COMPOUND SIMILARITY PREDICTION

# A PROJECT REPORT-2

Submitted in partial fulfillment of the requirements for the award of the degree of

## Bachelor of Technology

*in*

COMPUTER SCIENCE AND ENGINEERING

### BY

**M.JYOTHSNA K. ASHOK KUMAR**

**Roll No: 18331A05A2 Roll No: 19335A0509**

**P. SAI JAHNAVI S.NIKHILA**

**Roll No: 18331A05B5 Roll No: 18331A05B3**

**Under the Supervision of**

**Dr. P. SRINIVASA RAO**

**Associate Professor**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING MVGR COLLEGE OF ENGINEERING (Autonomous) VIZIANAGARAM-535005, AP (INDIA)**

**(Accredited by NBA, NAAC, and Permanently Affiliated to Jawaharlal Nehru Technological University Kakinada)**

# CERTIFICATE



This is to certify that the project report entitled **“Compound Similarity Prediction”** being submitted by **M.JYOTHSNA, K.ASHOK KUMAR, P. SAI JAHNAVI, S.NIKHILA** bearing registered numbers 18331A05A2, 19335A0509, 18331A05B5, 18331A05B3 respectively, in partial fulfillment for the award of the degree of “**Bachelor of Technology” in Computer Science and Engineering** is a record of Bonafide work done by them under my supervision during the academic year 2021-2022.

**HOD CSE Supervisor**

**Dr. P. Ravi Kiran Varma Dr. P. Srinivasa Rao**

**Associate Professor Associate Professor**

**Department of CSE Department of CSE**

**MVGR College of Engineering MVGR College of Engineering**

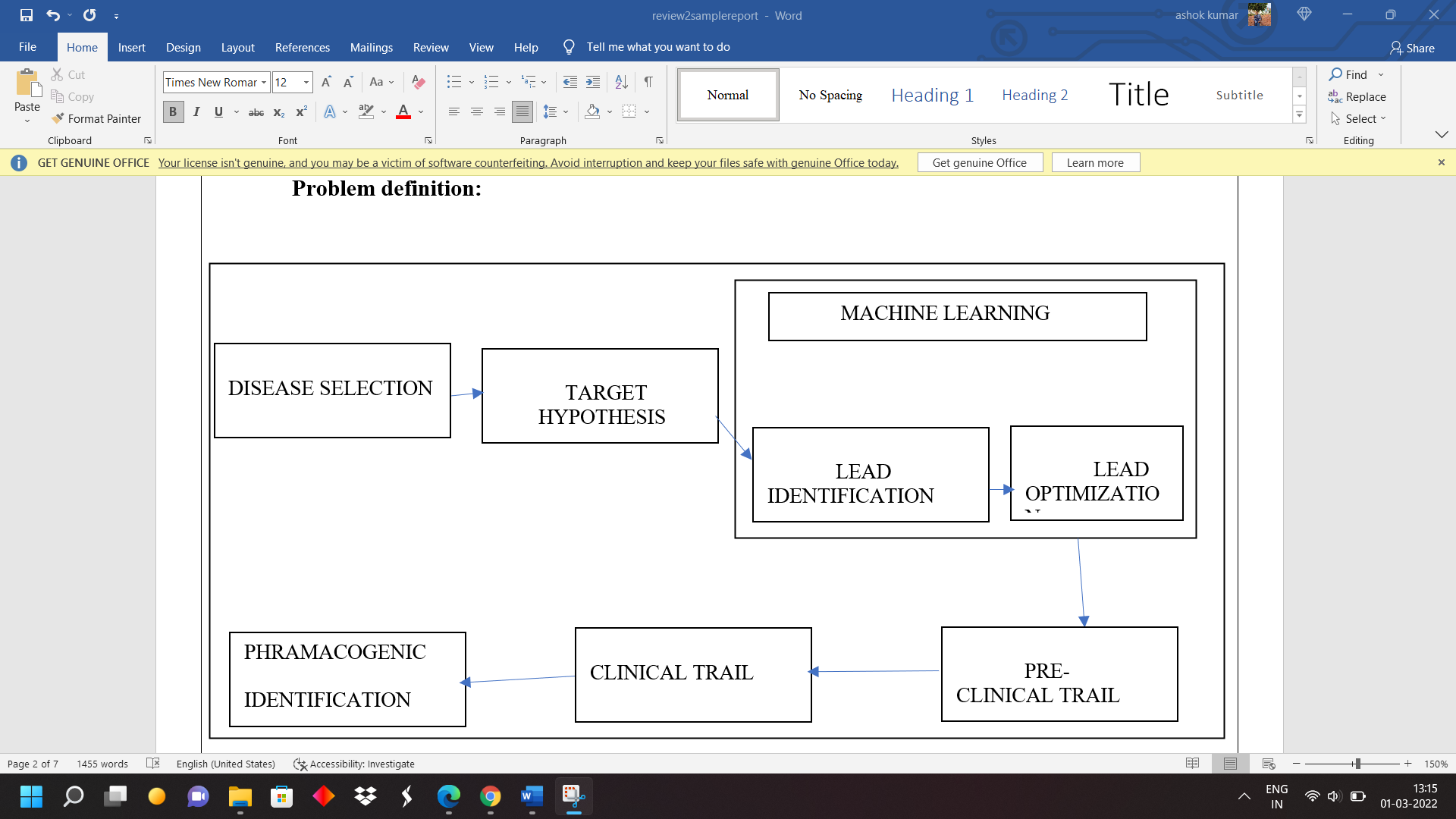
**Literature survey**

Hakime Ozturk *et al*[1], explained an approach on offensive content clustering in compound similarity in Drug Discovery using Machine learning. the Information about the design and discovery of drugs with the column called SMILES. SMILES(Simplified Molecular Input Entry Specification) consists of all the chemical formula of each and every compound. In this agglomative clustering, is used to combine the similar compounds. With this review, they have summarized the impact of NLP on bio/cheminformatics to encourage this already interdisciplinary field to take advantage of recent advances. Hakime Ozturk *et al*[2], Explained the study of SMILES-based Compund similarity functions for drug-target interaction prediction. Molecular structures can be represented as strings of special characters using SMILES. Since each molecule is represented as a string, the similarity between compounds can be computed using SMILES-based string similarityfunctions. Most previous studies on drug-target interaction prediction use 2D-based compound similarity kernels such as SIMCOMP. Martin Vogt *et al*[3] Described Python package for modelling Tanimoto similarity

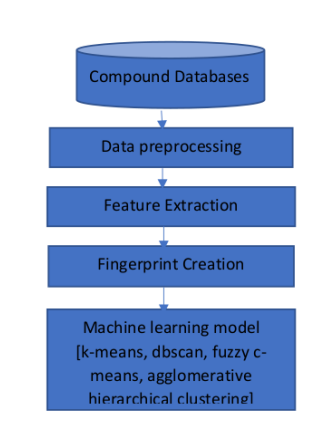
value distributions”. It described the information about how to use the ccbmlib Python package which is a collection of modules for modelling similarity value distributions. From this paper invoked that this is used to assess the statistical significance of tanimoto coefficients and evaluate how molecular similarity is reflected when different fingerprint representations are used and conditional significance score to estimate where a test compound would be ranked in a search and the resulting models have been evaluated for RDKit fingerprints, taking a collection of ChEMBL compounds as a reference data set. B. Zagidullin *et al*[4] describes the Application of machine and deep learning methods in drug discovery and cancer research has gained a considerable amount of attention in the past years. As the field grows, it becomes crucial to systematically evaluate the performance of novel computational solutions in relation to established techniques. To this end, we compare rule-based and data-driven molecular representations in prediction of drug combination sensitivity and drug synergy scores using standardized results of 14 high-throughput screening studies, comprising 64 200 unique combinations of 4153 molecules tested in 112 cancer cell lines. We evaluate the clustering performance of molecular representations and quantify their similarity by adapting the Centered Kernel Alignment metric. Our work demonstrates that to identify an optimal molecular representation type, it is necessary to supplement quantitative benchmark results with qualitative considerations, such as model interpretability and robustness, which may vary between and throughout preclinical drug development projects. B. Zagidullin *et al*[4] compare rule-based and data-driven molecular representations in prediction of drug combination sensitivity and finding drugsynergy scores using standardized results and later on to evaluate the clustering performance of molecular representations and quantify their similarity by adapting the Centred Kernel Alignment metric (identify an optimal molecular representation type (CKA)). Daniel Probst *et al*[5] explained that MAP4 is a new molecular fingerprint suitable for drugs and can be adopted as a universal fingerprint to describe and search chemical space so here it takes MinHasing. Ravi Manne *et al*[6] examined machine learning and deep learning techniques which helppharma industry in all stages of drug discovery which includes target validation, prognosticbookmarks, technical traits, in which these are totally seven phases in the process of drugdiscovery.Suresh Dara *et al*[7] Described the feasible literature on drug discovery through ML tools and techniques that are enforced in every phase of drug development to accelerate the research process. Target validation, prognostic biomarkers, digital pathology are considered as the problem statements where In clinical trials, absolute and methodological data must be generated to tackle many puzzles in validating ML techniques, improving decision-making, promoting awareness in ML approaches, and deducing risk failures in drug discovery. [8]“Anaconda and Jupyter Note Setup” to install anaconda for my execution where, the Anaconda distribution includes the Conda package manager in addition to the preconfigured Python packages and other tools. Anaconda Navigator is a GUI tool that is included in the Anaconda distribution and makes it easy to configure, install, and launch tools such as Jupyter Notebook.

**Problem definition:** Drug discovery is a very complicated process as it involves a huge investment of time and money. However, even after a huge struggle, the success rate is very low. Many long-term research projects may end up fruitless resulting in wastage of enormous efforts. The drug target designing combine machine learning that improve the quality of drugs discovered. Drug discovery involves seven step process that includes disease selection, target hypothesis, lead identification, lead optimization, pre- clinical trial, clinical trial, pharmacogenetic identification. Machine learning can be applied to identify the drug targets and in optimization of lead compound

**System Model:**

****

(i)System Model

****

1. Pre-Drug Discovery Process (Disease identification)

2. Modern Drug Discovery Process: The discovery process includes four important processes such as, target identification and validation, lead identification, lead optimization and pre-clinical trials.

a) Target Identification & Validation

used to identify target molecule which can be either gene or protein

b) Lead Identification

Lead identification also helps to see which molecules bind strongly to the target.

c) Lead Optimization

This phase results in finding the drug candidate from the lead identified compound. The goal is a process of refining the chemical structure of a confirmed. Hit to improve its drug characteristics.

d)Pre- Clinical Trial

an important phase to check whether the compound is working correctly or not

e) Clinical Trial

primary phase which will be fastest and safest way to find treatments.

Trials can be done in five ways such as, prevention trials, screening trials, diagnostic trials, treatment trial and quality of life trials.

**METHODOLOGY**

**k-means clustering** - A classification method that classifies data into k groups by

minimizing within-group distances to the centroid. *k*-means clustering is a method of vector

quantization, originally from signal processing, that aims to partition *n* observations into *k*

clusters in which each observation belongs to the cluster with the nearest mean, serving as a

prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. *k*-

means clustering minimizes within-cluster variances (squared Euclidean distances), but not

regular Euclidean distances, which would be the more difficult Weber problem: the mean

optimizes squared errors, whereas only the geometric median minimizes Euclidean distances.

For instance, better Euclidean solutions can be found using k-medians and k-medoids.

A) Procedure of K-mean Algorithm K-mean distributes all objects to K number of clusters at random;

1) Calculate the mean value of each cluster, and use this mean value to represent the cluster;

2) Re-distribute the objects to the closest cluster according to its distance to the cluster center;

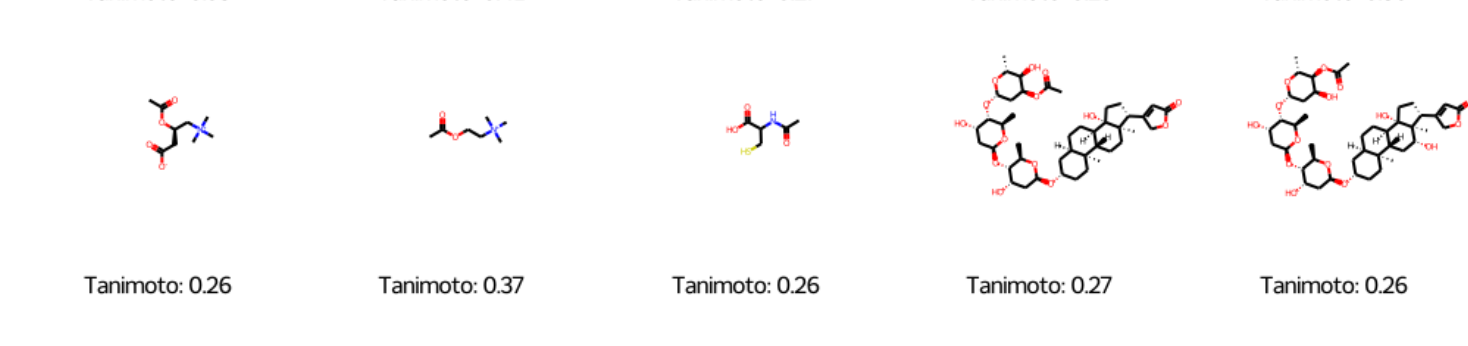
3) Update the mean value of the cluster, say, calculate the mean value of the objects in each cluster;

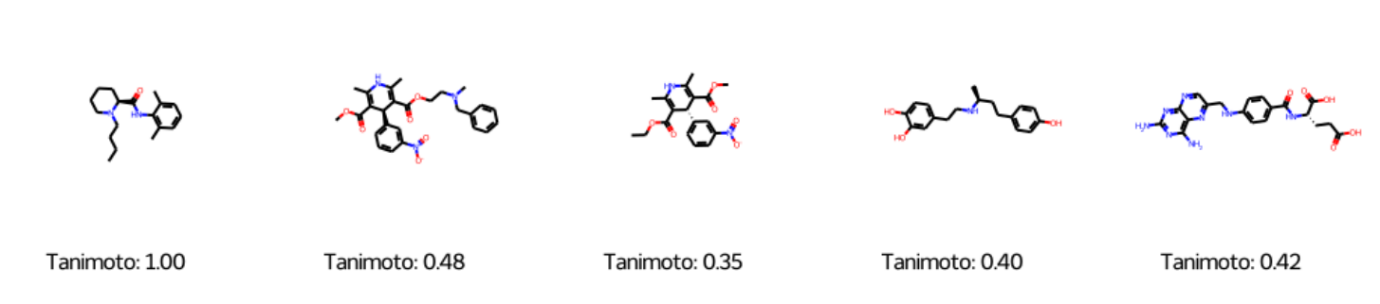
4) Calculate the criterion function E, until the criterion function converges.

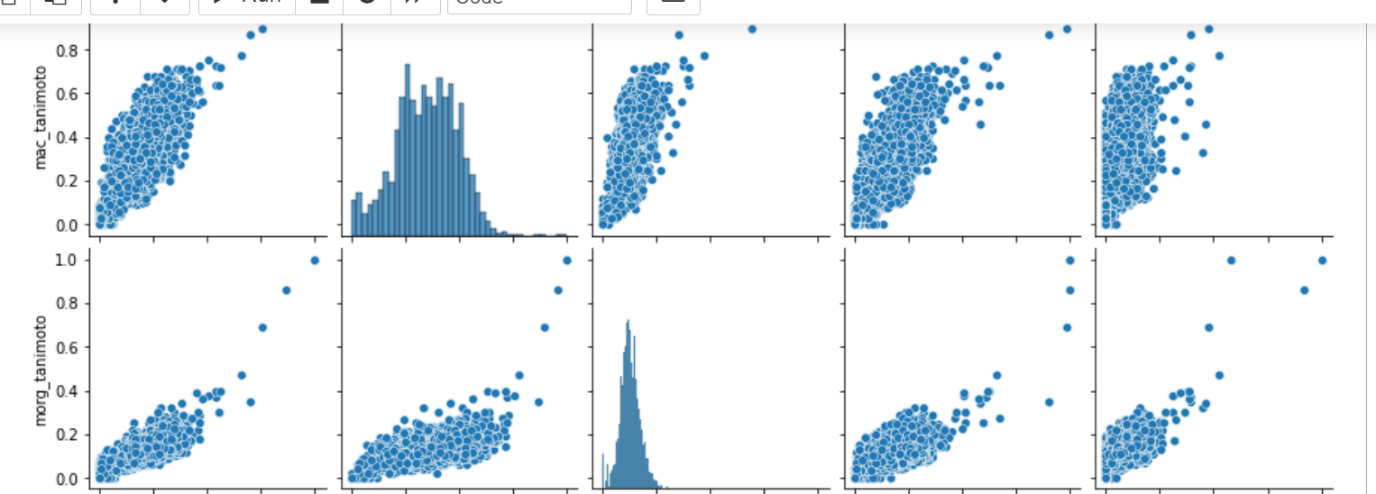
Usually, the K-mean algorithm criterion function adopts square error criterion, defined as:

E = ∑i=0k ∑ p ∊Ci | p - mi|2

In which, E is total square error of all the objects in the data cluster, p is given data object, mi is mean value of cluster Ci (p and m are both multi dimensional).







**Hierarchical clustering** - A classification method that builds a hierarchy of clusters

by agglomerative clustering e.g., merging smaller clusters or divisive clustering e.g.,

splitting a large cluster to smaller ones. *Hierarchical clustering,* also known as

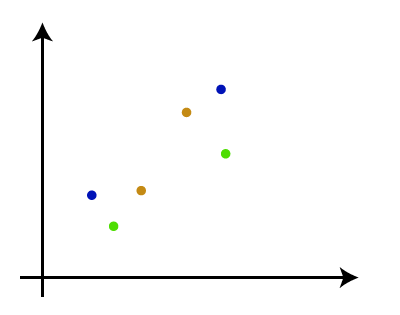
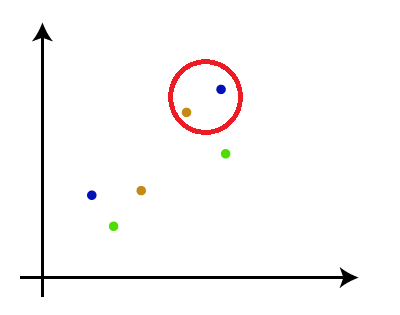
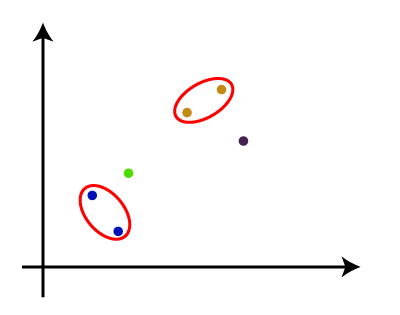
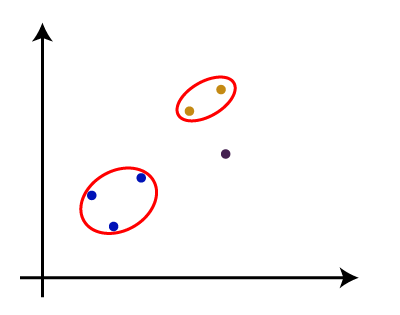
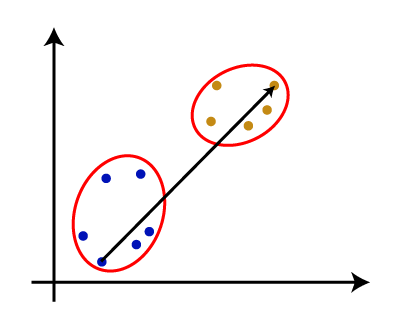
*hierarchical cluster analysis,* is an algorithm that groups similar objects into groups

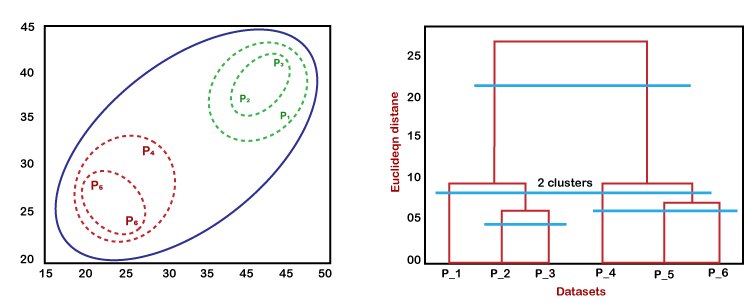
called *clusters*.

Algorithm:

1. Begin initialize c, c1 = n, Di = {xi}, i = 1,…,n ‘
2. Do c1 = c1 – 1
3. Find nearest clusters, say, Di and Dj
4. Merge Di and Dj
5. Until c = c1
6. Return c clusters
7. End

This algorithm begins with n clusters initially where each data point is a cluster. With each iteration, the number of clusters reduces by 1 as the 2 nearest clusters get merged. This process continues until the number of clusters reduces to the predefined value



**Principal component analysis** - A statistical method that uses orthogonal procedure

to transform a set of correlated features to new independent variables called principal

components. Principal Component Analysis (PCA) is an unsupervised machine learning

technique that attempts to derive a set of low-dimensional set of features from a much larger

set while still preserving as much variance as possible. Perhaps the two main applications of

PCA are. Variable selection. Visualizing High-Dimensional.

Independent component analysis -A statistical method that separates a

multivariable output into statistical independent additive components. Independent

Component Analysis (ICA) is­ a machine learning technique to separate independent sources

from a mixed signal. Unlike principal component analysis which focuses on maximizing the

variance of the data points, the independent component analysis focuses on independence,

i.e., independent components.

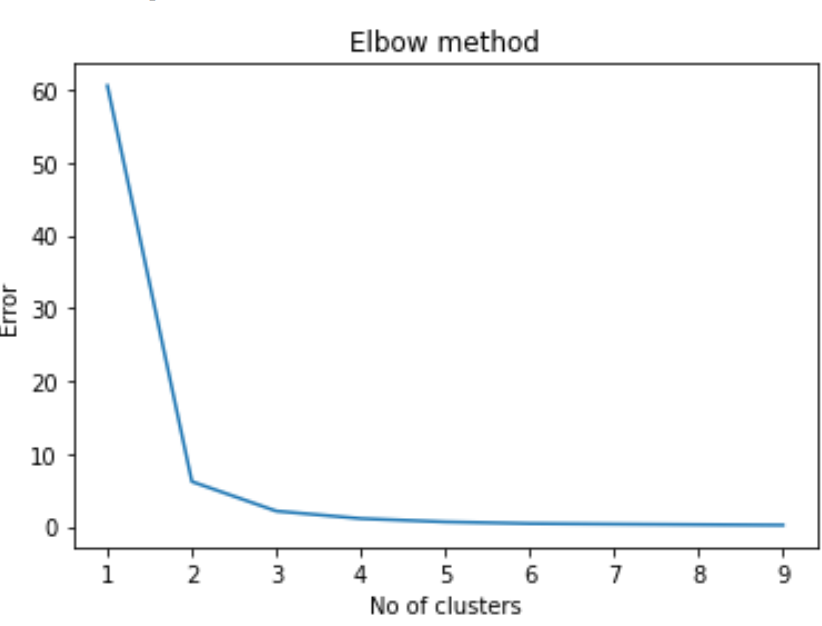
**Inputs:**

Dataset: MOL V3000 sdf file.

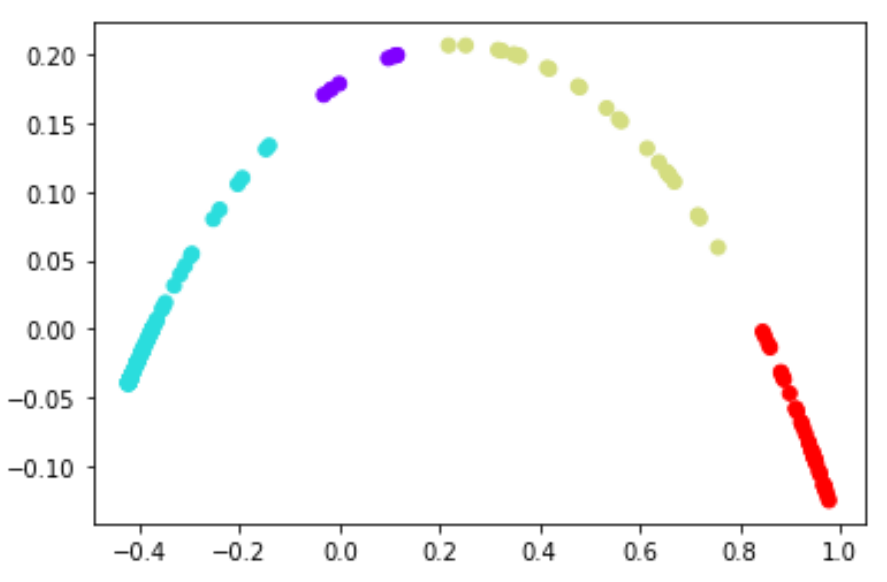
Source: <https://drugcentral.org/download>.

**\*Note: If you want to see the content of the file either translate into .xl format or with the help of marvin view.**

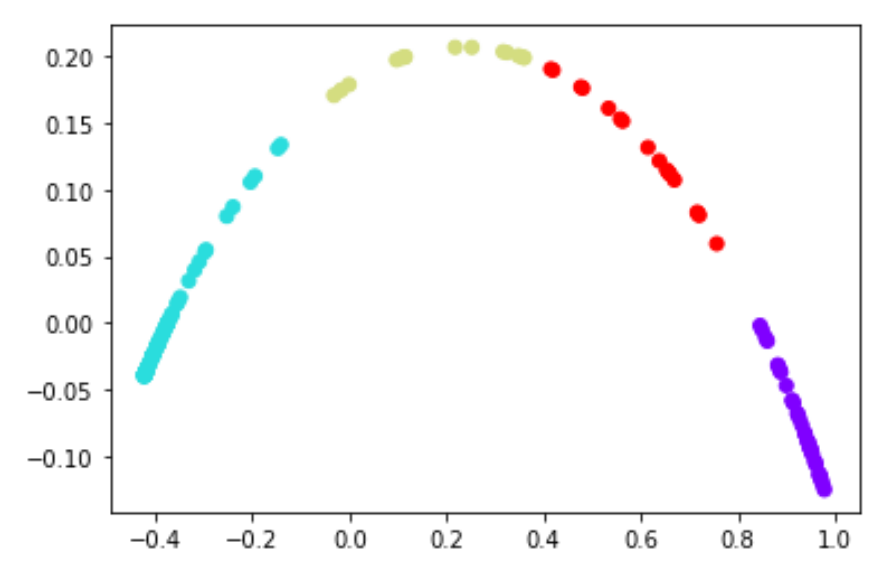
**Output:** We can analyse the similar compounds and can estimate the amount of dose required by the patient**.**

****

It is an elbow method representation to know the rate of error of total no. of clusters.



It shows the Density Based Clustering. It clustered into Four categories.

****

It Shows the K-Means Clustering where the percentage of compounds into group are differ from the Density Based Clustering.

**Hardware AND Software Requirements To Execute Project:**

**Hardware Requirements:**

Processor: Base Frequency- 1.1 GHZ

Ram: 4 GB

Hard Disk: 1 TB

**Software Requirements**:

Operating System: Windows.

Language: Python.

Tool: Anaconda Navigator, Jupyter Notebook, Gephi.

* In Jupyter Notebook we need to create rdkit environment.

Database: MY SQL.

Browser: Chrome, Mozilla, Opera Etc.

Technologies: Image Processing, Convolutional Neural Networks, Deep Learning, Machine Learning.

**Project Team**

|  |  |  |
| --- | --- | --- |
| Name | Roll No | Signature |
| M. Jyothsna | 18331A05A2 |  |
| K. Ashok Kumar | 19335A0509 |  |
| P. Sai Jahnavi | 18331A05B5 |  |
| S. Nikhila | 18331A05B2 |  |

**Supervisor:**

**Dr. P. SRINIVASA RAO**