains **16,087** entries representing different chemical compounds and their associated properties. It is approximately **1.7 MB** in size.

- **Relevance to the Project:**
  This dataset is highly relevant to our project goal of building an AI-powered chemical compound discovery system using HSNW (Hierarchical Navigable Small World graphs). The SMILES (Simplified Molecular Input Line Entry System) format encodes molecular structures in a linear string, making it suitable for vectorization and recommendation systems. By leveraging SMILES data, we aim to efficiently recommend or discover new compounds with similar properties or structures based on molecular similarity.

- **Structure:**
  The dataset includes the following columns:

  - **Molecule**: The name of the molecule (or an ID)
  - **SMILES**: The SMILES string representing the molecular structure
  - **MolecularWeight**: The molecular weight of the compound
  - **TPSA**: Topological Polar Surface Area
  - **LogP**: Octanol-water partition coefficient (a measure of hydrophobicity)
  - **NumHDonors**: Number of hydrogen bond donors
  - **NumHAcceptors**: Number of hydrogen bond acceptors

These features allow us to compute similarity between compounds and train a recommender engine that can suggest novel chemical structures based on a query compound's profile.