

Department of Informatics

School Of Computer Science

UNIVERSITY OF PETROLEUM & ENERGY STUDIES,

DEHRADUN- 248007, Uttarakhand

SOFTWARE REQUIREMENTS SPECIFICATION

For

Geyser Eruption Analysis by the use K-mean clustering algorithm

11th November 2022

Under the guidance of

**Mr. Bhagwant Singh**

Prepared By

|  |  |  |
| --- | --- | --- |
| **Specialization** | **SAP ID** | **Name** |
| B. Tech CSE(Big Data) | 500085576 | Anushka Sharma |
| B. Tech CSE(Big Data) | 500087652 | Garvita Adhikari |
| B. Tech CSE(Big Data) | 500082524 | Krish Aggarawal |
| B. Tech CSE(Big Data) | 500082352 | Udbhav Singh Sengar |

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Topic** | | **Page No** |
| 1 | Introduction |  |
|  | 1.1 Purpose of the Project | 3 |
|  | 1.2 Target Beneficiary | 3 |
|  | 1.3 Project Scope | 3 |
|  | 1.4 References | 3 |
| 2 | Project Description |  |
|  | 2.1 Reference Algorithm | 3 |
|  | 2.2 Data/ Data structure | 3 |
|  | 2.3 SWOT Analysis | 4 |
|  | 2.4 Project Features | 4 |
|  | 2.5 User Classes and Characteristics | 4 |
|  | 2.6 Design and Implementation Constraints | 4 |
|  | 2.7 Design diagrams | 5 |
|  | 2.8 Assumption and Dependencies | 6 |
| 3 | System Requirements |  |
|  | 3.1 User Interface | 6 |
| 4 | Non-functional Requirements |  |
|  | 4.1 Performance requirements | 6 |
|  | 4.2 Software Quality Attributes | 6 |
| Appendix A: Glossary | | 6 |
| Appendix C: Issues List | | 6 |

# INTRODUCTION

## Purpose of the project:

To examine data of various time stretches at which geyser eruptions occur in order to compute and compare the time interval of eruption. Geysers provide a natural laboratory to study multiphase eruptive processes.

## Target beneficiary:

To deliver services to the Meteorological Department.

## Project Scope:

When the geyser erupt , thermal energy convert into kinetic energy. So the kinetic energy can be used for industrial purpose.

## Reference:

<https://education.nationalgeographic.org/resource/geyser>

<https://kasumisanchika.medium.com/k-means-clustering-for-old-faithful-geyser-eruptions-analysis-74f3df8979b1#:~:text=Here%20comes%20K%2Dmeans%20Clustering&text=The%20clustering%20mechanism%20is%20done,to%20find%20subsets%20of%20features.&text=K%2DMeans%20is%20a%20clustering,number%20of%20non%2Doverlapping%20subgroups>.

# PROJECT DESCRIPTION

## Reference algorithm:

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means [clustering](https://www.javatpoint.com/clustering-in-machine-learning) algorithm mainly performs two tasks:

* Determines the best value for K center points or centroids by an iterative process.
* Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

## Characteristics of data:

The dataset is in such a way that it contains:

* 271 data points.
* 2 Attributes containing waiting time between two consecutive geyser eruptions (in minutes) and the duration of the eruption (in minutes) for a given geyser.

This data is organized in numeric data format (.txt file). Each file stores a struct containing the following fields for a given geyser:

* Waiting time between two consecutive geyser eruptions: (in minutes)
* Duration of the eruption :(in minutes)

## SWOT Analysis:

**Strength:**

1. Many real-world opportunities are associated with it.

2. Highly scalable.

**Weakness:**

1. Scaling with a number of dimensions.

2. Clustering data of varying sizes and density

**Opportunities:**

Time is important for everyone, so it gives an edge over other technologies for saving your time.

**Threat:**

Currently access of data is not supervised with any authentication thus this poses a threat to privacy of the user.

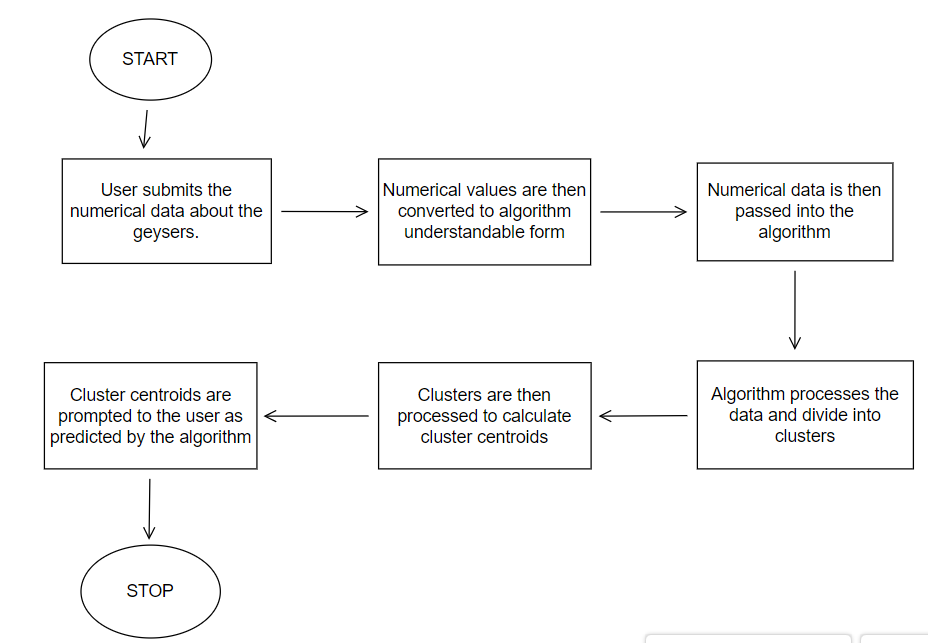
## Project Features

## A user can give the data to be analyzed through a text file as input which must contain waiting time between two consecutive geyser eruptions (in minutes) and the duration of the eruption (in minutes) for a given geyser. After that our model will explore the spans at which explicit geyser erupts and submits numerical scatter plots for the time stretches by the utilization of K mean clustering calculation .

## Design and implementation constraints:

1. Precise numerical data.

## Design diagrams

****

## Assumptions and Dependencies:

User provides the dataset according to the data format which is accepted by the system.

Data provided is taken in minutes.

# SYSTEM REQUIREMENTS

## User interface:

## Any IDE suitable for C++ can also be used as an interface.

# NON-FUNCTIONAL REQUIREMENTS

## Performance requirements:

1. For K means clustering, an IDE with a powerful compiler required.
2. Operating System- we have used windows operating system, though any operating system capable of running C++ code can be used.
3. Highly accurate numerical floating point data.

## Software quality attributes:

1. Availability: The system will consistently be accessible for access at 24 hours, 7 days per week.
2. Additionally, in the event of any significant framework failing, the whole program can be executed on other systems with slightly changed results which don’t seriously influence the whole project.
3. Maintainability: The developer should maintain correct relations in the algorithms.
4. Flexibility: A straightforward however quality code will be created to make it straightforward and can accept all kinds of numeric data as input and also can run on different IDEs.
5. Usability: The data should be of the format as required by the system.

# APPENDIX A: GLOSSARY

IDE: Integrated development environment

**END**