DWDM ALGORITHM IMPLEMENTATION

TEAM-22(AIML-B)

Y.Suvarna(22b01a42c0) K.Trisha(22b01a42b0) P.Harini(22b01a4289)

# PROBLEM STATEMENT

Cluster students ba sed on acad em ic p erform ance using K-m ean s or

K-m ed oid s. Evaluate c lu ster qu ality with silh ouette score. Visualize c lusters using scatter p lots or d end rogram s. Aim to id entify hig h and low

perform ing stud ent group s for targ eted interventions. Enhan ce stud ent

ou tcom es throug h ta ilo red supp ort strateg ies.

# INTRODUCTION TO DATA SET



**The d ataset cap tures various attributes related to stud ents' academ ic**

**perform ance and d em og raphic inform ation. It includ es featu res such as gend er, nationality, p lace of b irth, stag e/g rade level, top ics studied,**

**sem ester, and several p erform an ce m etrics like raised h ands, v isited**

**resources, announcem ents view , and discussion p articipation. Add itionally, it includ es inform a tion on p arental involvem ent an d satisfaction, as w ell a s student ab sence d ays. This data set offers a com p rehensive v iew of**

**student en gag em ent and achievem ent in an ed ucational setting , provid in g**

**valuab le insig hts for analyzing a nd im p rovin g ed ucationa l outcom es.**

1. DESCRIBE THE DATA SET

The data set consists of inform ation related to stu dents' acad em ic

p erform a nce and d em og rap hic cha ra cteristics. It includes features such as:

* 1. **G ender: Indicates the g ender of the stud ent ( m ale or fem ale).**
  2. **N ationality: Rep resents the na tionality of the stu dent.**
  3. **Place of Birth: Specifies the p lace of b irth of th e stud ent.**
  4. **Stag e ID: In dicates the stag e or level of ed ucation (e.g ., low er level, m idd le school).**
  5. **G rad eID: Sp ecifies the grad e level of the stud ent ( e.g ., G -0 4, G -0 7 ).**
  6. **Section ID: Rep resents the section or c lass of th e stud ent.**
  7. **Top ic : Indicates the su bject or top ic stud ied b y the stud ent.**
  8. **Sem ester: S pecifies the sem ester d uring w hich the da ta w as recorded .**
  9. **Relation: In dica tes the rela tion of the respond ent ( parent) to the stud ent ( e.g ., father, m other).**

1 0. Raised Ha nds: Rep resents the num b er of tim es the student raised their hand in c lass.

1 1. Visited Resources: Sp ecifies the nu m ber of ed uca tional resources visited by the stu dent.

1 2. Announcem ents View: Rep resents the num b er of ann ouncem ents viewed b y the stud ent.

1 3. Discussion: Indicates the level of p articipa tion in c lass d iscussions.

1 4. ParentAnsw ering Survey: Ind icates w hether the p arent resp ond ed to the survey ( Y es/N o).

1 5. ParentschoolSatisfa ction : Sp ecifies the parent's satisfaction level w ith the school ( G ood /Bad ).

1 6. Stud entAbsence Days: Indicates the n um ber of student ab sence d ays ( U nd er-7 , Above-7 ).

1 7. Class: Represents the overall p erform ance class of the stud ent ( e.g.

, M for M ed iu m , L for Low , H for Hig h).

This da taset p rovid es a com prehen sive view of stud ents' acad em ic

engag em ent, p arental involvem en t, an d overall p erform an ce, enab ling analysis a nd insights into factors affecting stud ent su ccess in a n

education al environm ent.

# PREPROCESSING STEPS APPLIED WITH DETAILS

Prep rocessing step s for the given d ataset typ ically involve c lean ing ,

transform ing , and org anizing the data to prepa re it for analysis or m odeling . H ere's a general outline of prep rocessing step s you m ight take:

## Handling Missing Values:

Check for m issing values in each colum n. Decid e on a stra teg y to hand le m issing va lues, such as:

* + **Rem oving row s w ith m issing valu es.**
  + **Im p uting m issing values w ith m ean, m edian, or m od e.**

## Data Encoding:

* + **Convert categorical variab les in to num erical form at if necessa ry. T his can be d one using techniques like one-hot encod ing or la bel encod ing.**
  + **Decide wh ether encod ing is need ed for a ll categ orical variab les or only**

certain on es ba sed on their nature and the m ach ine learning a lg orithm to b e used.

## Feature Scaling:

* + **Check the scale of num erical features. If the features are on different scales, consider scaling them to a sim ilar rang e.**
  + **Com m on techn iques for featu re scaling include M in -M ax scaling or stand ard ization ( z-score norm alization).**

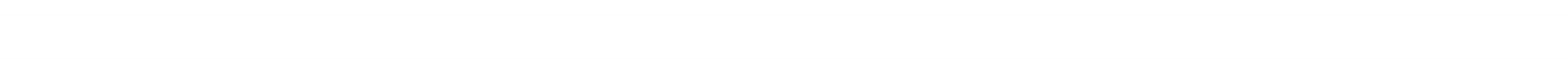
## Handling Categorical Variables:

* + **Analyze categorical variab les and decid e w hether they need to b e transform ed .**
  + **Ap ply one-hot encoding or label encod ing to convert categ orica l va riab les into a form at suitab le for m od eling .**

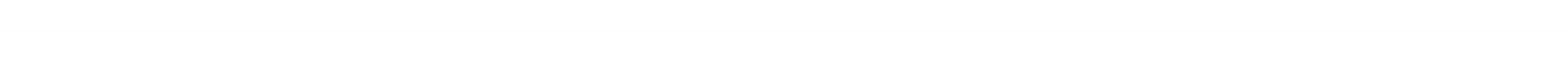
## Data Splitting:

Sp lit the d ataset into train ing and testing sets to evaluate m od el perform ance.

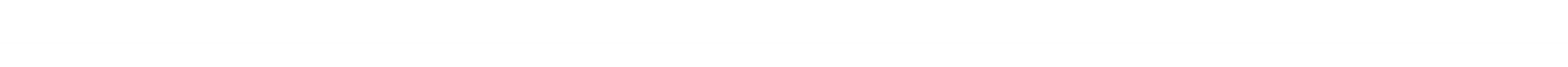
The typ ica l sp lit ratio is a round 70 -3 0 or 8 0-20 , dep end ing on the size of the d ataset.



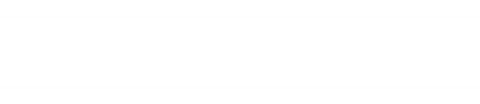
**These are som e com m on preprocessing step s, b ut the sp ecific steps you take**



**m ay vary d ep en ding on the na ture of your d ataset, the m ach ine learning**



**alg orithm you p lan to use, an d the sp ecific requirem ents of your analysis or**



**m odelin g task.**

# ALGORITHM IMPLEMENTATION

im port pa nda s as p d

from sklearn.cluster im p ort Ag glom era tiveClustering from sklearn.m etrics im p ort silhouette\_score

im port m atp lotlib .p yp lot as plt

from scip y.cluster.hierarchy im port dend rog ram , lin kage

# Load d ataset

d ata = p d .read\_csv('/content/km aenssam p le.csv') # Replace 'your\_d ataset.csv' w ith the actual filenam e or pa th

# Select relevant colum ns for clusterin g

colum ns\_for\_clustering = ['raised hand s', 'Vis ITedResou rces', 'An noun cem ents V iew ', 'D iscussion']

# Extract the d ata for clustering X = d ata[colum ns\_for\_clu stering ]

# Determ ine the optim al num b er of clusters u sing the silhou ette score silh ouette\_scores = []

for n\_clusters in ra nge(2 , 1 0):

clusterer = Agg lom erative Clustering (n\_clusters=n\_clusters, linkag e='w ard ') cluster\_lab els = c lusterer.fit\_pred ic t(X )

silhouette\_avg = silhouette\_score(X , cluster\_la bels) silhouette\_scores.app end( silh ouette\_avg )

# Plot silh ouette scores

p lt.plot(rang e(2 , 10 ), silhou ette\_scores, m arker='o') p lt.xlab el('Nu m ber of clusters')

p lt.ylab el('S ilhouette Score')

p lt.title ('Silhouette Score vs Nu m ber of Clusters') p lt.show ()

# Based on th e silhouette score, let's ch oose the num b er of clusters

optim al\_n\_clusters = silhou ette\_scores.index( m ax(silhouette\_scores)) + 2 #

Ad ding 2 to the ind ex to account for the starting cluster num b er

# Ap p ly Ag glom era tive clustering w ith op tim a l num b er of c lu sters

c lusterer = Ag glom erative Clustering (n\_clusters=op tim al\_n\_clusters, lin kage='ward ')

c luster\_lab els = clusterer.fit\_pred ic t(X )

# Visualize c lusters usin g scatter plot p lt.fig ure( fig size=( 1 0, 6 ))

for i in rang e( optim al\_n\_clusters):

p lt.scatter(X .iloc[cluster\_labels = = i, 0 ], X.iloc[cluster\_lab els == i, 1 ], lab el=f'Cluster {i+1 }')

p lt.xlab el('raised hand s')

p lt.ylab el('V is ITed Resources')

p lt.title ('Clustering of Stud ents b ased on Acad em ic Perform ance') p lt.leg end ()

p lt.show ()

# Visualize c lusters usin g d end rogram p lt.fig ure( fig size=( 1 0, 6 ))

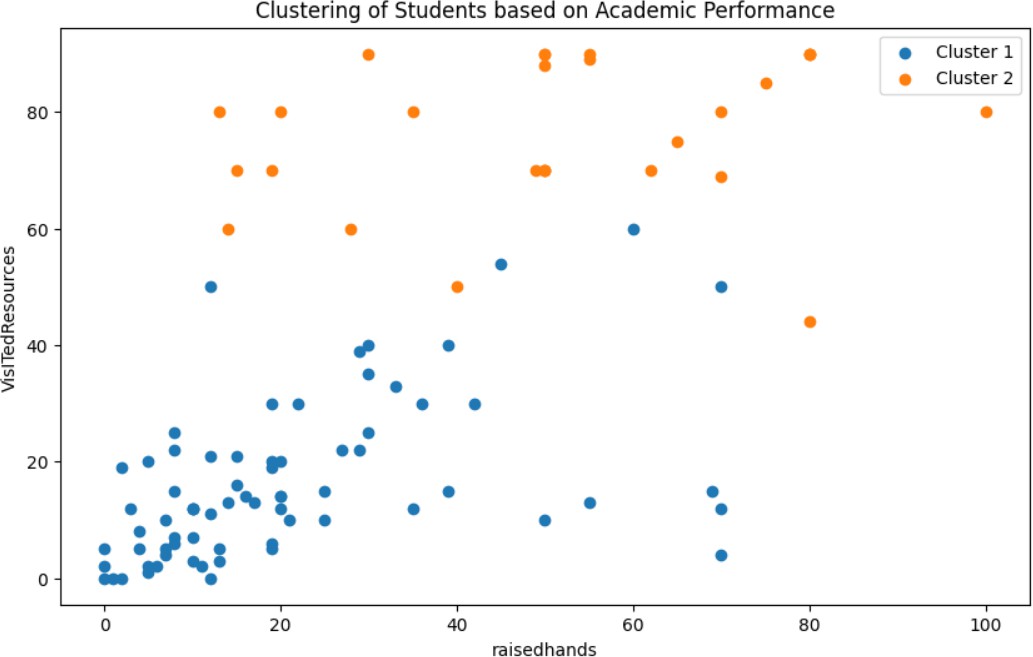
lin kage\_m atrix = linkage( X, m ethod ='w ard ') d end rogram (lin kage\_m atrix)

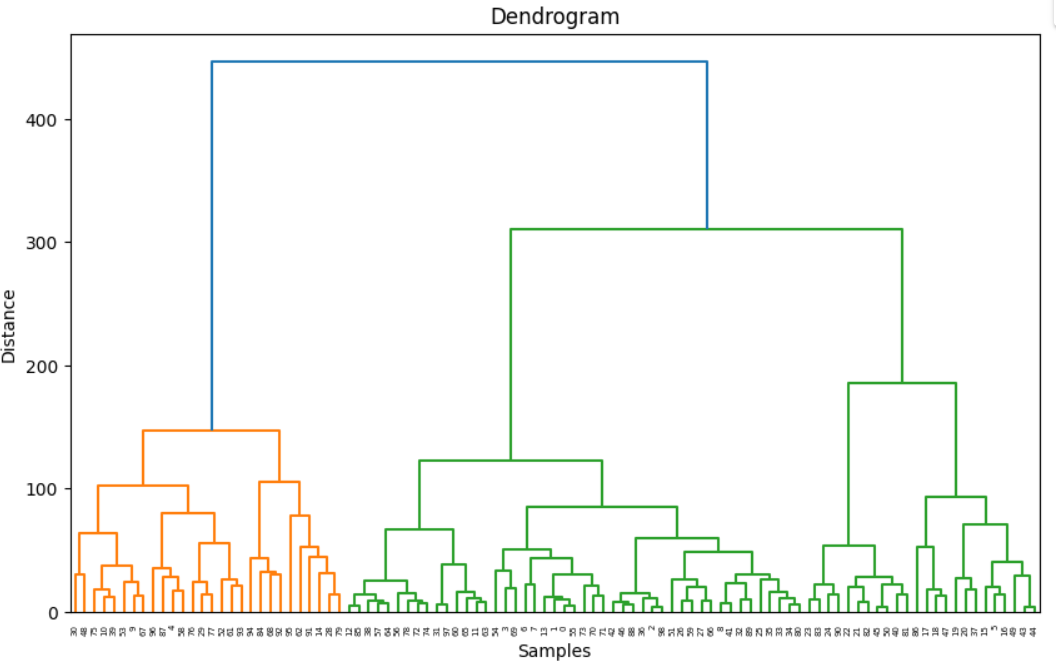
p lt.title ('Dend rog ram ') p lt.xlab el('Sam p les')

p lt.ylab el('Distance') p lt.show ()

# output:







Conclusion:

The analysis c lustered stud ents b ased on acad em ic eng ag em ent usin g features like raising ha nds, v isiting resources, an noun cem ents v iew , an d

d iscussion p articip ation. T he optim al num b er of clusters w as d eterm in ed

using silhouette scores. Th e visualization sh ow ed d istinct c lusters of stud ents w ith sim ila r eng agem ent p atterns. These c lusters can help ed ucators id entify g roup s need in g intervention and tailor sup p ort strategies. F urther analysis

could exp lore aca dem ic p erform an ce d ifferences am ong c lusters, aiding in targ eted edu cational interventions and resource alloca tion.