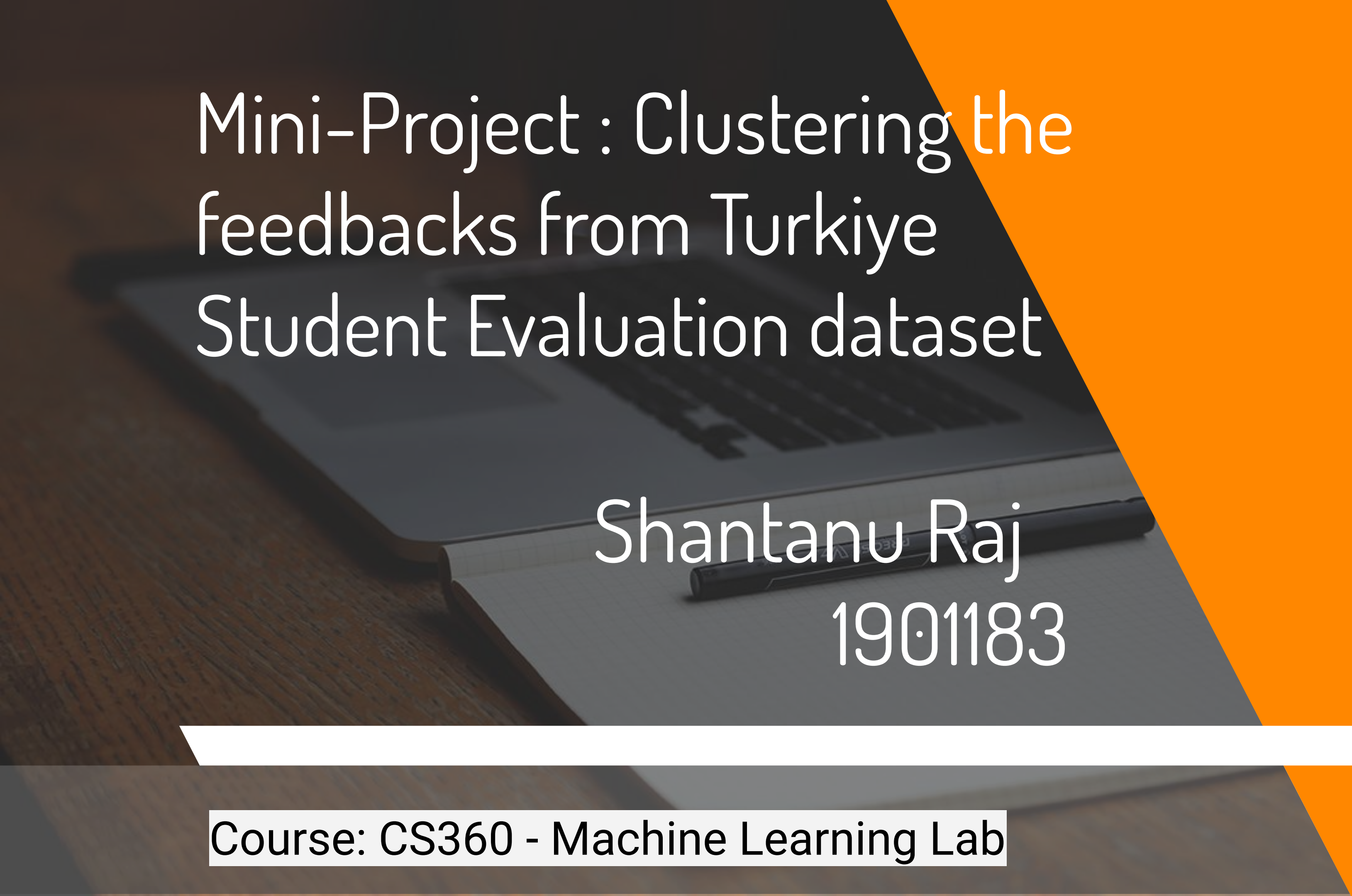


Mini-Project : Clustering the feedbacks from Turkiye Student Evaluation dataset

Shantanu Raj
1901183

Course: CS360 - Machine Learning Lab

Problem Definition

.An Unsupervised Learning Problem

.Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets.

.These algorithms discover hidden patterns or data groupings without the need for human intervention.

Problem- Use Different Clustering Technique to deal with given data

Different Clustering Technique Used in this project to cluster the data are :

1. K-Means
2. K-Medoid
3. Fuzzy C-Means (FCM)
4. Self Organizing Map (SOM)

About Dataset

- This data set contains a total 5820 evaluation scores provided by students from Gazi University in Ankara (Turkey).
- There is a total of 28 course specific questions and additional 5 attributes.
- Q1-Q28 are all Likert-type, meaning that the values are taken from {1,2,3,4,5}
- There are 3 instructors who are teaching 13 different courses
- 5820 patterns and 28 features are present in the dataset

- Additional 5 Attributes:
- instr: Instructor's identifier; values taken from {1,2,3}
- Class: Course code (descriptor); values taken from {1-13}
- Repeat: Number of times the student is taking this course; values taken from {0,1,2,3,...}
- attendance: Code of the level of attendance; values from {0, 1, 2, 3, 4}
- difficulty: Level of difficulty of the course as perceived by the student; values taken from {1,2,3,4,5}

Possible Reasons for dataset preparation:

.To predict course difficulty based on survey responses received by students from turkish university upon their completion of their respective courses.
.To improve instructor's performance based upon survey responses

Some important points regarding Dataset

.There are no Null values in the dataset
.Instructor 3 has taken more number of courses
.Course number is mostly liked and course no 12 is mostly disliked by students
.Course number 13 is mostly repeated by students
.Best Rating are given from class 2 and worst ratings are given from class 4 students
.According to the Student ratings we see that Instructor 1 and 2 are performing well and got similar ratings but Instructor 3 got less ratings
.we can recommend the instructor 3 for check on course 4 and 13.

Literature Survey 1 : Mr. D. Selvapandian , Mr. Thamba Meshach(December 25,2020)

An Efficient Sentiment Analysis on Feedback Assessment from Student to Provide Better Education

- Opinion mining concept is deployed to predict the trainer evaluation with student's feedback.
- To examine this feedback concept, where opinion examination helps to distinguish how students are communicated in writings and whether the articulations demonstrate positive (ideal) or negative (troublesome) and conclusions toward the subject.
- In this research work efficient fusion based neural network (EF-NN) classifier is introduced to predict the frequent context patterns used in the student feedback dataset.

- Student feedback data set is extracted based on attribute features like the interaction between the students, examination, and notes given, etc.
- Experimental results can be evaluated on weka toolbox based on this result negative and positive details are collected to improve the efficiency of teaching by faculty to provide the enhanced training.
- Finally, the result of the accuracy, recall, and precision is compared with the existing K-means clustering method.
- The performance of the proposed model is compared with the existing k means clustering technique it increases the 93% accuracy, 89% precision, and 91 % recall

Literature Survey 2 : V. Vijayalakshmi and A. Prakash(1 July 2020)

Developing an Optimized Feature Selection Process for Designing Efficient Content Management System using Educational Data

- In order to obtain better quality of e-learning, the pedagogical attributes have to be evaluated continuously for achieving collaborative environment.
- Due to the growing interest in e-learning technologies, the acquisition of relevant knowledge from educational information is a challenging task.
- Data mining is the field that seamlessly helps us to explore the knowledge from different evaluation aspects. Feature selection is an important step in the data mining process.
- This paper explores an innovative process to select the features using pedagogical data. The proposed framework composed of two phases, data pre-processing and feature selection.
- In pre-processing phase, single linked list is been employed to remove the duplicates values, so as to enhance the memory computational process.
- Canonical Correlation Analysis (CCorA) is used for deriving the relationship among the attributes and knowledge for the given education data.
- Experimental analysis is carried out in two datasets of UCI machine repository, namely, teaching assistant evaluation and Turkiye student evaluation dataset.
- Evidently, the results say that the proposed work is concentrated on selecting the features required for content management systems across two different datasets which paves the path for researchers in the field of e-learning.

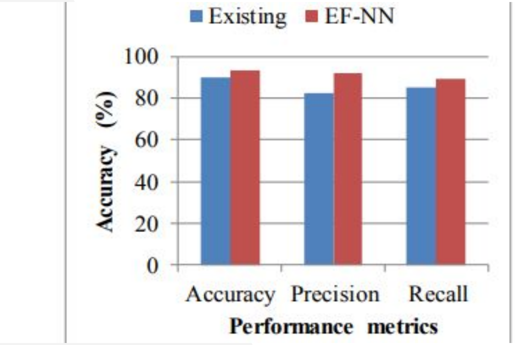
Literature Survey 3: Ahmed Mohamed Ahmeda, Ahmet Rizanerc (30 August 2019)

Using data mining to predict instructor performance

- This paper focuses on predicting the instructor performance and investigates the factors that affect students' achievements to improve the education system quality.
- Turkey Student Evaluation records dataset is considered and run on different data classifier such as J48 Decision Tree, Mlp, Naïve Bayes, and Sequential Minimal Optimization.
- Comparison of all the four classifiers is conducted to predict the accuracy and to find the best performing classification algorithm among all.
- The conclusions of this study are very promising and provide another point of view to evaluate student performance
- It also highlights the importance of employing data mining tools in the field of education.
- The results show that using the attribute evaluation method on the dataset increases the prediction performance accuracy.
- We concluded that using data of student evaluation for courses is useful to predict the factors that affect their achievement and also to predict instructors' performance.
- Moreover, it is another point of view to improve educational quality which is vital to attract students while most of the researchers used CGPA and internal assessment attributes to predict students' performance to enhance educational system

Some Graphs of Literature Survey

L1



L2

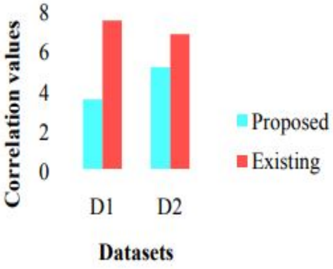


Figure 4: Information analysis between proposed and existing system

L3

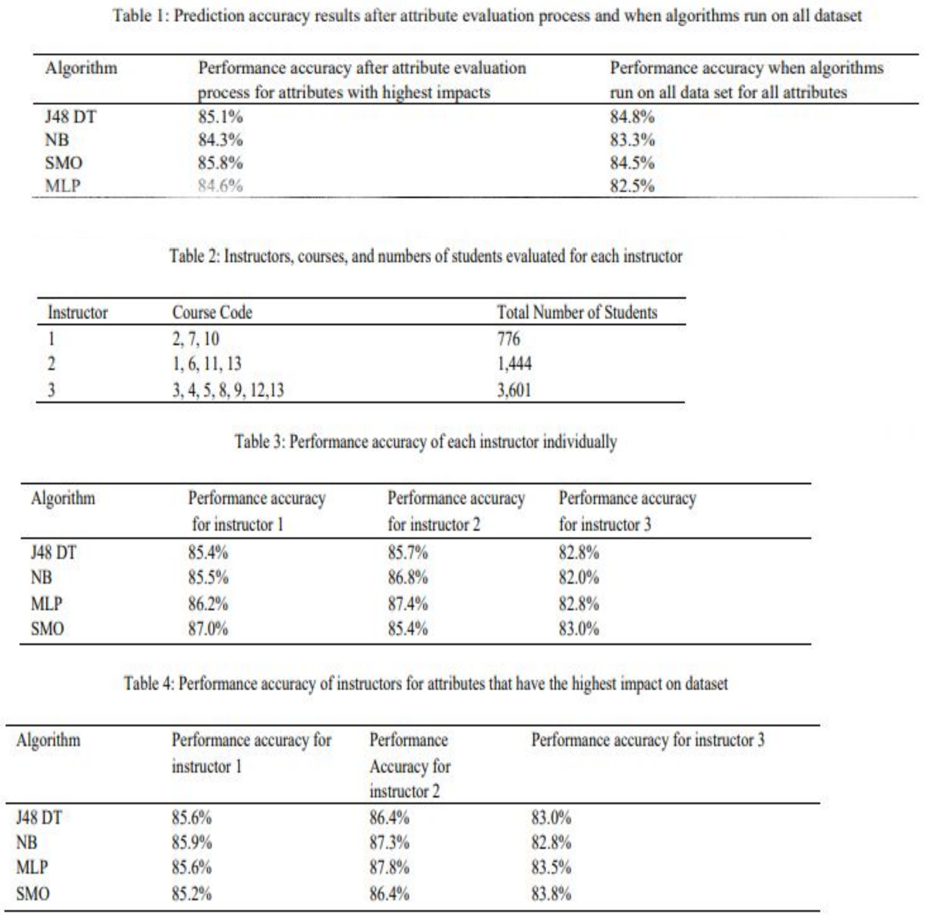


Table 1: Prediction accuracy results after attribute evaluation process and when algorithms run on all dataset

Algorithm	Performance accuracy after attribute evaluation process for attributes with highest impacts	Performance accuracy when algorithms run on all data set for all attributes
J48 DT	85.1%	84.8%
NB	84.3%	83.3%
SMO	85.8%	84.5%
MLP	84.6%	82.5%

Table 2: Instructors, courses, and numbers of students evaluated for each instructor

Instructor	Course Code	Total Number of Students
1	2, 7, 10	776
2	1, 6, 11, 13	1,444
3	3, 4, 5, 8, 9, 12,13	3,601

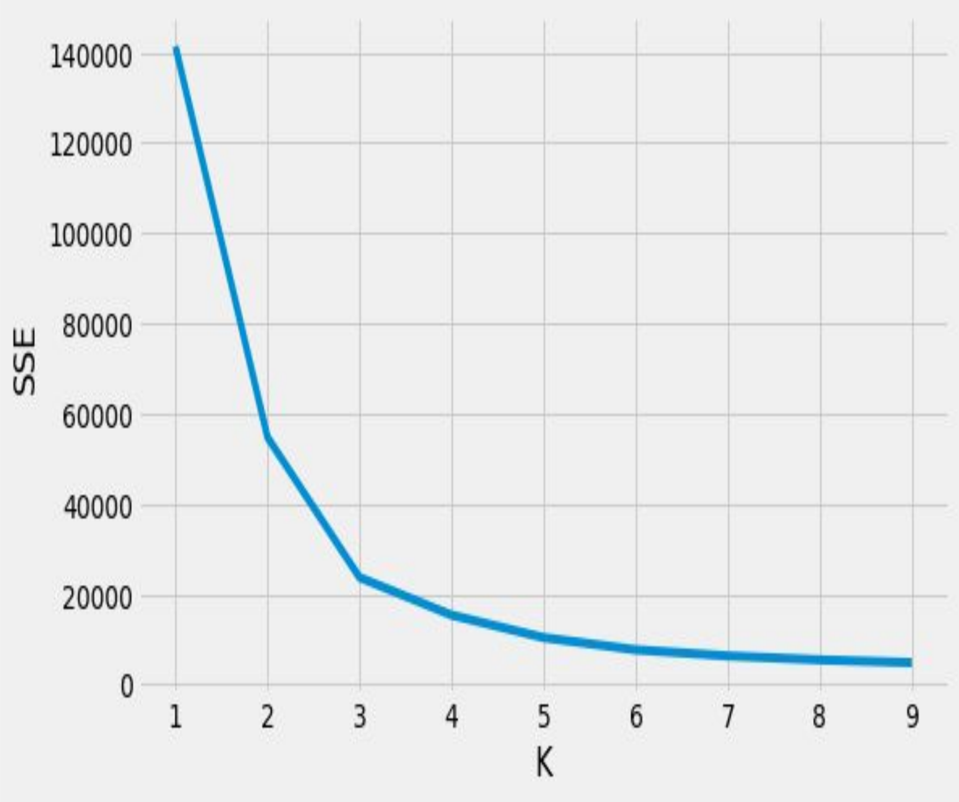
Table 3: Performance accuracy of each instructor individually

Algorithm	Performance accuracy for instructor 1	Performance accuracy for instructor 2	Performance accuracy for instructor 3
J48 DT	85.4%	85.7%	82.8%
NB	85.5%	86.8%	82.0%
MLP	86.2%	87.4%	82.8%
SMO	87.0%	85.4%	83.0%

Table 4: Performance accuracy of instructors for attributes that have the highest impact on dataset

Algorithm	Performance accuracy for instructor 1	Performance Accuracy for instructor 2	Performance accuracy for instructor 3
J48 DT	85.6%	86.4%	83.0%
NB	85.9%	87.3%	82.8%
MLP	85.6%	87.8%	83.5%
SMO	85.2%	86.4%	83.8%

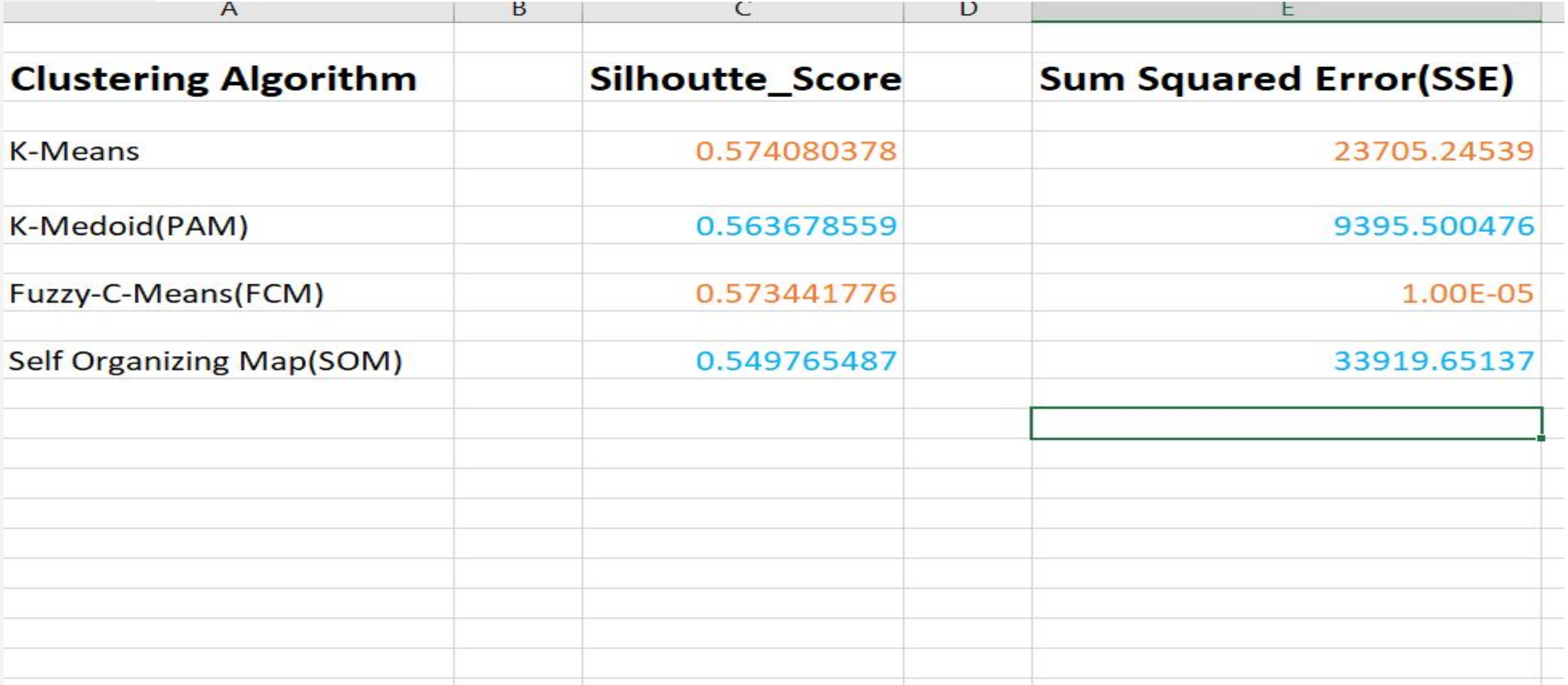
Elbow Plot



Elbow Plot:

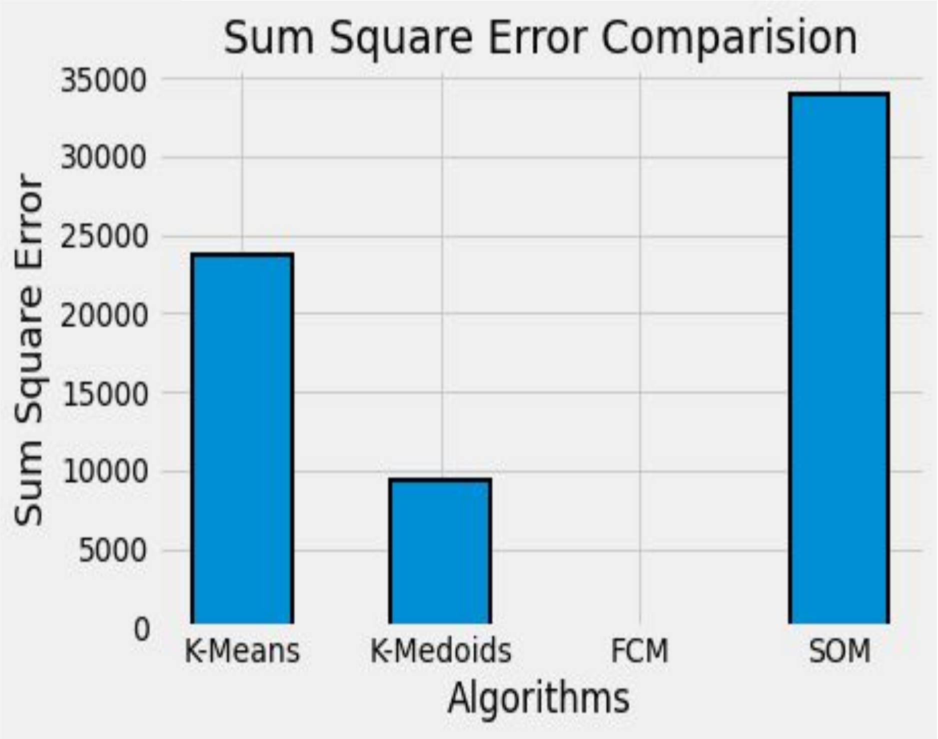
- The elbow method uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters
- We would choose a value of k where the SSE begins to flatten out and we see an inflection point. When visualized this graph would look somewhat like an elbow
- From this graph , we can decide number of clusters(k) = 3

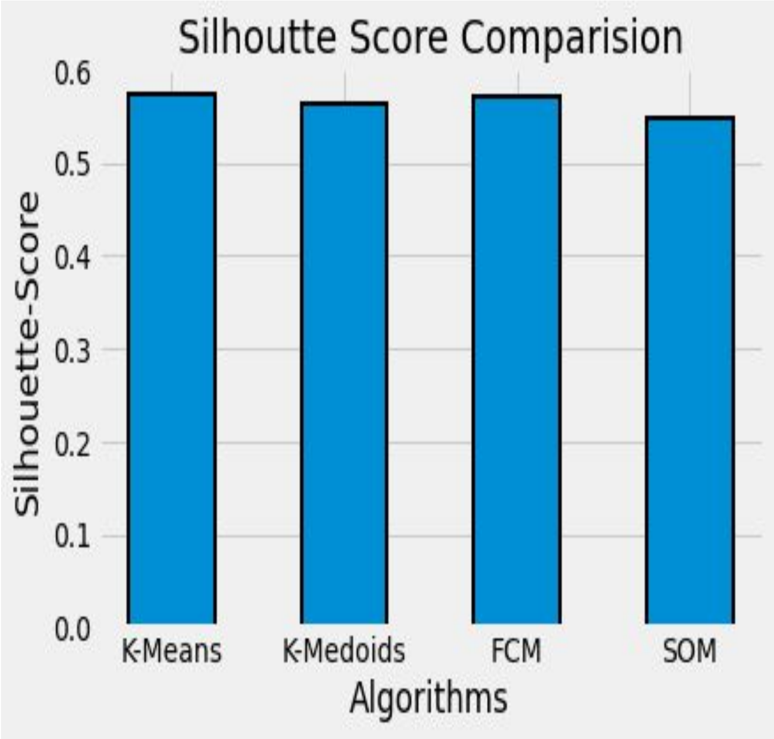
Results



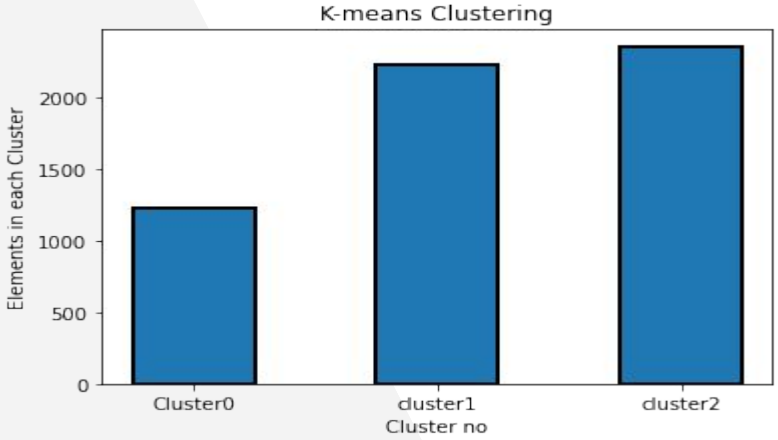
Clustering Algorithm	Silhoutte_Score	Sum Squared Error(SSE)
K-Means	0.574080378	23705.24539
K-Medoid(PAM)	0.563678559	9395.500476
Fuzzy-C-Means(FCM)	0.573441776	1.00E-05
Self Organizing Map(SOM)	0.549765487	33919.65137

Results

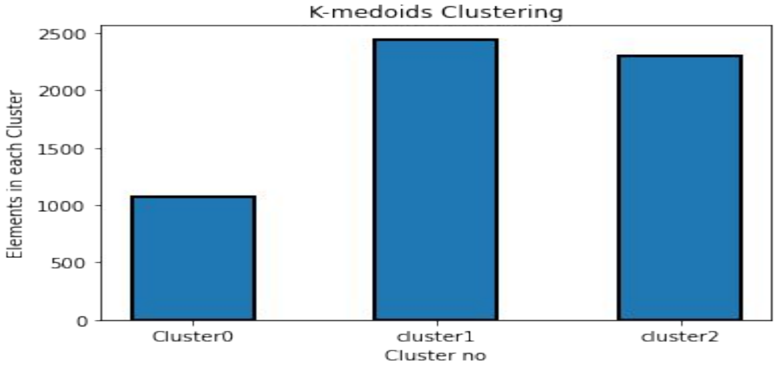

Sum Square Error Comparision
Sum Square Error
35000
30000
25000
20000
15000
10000
5000
0
K-Means
K-Medoids
FCM
SOM
Algorithms


Silhoutte Score Comparision
Silhouette-Score
0.6
0.5
0.4
0.3
0.2
0.1
0.0
K-Means
K-Medoids
FCM
SOM
Algorithms

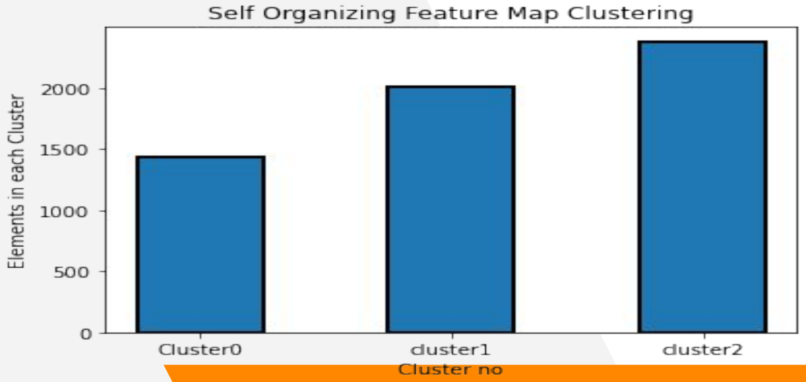
Results



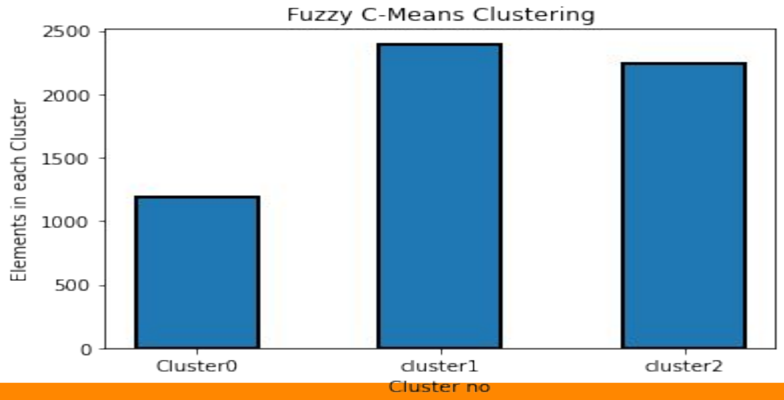
K-means Clustering



K-medoids Clustering



Self Organizing Feature Map Clustering



Fuzzy C-Means Clustering

Drawbacks faced..

- Difficult to predict K-Value in k-means algorithm
- Different initial partitions can result in different final clusters in K-means
- PAM has obtain different results for different runs on the same dataset because the first k medoids are chosen randomly.
- PAM has taken more time to fit the model in comparison with other algorithms.
- SOM and PAM has taken more time to Converge.
- During training the model with full dataset in PAM, Elbow plot remains constant.

Result Analysis

- Initially i have reduced the whole 28 dimension to 2 dimension,It co-relates the features that are similar and clubs them into 2d plot by maintaining the quality.

- In terms of SSE , Fuzzy C-Means performed exceptionally well.
- In terms of Silhouette Score , All four algorithms performed well but K-means performance is best.

- In terms of Convergence time , Som performance is best.

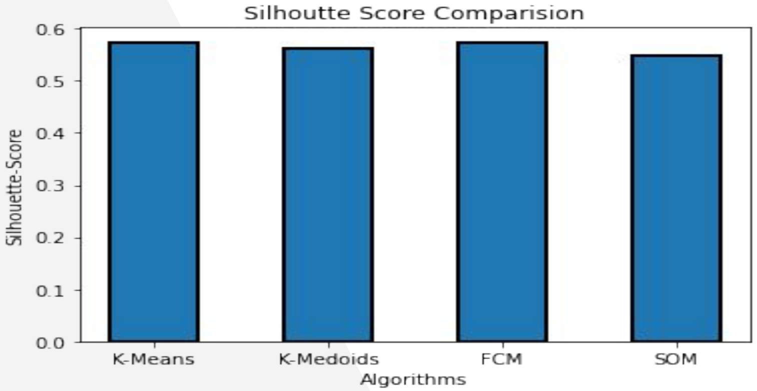
- From Overall Analysis , we can say that for this given dataset with reduced features k-means performance is best due to more score value.

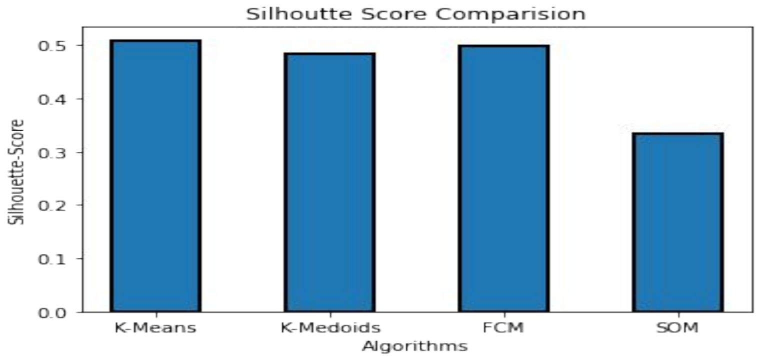
Novelty

Before applying any clustering algorithm to cluster the data proceed in this manner to get better performance:

- first normalize the given data using inbuilt functions
- then reduce features to 2 features
- then apply different clustering algorithm on the reduced dataset
- the performance of different clustering algorithm will be better on the dataset prepared in this manner than any other dataset.

Novelty





This represents the score of different algorithm on the dataset prepared in this manner:
.first normalize the given data
.then reduce its feature to 2 feature from normalized data
.then apply different clustering algorithms on reduced data

This represents the score of different algorithm on the dataset prepared in this manner:
.first reduce its feature to 2 feature from given dataset
.then apply normalization on reduced data
.then apply different clustering algorithms on normalized data