Software Requirements Specification

For

Student Segmentation using Clustering Algorithm

15-11-2021

Under

Dr.Sujoy Chatterjee

Prepared by

Specialization	SAP ID	Name
AI & Ml	500075940	Rohan Nyati
AI & Ml	500076347	Rajneesh
AI & Ml	500075224	Shantanu Jaswal

UPES

UNIVERSITY WITH A PURPOSE

Department of Informatics
School Of Computer Science

UNIVERSITY OF PETROLEUM & ENERGY STUDIES,
DEHRADUN- 248007. Uttarakhand

Table of Contents

Revision History

Date	Change	Reason for Changes	Mentor Signature
-	-	-	-

General Instructions:

1. Font should be Time new Roman 12
2. Main heading should be All Capital with Times New Roman 14
3. Sub-Heading should be Times new roman 12 , Underline
4. Line gap should be 1.15
5. Justified alignment should be used for all text
6. Content inside a table should be Times New Roman 10
7. Caption for both Table and Figure should be Times New Roman 11
8. Add Source for all Images used.

1	INTRODUCTION	
	1.1 Purpose of the Project	Describe the scope of this project by stating and justifying the problem statement of the project. Present will clear motivation to execute the project.
	1.2 Target Beneficiary	Identify the prime beneficiaries of the project.
	1.3 Project Scope	Provide a short description of the area of application of the software, include relevant benefits, objectives, and goals. State clearly the requirement and deliverables of the project.
	1.4 References	List all documents or Web addresses to which this SRS refers.
2	PROJECT DESCRIPTION	
	2.1 Reference Algorithm	State the reference algorithm for the project and identify the required data structure (**Mandatory for Minor1**) Or/Add design algorithm justifying the methodology of the project
	2.2 Characteristic of Data	Present with the characteristic of the dataset used for the project. Provide the primary and secondary source of the data, along with sampling techniques. Explain the statistical method used for data processing (**if any**).
	2.3 SWOT Analysis	Present with a justification to support your project.
	2.4 Project Features	Summarize the major features the product contains or the significant functions that it performs or lets the user perform. (Level 2 USE Case diagram)
	2.5 User Classes and Characteristics	Identify the various user classes that you anticipate will use this product.

	2.6 Design and Implementation Constraints	Present hardware boundary conditions (timing requirements, memory requirements); interfaces to other applications; specific technologies, and tools to be used; parallel operations; language requirements; communications protocols; security considerations; design conventions or programming standards.
	2.7 Design diagrams	Present all the required Diagram (USE –Case, Class Diagram, Activity, Sequence, Data Flow diagram and State Diagram. (Major project should include Collaboration and Deployment Diagram too)
	2.8 Assumption and Dependencies	List any assumed factors (as opposed to known facts) that could affect the requirements stated in the SRS. Also identify any dependencies the project has on external factors.
3	SYSTEM REQUIREMENTS	
	3.1 User Interface	Define the software components for which a user interface is needed.
	3.2 Software Interface	Describe the connections between modules. Describe the services needed and the nature of communications. Describe detailed application programming interface protocols.
	3.3 Database Interface	Explain the Database management system used
	3.4 Protocols	Describe the requirements associated with any protocol deployed in the project. Specify any communication security or encryption issues, data transfer rates, and synchronization mechanisms
4	NON-FUNCTIONAL REQUIREMENTS	
	4.1 Performance requirements	If there are performance requirements for the product under various circumstances, state them. Specify the timing relationships for real time systems. State performance requirements for individual functional requirements or features
	4.2 Security requirements	Specify any requirements regarding security or privacy issues surrounding use of the product or protection of the data used or created by the product. Define authentication, verification and validation of the system. Refer to any external policies or regulations containing security issues that affect the product.
	4.3 Software Quality Attributes	Explain: adaptability, availability, correctness, flexibility, interoperability, maintainability, portability, reliability, reusability, robustness, testability, and usability.
5	Other Requirements	Define any other requirements not covered elsewhere in the SRS.
Appendix A: Glossary		Define all the terms necessary to properly interpret the SRS, including acronyms and abbreviations.
Appendix B: Analysis Model		Pertinent analysis models used for this project

Appendix C: Issues List	This is a dynamic list of the open requirements issues.

INTRODUCTION

There has been a lot of deal about the coronavirus that started in 2019 and due to that the students have not been able to join the college for studies and colleges have been running in online mode for more than a year .

As the colleges have been planning to open again after one harsh year we are planning to create a project that helps colleges to get an idea about how many students are fully vaccinated and will be joining the college after this pandemic. We will be creating some random data on which we will be implementing clustering algorithm that will help us get predictions about the current mental state of the students in terms of academic knowledge and how much they were able to learn and understand during this pandemic on the basis of which colleges will be able to predict which student needs how much help and how many extra classes will be required for that student and will be able to group them on the basis of that , thus help them get back on track .

Purpose of the Project :

If there are 1,2,3,.......,n number of students :

1. How to segment students on the basis of fully vaccinated and not vaccinated .
2. How to segment those students into different clusters (C_1,C_2,C_3…)
3. How to map new students into particular clusters (P_1,P_2,...)

Target Beneficiary :

The target beneficiaries of this project are students , who are able to state their basic understandings of topics and current academic knowledge .
College departments are able to identify students' state of mind and be able to predict how much extra effort the college needs to provide to students .

Project Scope :

The Motivation behind opting for this is that for the past one year, students have been studying in online mode and because of which they are facing difficulties in some areas. So our motivation here is to access the college in managing such students and assisting them to get back on track.

Basic Objective of the Project is too :

1. Clear understanding of Clustering algorithms
2. To create a way for colleges to obtain a better idea about how much help a student requires after returning to college to improve their academic status .
3. To create clusters of students on the basis of their current academic knowledge .

References :

1]A Clustering Method Based on K-Means Algorithm Youguo Li, Haiyan Wu Department of Computer Science Xinyang Agriculture College Xinyang, Henan 464000,China December 2012
https://www.researchgate.net/publication/271616608_A_Clustering_Method_Based_on_K-Means_Algorithm/link/57da70fc08aeea1959316130/downloa

[2]Dynamic Incremental K-means Clustering Bryant Aaron, Dan E. Tamir Department of Computer Science, Texas State University, San Marcos, Texas, USA, Naphtali D. Rishe, and Abraham Kandel School of Computing and Information Sciences Florida International University Miami,Florida,USA 2014 International Conference
http://cake.fiu.edu/Publications/Aaron+al-14-DK.Dynamic_Incremental_K-means_Clustering_IEEE-downloaded.pdf

[3] A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu Institute for Computer Science, University of Munich Oettingenstr. 67, D-80538 Miinchen, German© 1996
https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

[4] International Journal of Enterprise Computing and Business Systems Analysis and Study of Incremental DBSCAN Clustering Algorithm SANJAY CHAKRABORTY Prof. N.K.NAGWANI National Institute of Technology National Institute of Technology (NIT) Raipur, CG, India.Vol. 1 Issue 2 July 2011
https://arxiv.org/ftp/arxiv/papers/1406/1406.4754.pdf

PROJECT DESCRIPTION

Reference Algorithm :

Algorithm: DBSCAN: a density-based clustering algorithm

DBSCAN is a density based algorithm used for clustering that requires only one input parameter and supports the user in determining an appropriate value for it. It discovers clusters of arbitrary shape. Finally, DBSCAN is efficient even for large spatial databases.

Input :

- D: a data set containing n objects,
- r: the radius parameter, and
- MinPts: the ϵ-neighborhood density threshold.

Output : A set of density-based clusters.

Method:

```
mark all objects as unvisited;
do
    randomly select an unvisited object p;
    mark p as visited;
    if the $\epsilon$-neighborhood of p has at least MinPts objects
        create a new cluster C, and add p to C;
        let N be the set of objects in the $\epsilon$-neighborhood of p;
        for each point p’ in N
            if p’ is unvisited
                mark p’ as visited;
                if the $\epsilon$-neighborhood of p’ has at least MinPts points, add those points to N;
            if p’ is not yet a member of any cluster, add p’ to C;
        end for
        output C;
    else mark p as noise;
until no object is unvisited
```

Algorithm: k-means. The k-means algorithm for partitioning, where each cluster’s center is represented by the mean value of the objects in the cluster.

K-Means algorithm is a cluster algorithm, and is proposed by J.B.MacQueen. This algorithm which is unsupervised is usually used in data mining and pattern recognition. Aiming at minimizing cluster performance index, square-error and error criterion are foundations of this algorithm. The K-Means algorithm based on dividing has advantages of briefness, efficiency and certainty.

Input: k: the number of clusters, D: a data set containing n objects.
Output: A set of k clusters.
Method:
(1) arbitrarily choose k objects from D as the initial cluster centers;
(2) repeat
(3) (re)assign each object to the cluster to which the object is the most similar, based on mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5)until no change

Characteristic of Data :

```
RangeIndex: 1499 entries, 0 to 1498
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   sno             1499 non-null   int64  
 1   Name            1499 non-null   object 
 2   vaccinated      1499 non-null   int64  
 3   joining campus  1499 non-null   int64  
 4   SEM3            1499 non-null   float64
 5   SEM4            1499 non-null   float64
dtypes: float64(3), int64(3), object(1)
memory usage: 82.1+ KB
```

Total 6 columns including 3 integer entries , 2 float entry and 1 object entry.

The dataset contains 1499 values , including sno, name , vaccination information, joining campus preference,semester (3,4) sgpa of students.

The dataset is in CSV(comma separated values) format and Text format .

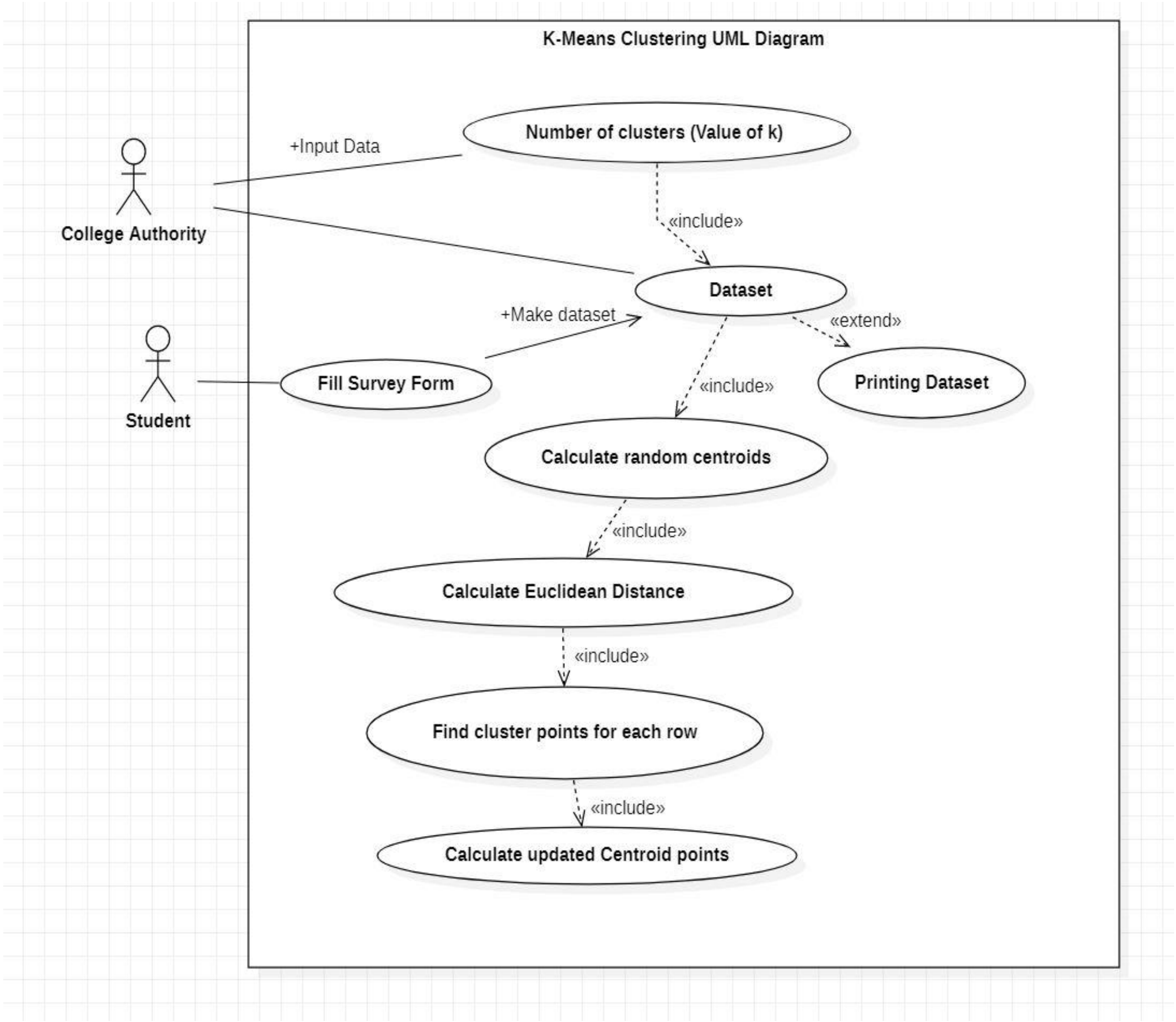
SWOT Analysis :

Strengths	Weaknesses	Opportunities	Threats
Clustering simplifies the management of large or rapidly growing systems.	The clustering result sensitive to the type of kernel and its parameters	Algorithms can be used in a single file .	Very large dataset
Machines provide greater processing power	Time complexity being high	Code will work smoother on python.	Accuracy will suffer if the values are very close in the dataset .
As your user base grows and report complexity increases, your resources can grow.	Not suitable for very large-scale data.	Data Visualization can be done using python and machine learning .	May not work on low end systems.

Project Features :

This project enables the user to segment the student data into clusters using a machine learning algorithm , in order to aid decision makers and authorities to make efficient decisions. This project is developed using c++ programming language.

Use Case diagram :


K-Means Clustering UML Diagram
College Authority
Student
+Input Data
Number of clusters (Value of k)
«include»
Dataset
+Make dataset
«extend»
Fill Survey Form
«include»
Printing Dataset
Calculate random centroids
«include»
Calculate Euclidean Distance
«include»
Find cluster points for each row
«include»
Calculate updated Centroid points

User Classes and Characteristics :

User Classes:

Colleges : The College authorities and management can make use of the project.

Students : For educational purposes and to find where they stand among the other peers.

Characteristics:

#include <iostream>- iostream is the header file which contains all the functions of the program like cout, cin etc.

#include <fstream>- This data type represents the file stream generally, and has the capabilities of both ofstream and ifstream which means it can create files, write information to files, and read information from files.

#include <vector>- By writing #include <vector> , you are telling the compiler to not only use your own code, but to also compile a file called vector .

#include <math.h> - The math.h header defines various mathematical functions.

#include <chrono>-This Chrono library is used for date and time.

#include <algorithm>-The header <algorithm> defines a collection of functions especially designed to be used on ranges of elements.

Design and Implementation Constraints :

Systems capable of executing c and cpp languages and installed with c and cpp libraries.

Systems installed with compiler MSV C++ 11.0 and higher.

Systems installed with basic cpp development IDE(windows and mac) or terminal(ubuntu and linux) to execute the program.

SOFTWARE REQUIREMENTS

Name of Component		Specification
Operating System		Windows 10, Macintosh
Front end		C , C++ Programming Language
IDE Required		Visual Studio Code/XCode

HARDWARE REQUIREMENTS

Name of Component		Specification
Processor		Intel(R) Core(TM)i5-3210M CPY @ 2.50GHz 2.50
RAM		4GB
Hard Disk		500GB HDD or 250GB SSD
Mouse		2 or 3 Button mouse
Keyboard		101 Key Keyboard