Geyser Eruption Analysis by the use K-mean clustering algorithm

A Project Submitted
In Partial Fulfillment of the Requirements
For the Course of
Minor Project - I
In
Third year – Fifth Semester of
Bachelor of Technology
Specialization
In
Big Data

Under

Mr.Bhagwant Singh

By

SAPID	Roll No	Name
500085576	R2142201903	Anushka Sharma
500087652	R2142201887	GarvitaAdhikari
500082524	R214220624	Krish Aggarawal
500082352	R2142201245	Udbhav Singh Sengar

DEPARTMENT OF INFORMATICS
SCHOOL OF COMPUTER SCIENCE
UNIVERSITY OF PETROLEUM AND ENERGY STUDIES, BIDHOLI,
DEHRADUN, UTTRAKHAND, INDIA

December, 2022

TABLE OF CONTENT

S. No.	Content	Page
1	Introduction	3
1.1	ProblemStatement	4
1.2	Motivation	4
1.3	Objectives	4
2	LR	5
3	Methodology	6-7
4	Hardware Requirements	7
5	Software and Libraries requirements	7
6	Design Diagrams	8
7	Result and Output	9
8	References	10
9	Appendix	11

1. Introduction

Machine Learning is the subfield of computer science, according to Arthur Samuel. In 1959, he gave the statement, "computers are having the ability to learn without being explicitly programmed". Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions based on data. Machine Learning algorithms are the programs that can learn the hidden patterns from the data, predict the output, and improve the performance from experiences on their own. Different algorithms can be used in machine learning for different tasks, such as K-Means Clustering algorithm that can be used to solve the clustering problems in machine learning or data science.

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties. 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.

This paper focuses on Geyser Eruption Analysis by the use K-mean clustering algorithm. The problem statement concerns 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. In this paper the 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.

1.1 Problem Statement

- This dataset contains waiting time between two consecutive geyser eruptions (in minutes) and the duration of the eruption (in minutes) for a given geyser.
- Our model will explore the spans at which explicit geyser erupts.
- User submits numerical scatter plots for the time stretches by the utilization of K mean clustering calculation.
- When the geyser erupt, thermal energy convert into kinetic energy. So the kinetic energy can be used for industrial purpose.
- To comprehend geysers better and to anticipate parts of their way of behaving.

1.2 Motivation

- To examine data of various time stretches at which geyser eruption occurs in order to compute and compare the mean time interval of eruption
- Geysers provide a natural laboratory to study multiphase eruptive processes
- Time is important for everyone, so it gives an edge over other technologies for saving your time.
- Many real-world opportunities are associated with it.

1.3 Objectives

- This model focuses on Geyser Eruption Analysis by the use K-mean clustering algorithm.
- The problem statement concerns to examine data of various time stretches at which geyser eruptions occur in order to compute and compare the time interval of eruption.
- 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.
- Geysers provide a natural laboratory to study multiphase eruptive processes

2. Literature Review

In 2022, William F. Fagan, Reservoir computing, a form of machine learning, was used to characterize the collective behavior of 10 Yellowstone geysers. Network-level predictions of geyser behavior offered substantial improvements over individual time series. Inter-geyser distance and geyser morphology helped shape the strength of interconnections among thermal features. These improvements suggest that geysers are not independent and instead reflect the existence of a complex interconnected subsurface groundwater system.

In2022, Diana Rauwolf,this paper is based on inspection paradox of renewal theory which states that, in expectation, the inspection interval is larger than a common renewal interval, in general. For a random inspection time, which includes the deterministic case, and a delayed renewal process, representations of the expected length of an inspection interval and related inequalities in terms of covariance are shown. Data sets of eruption times of Beehive Geyser and Riverside Geyser in Yellowstone National Park.

In 2019, Thomas R. Walter ,a catalog of 73,466 eruptions of Strokkur geyser was created in Iceland, from a 1 year seismic data set. It was found that 50,135 single eruptions but only 1 sextuple eruption, while the mean waiting time increased from 3.7 min after single eruptions to 16.4 min after sextuple eruptions. Through this research it was found that the waiting time after an eruption can be predicted, while future eruption type or amplitude cannot.

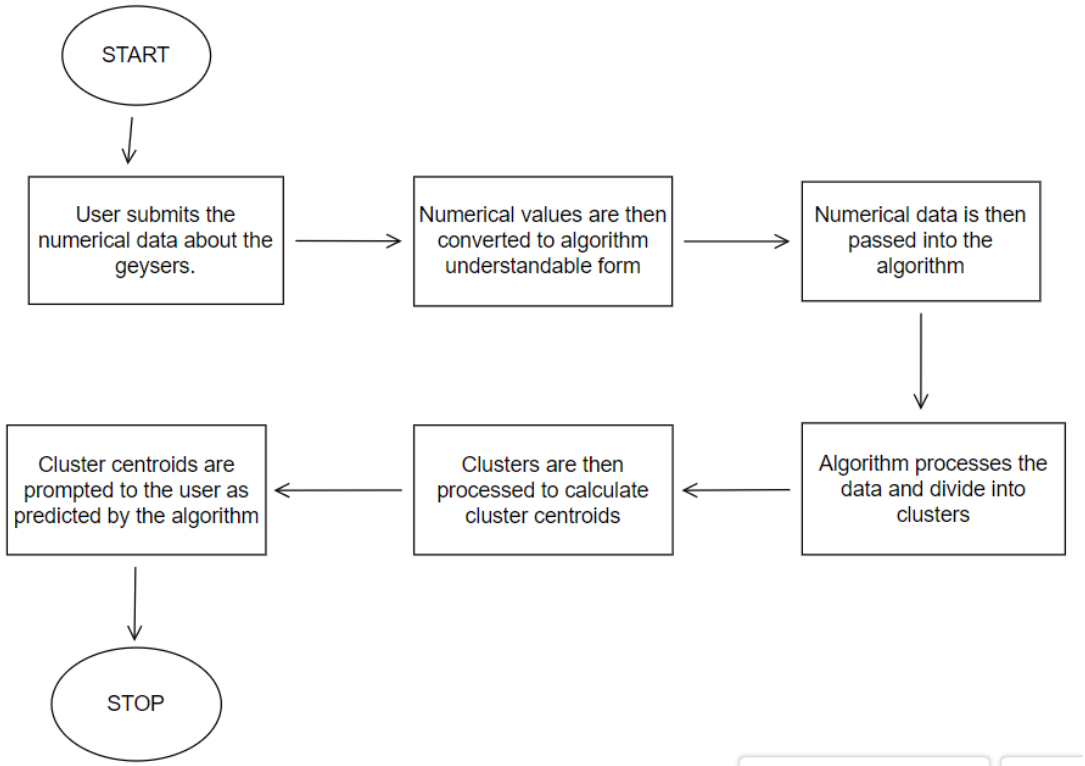
In 2019, Karplus, M. S., this paper focuses on three open questions about geysers in general and Steamboat in particular: (1) what initiates an active phase? (2) What controls the intervals between eruptions within active phases? (3) What controls the height of the eruption column? The coupled conduit and plume dynamics model suggests that a deeper reservoir produces a greater vapor mass fraction and therefore a greater exit velocity.More robust monitoring of geyser systems is needed to understand eruption variability and additional constraints on plumbing geometry would improve quantification of eruption dynamics.

In 2017 , Manga ,M, this paper mainly focuses on why do geysers exist? What determines eruption intervals, durations, and heights? What initiates eruptions?Eruptions are driven by the conversion of thermal to kinetic energy during decompression. Larger and deeper cavities permit larger eruptions and promote regularity by isolating water from weather variations.

In 2015, Namiki ,A, this paper tells that Geysers can be hydraulically connected through permeable pathways to other hot springs. The level of complexity of geyser eruptions may be controlled by the underground geometry. Over time geysers change periodicity, develop new thermal features, and shift interactions.

3. Proposed Methodology

Block Diagram



K-mean Clustering

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.

This implementation is performed using C++ & Python programming language with related libraries to achieve the task. For the K-means clustering algorithm, some specific values are assigned for K, and clusters have been created from the standardized data. Because the features could not be in the same measurement units, standardizing data involves including data with a zero mean and a one standard deviation. And the centroid calculated from the dataset are then plotted in the graph

Data Visualization

The process of finding trends and correlations in our data by representing it pictorially is called Data Visualization. Data visualization is a field in data analysis that deals with visual representation of data. It graphically plots data and is an effective way to communicate inferences from data.Using data visualization, we can get a visual summary of our data. With pictures, maps and graphs, the human mind has an easier time processing and understanding any given data. Data visualization plays a significant role in the representation of both small and large data sets, but it is especially useful when we have large data sets, in which it is impossible to see all of our data, let alone process and understand it manually.

Scatter Plots

In this project we have visualized the data using scatter plots formed by python programming language. Now let us know about, "what is a Scatter Plot?" Scatter plots are used when we have to plot two or more variables present at different coordinates. The data is scattered all over the graph and is not confined to a range. Two or more variables are plotted in a Scatter Plot, with each variable being represented by a different color.

4. Hardware Requirements:

- PC/Laptop
    - ❖ Processor – 1.4GHz (32/64 bit)
    - ❖ 2 GB RAM
    - ❖ Memory – 5 GB
- Computer Network connection (Ethernet/Wi-Fi)

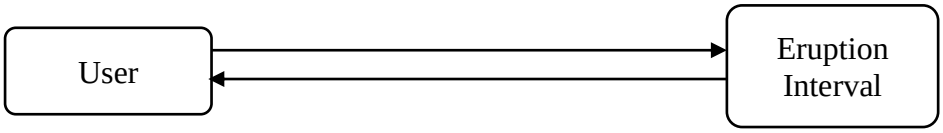
5. Software and Libraries requirements:

- Code Blocks (C++)
    - ❖ Vector
    - ❖ Math.h
    - ❖ Stdlib.h
    - ❖ Time.h
- Python 3.7.2
    - ❖ Matplotlib()
    - ❖ Plt.scatter()

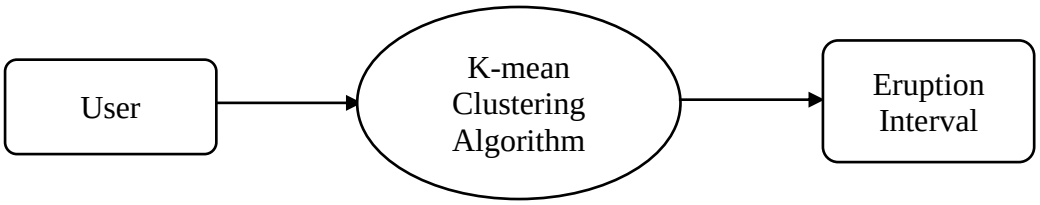
6. Design Diagrams

DFDs'

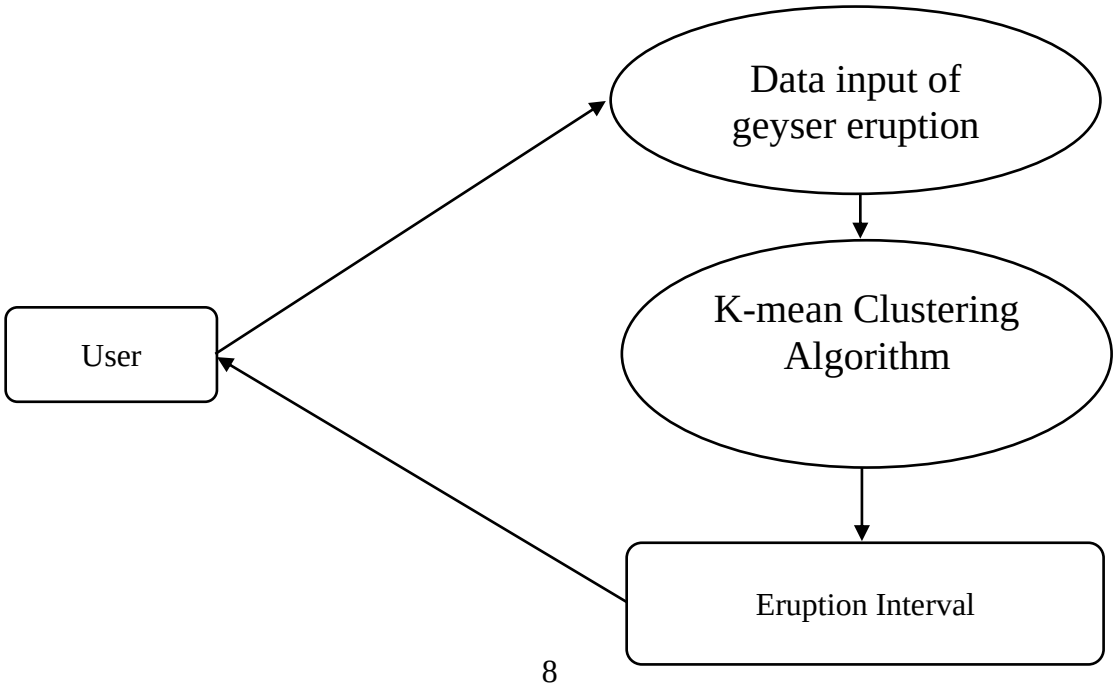
Level 0



Level 1



Level 2



7. Result and Output

- Analyzing the extracted data from web source.
- Identified the dataset to be categorical data.
- Plotting a scattered graph from this categorical data.

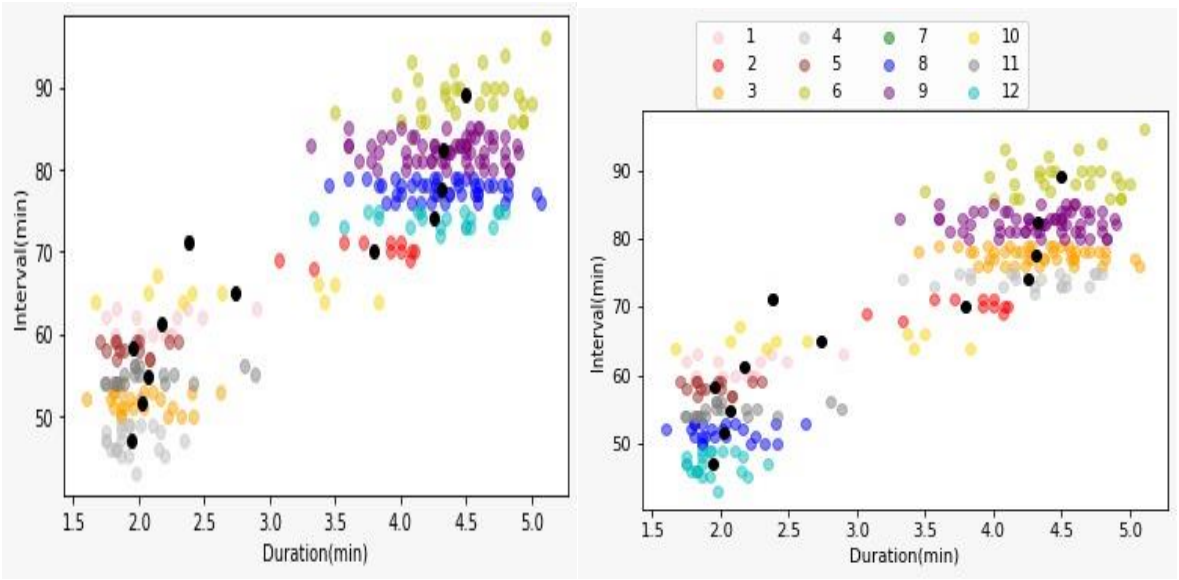
Outputs:-

```
C:\Users\HP\OneDrive\Desktop\geyser\bin\Debug\geyser.exe
Enter password: Minor123
Total Points 271
No. of clusters, Max iterations , Has name
2 100 0
Break in iteration 5

Cluster 1
Point 88: 4.517 80
Point 5: 4.533 85
Point 7: 4.7 88
Point 15: 4.7 83
Point 18: 4.8 84
Point 32: 4.467 77
Point 38: 4.833 80
Point 40: 4.783 90
Point 43: 4.567 84
Point 49: 4.633 82
Point 51: 4.8 75
Point 52: 4.716 90
Point 54: 4.833 80
Point 56: 4.883 83
Point 59: 4.567 77
Point 62: 4.5 84
Point 64: 4.8 82
Point 68: 4.7 78
Point 73: 4.5 79
Point 76: 5.067 76
Point 78: 4.567 78
Point 86: 4.933 88
Point 94: 4.817 78
Point 97: 4.667 84
Point 100: 4.9 82
Point 104: 4.5 83
```

```
C:\Users\HP\OneDrive\Desktop\geyser\bin\Debug\geyser.exe
Cluster 2
Point 2: 1.8 54
Point 4: 2.283 62
Point 6: 2.883 55
Point 9: 1.95 51
Point 11: 1.833 54
Point 14: 1.75 47
Point 16: 2.167 52
Point 17: 1.75 62
Point 19: 1.6 52
Point 21: 1.8 51
Point 22: 1.75 47
Point 27: 1.967 55
Point 33: 3.367 66
Point 36: 2.017 52
Point 37: 1.867 48
Point 39: 1.833 59
Point 42: 1.883 58
Point 44: 1.75 58
Point 47: 3.833 64
Point 48: 2.1 53
Point 50: 2 59
Point 53: 1.833 54
Point 55: 1.733 54
Point 58: 1.667 64
Point 61: 2.233 59
Point 63: 1.75 48
Point 65: 1.817 60
Point 69: 2.067 65
Point 72: 1.967 56
Point 75: 1.983 62
Point 77: 2.017 60
Point 84: 2.633 65
```

Scatter Plots:-



8. References

1. Quantifying Interdependencies in Geyser Eruptions at the Upper Geyser Basin, Yellowstone National ParkWilliam F. Fagan,Anshuman Swain,Amitava Banerjee,Hamir Ranade,Peter Thompson,Phillip P. A. Staniczenko,Barrett Flynn, JeffersonHungerford, Shaul Hurwitz

2. Quantifying the Inspection Paradox with Random Time, Diana Rauwolf, Udo KampsInstitute of Statistics, RWTH Aachen University https://www.tandfonline.com/doi/full/10.1080/00031305.2022.2151510

3. Eruption Interval Monitoring at Strokkur Geyser, Iceland. Eva P. S. Eibl,Sebastian Hainzl,Nele I. K. Vesely,Thomas R. Walter,Philippe Jousset,Gylfi Páll Hersir,Torsten Dahm

4. Reed, M. H. ; Barth, A. ; Girona, T. ; Hajimirza, S. ; Hurwitz, S. ; Karlstrom, L. ; Karplus, M. S. ; Manga, M. ; Muñoz-Saez, C. ; Rashtbehesht, S. H. ; Wu, S. M. “Multiparameter Study of Eruptive Behavior at Steamboat Geyser, Yellowstone ”

5. Hurwitz, S., & Manga, M. (2017). The Fascinating and Complex Dynamics of Geyser Eruptions. Annual Review of Earth and Planetary Sciences, 45(1), 31-59. http://dx.doi.org/10.1146/annurev-earth-063016-015605 Retrieved from https://escholarship.org/uc/item/5jd7x994

6. Munoz-Saez, C., Namiki, A., & Manga, M. (2015). Geyser eruption intervals and interactions: Examples from El Tatio, Atacama, Chile. Journal of Geophysical Research: Solid Earth, 120(11), 7490-7507. http://dx.doi.org/10.1002/2015JB012364 Retrieved from https://escholarship.org/uc/item/77k43002

APPENDIX A: GLOSSARY