Machine Learning Mini-project

Clustering the feedbacks from Turkiye Student Evaluation dataset

Course: CS360 - Machine Learning Lab

Dept. of Computer Science & Engineering
Indian Institute of Information Technology, Guwahati.

NAME : CHIKKE SRUJAN
ROLL N0 : 1901055

OUTLINE

- → Abstraction
- → Problem statement
- → Literature survey
- → Result & Result Analysis

Abstraction :

- 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.
- It has direct 28 questions on the course & instructors and Students have rated it on **Likert Scale(** meaning that the values are taken from {1,2,3,4,5}).
- While course difficulty level, type of course, number of times the students repeat the course, etc. has also been taken into consideration.
- The questionnaire includes items to evaluate the course and the instructor. Our investigation on the outcomes of this questionnaire has two distinct issues.
- The first is finding out what is important for Turkish students while judging a course and its instructor as mentioned.
- Along with this consequence, we also try to use clustering – a data mining technique – on the dataset to try evaluating a questionnaire in some other way than using the traditional statistical techniques.

- ➢ Correspondingly, it is seen that the attribute attendance that is correlated with both the other attributes – difficulty and retaking the course – and most of the questionnaire items has significance also in clustering.
- ➢ Therefore, the data is clustered into two and three groups according to the questionnaire items and attendance using k-means clustering.
- ➢ The correlation results show that the students emphasize the importance of attendance; the instructor’s being prepared for course, being open and respectful to the students.
- ➢ On the other hand, clustering can only be done by taking attendance into consideration. Since the difficulty and being a repeat student does not have significant difference, the cluster developed using these are not well defined.

Problem Statement :

❑ Clustering the data from Turkey student evaluation data set(Unsupervised Learning Problem).

❑ Grouping the whole data into sufficient number of clusters.

❑ Clustering Algorithms:K-Means,K_Medoid,Fuzzy & SOM clustering algorithms.

Literature Survey :

- ➢ **An Efficient Sentiment Analysis on Feedback Assessment from Student to Provide Better Education**
- ➢ 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.
- ➢ Opinion mining concept is deployed to predict the trainer evaluation with student's feedback.
- ➢ The proposed Efficient Fusion Based Neural Network (EF-NN) classifier initially, where the dataset is processed and transferred to attributes selection, then transform the data for processing the signals further into classification. The hybrid model produces the concept of sentimental data classification with high accuracy

- The exactness is determined for all motion picture audits accessible in the corpus to approve the proposed model archive 93% accuracy based on the number of feedbacks correctly predicted. The number of feedbacks achieved from the different reviewers on precision is 92% and recall is calculated on the true positive rate it has been achieved 89%.
- Reference : https://sci-hub.ee/10.1109/i-smac49090.2020.9243594
- Mr. D. Selvapandian ,
- Mr. Thamba Meshach W , Mr.K.S.Suresh ,
- Dr.R.Dhanapal,Dr.Jebakumar Immanuel.D.

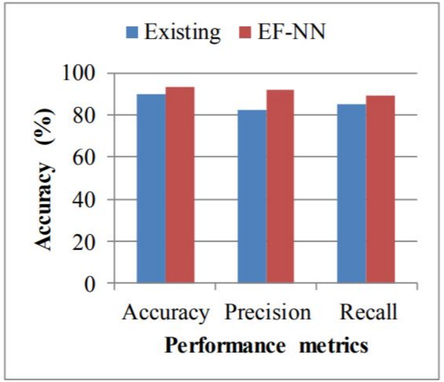


Fig.1. Performance metrics of accuracy, precision, and recall

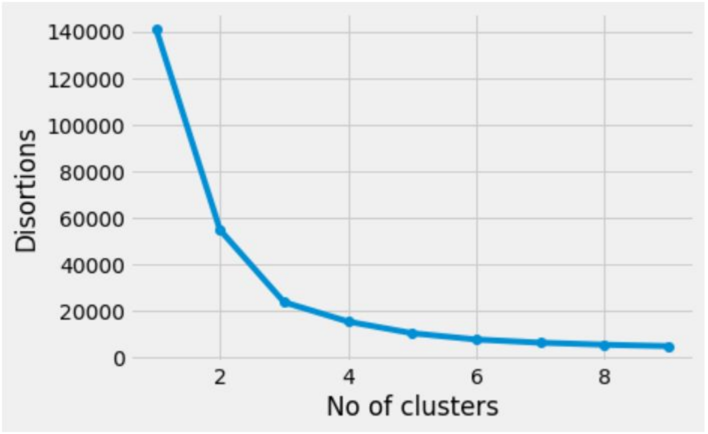
- **Clustering-Based EMT Model for Predicting Student Performance**
- In this paper they discussed about Educational data mining (EDM) .It is one of the research areas that uses data mining techniques on the educational dataset to extract valuable knowledge making it interpretable and easy to be understandable for decision making.
- Unsupervised techniques find the hidden knowledge from un labeled datasets by the inherited associations between records where the results represented as a label for future records. The most common unsupervised method is the clustering technique that splits data into groups/clusters based on their related properties
- Reference : (Link)

➢ 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.
- ➢ The conclusions of this study are very promising and provide another point of view to evaluate student performance
- ➢ 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.
- ➢ Furthermore removing the worst ranked attributes that have a lower impact on dataset increased the algorithms performance accuracies.
- ➢ Reference : **Ahmed Mohamed Ahmed,Ahmet Rizaner,Ali Hakan Ulusoyc**

Result & Result Analysis:

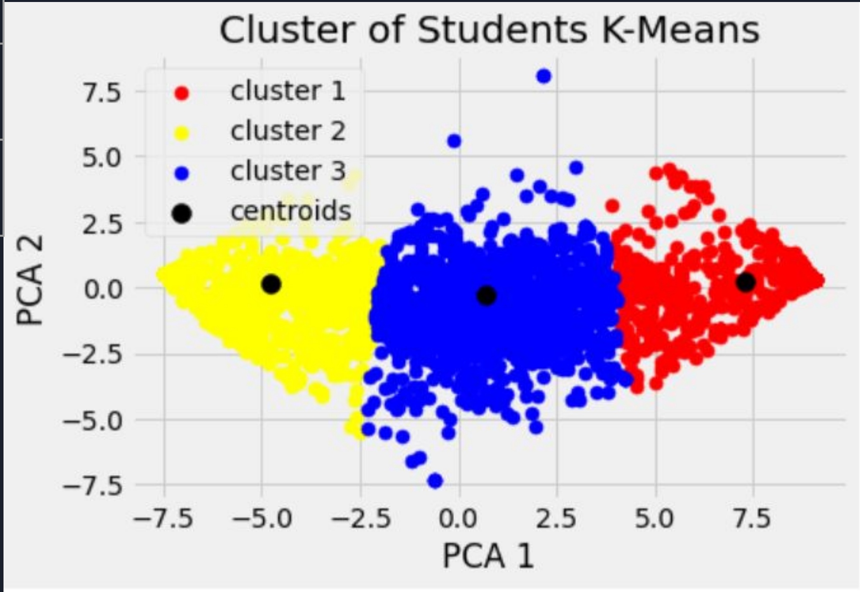
- **Elbow Method** :In cluster analysis, the **elbow method** is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.
- Based on elbow graph ,we can go for 3 clusters.



K-Means Clustering :

- SSE = 23705.143795
- SILHOUETTE SCORE = 0.57399421246334
-

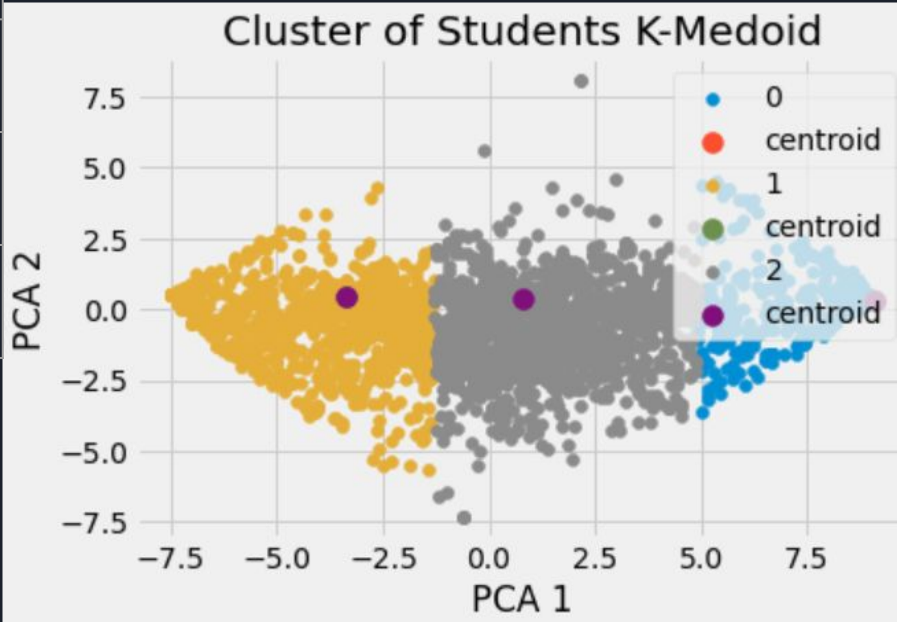
cluster	count
0	1230
1	2227
2	2363



K-Medoid clustering :

- SSE = 23705.143795
- SILHOUETTE SCORE = 0.563678530737
-

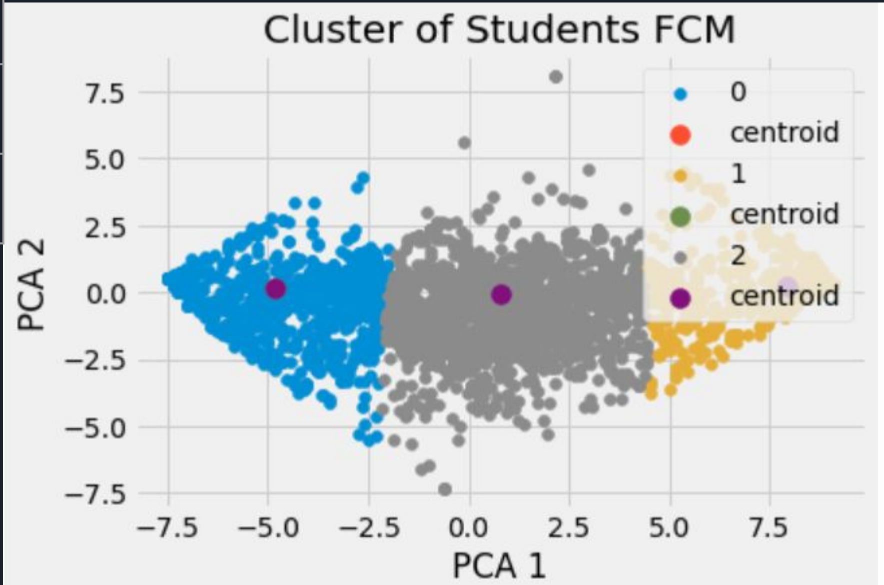
cluster	count
0	1077
1	2446
2	2297



Fuzzy C-Means (FCM) :

- SSE : 24177.871276025
- SILHOUETTE SCORE : 0.573441776316902
-

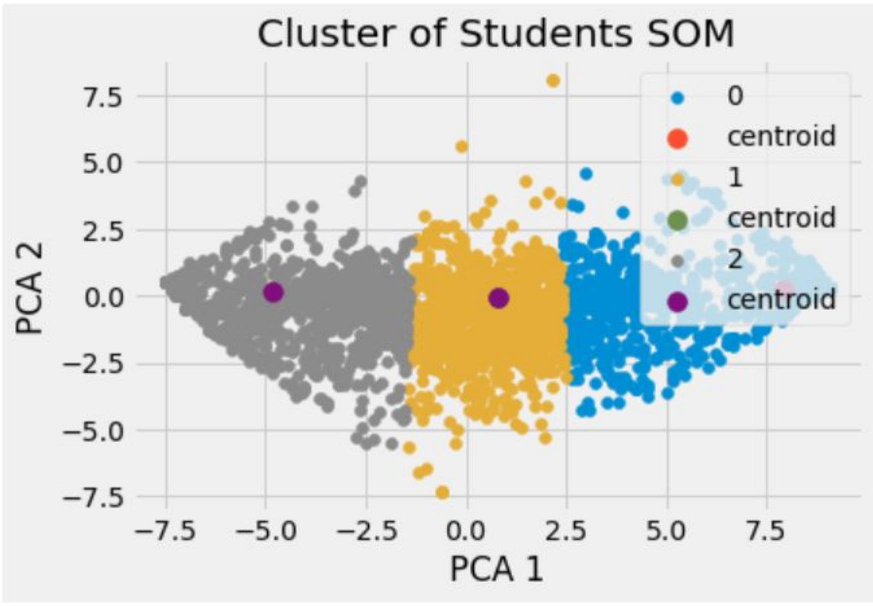
cluster	count
0	2394
1	1185
2	2241



Self Organising map(SOM) clustering :

- SSE = 36861.15281
- SILHOUETTE SCORE = 0.53913297918043
-

cluster	count
0	1636
1	1963
2	2221



Result Analysis :

- ➢ As our data has 28 dimensions , we need to reduce the number of dimensions to 2 by using Principle component Analysis . This will reduce the computational time and it also retains the information of the data set.
- ➢ We have calculated sum square error and silhouette score of 4 clustering algorithms to calculate performance of the models .
- ➢ According to Sum square error K-Medoid(9395.50) gives the best performance.
- ➢ According to Silhouette square K-Means(0.57399421246334) gives the best Performance.
- ➢ When it comes to convergence based on time, SOM(Self Organising Maps) gave the Best performance.
- ➢ From the Overall analysis we can tell that K-Means gave the best Performance for the given DataSet due to more score value.