Project Report

On

SCHOOL REVIEWS ANALYSIS AND RECOMMENDER SYSTEM

Submitted in partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics (PG-DBDA)

From Know-IT(Pune)

Guided by:

Anay Tamhankar Sir.

Prasad Deshmukh Sir.

Submitted By:

Nikhil Datar (230343025011)

Vaibhav Gangurde (230343025015)

Rushikesh Gade (230343025014)

Kunal Barapatre (230343025030)

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Nikhil Datar (230343025011)

Vaibhav Gangurde (230343025015)

Rushikesh Gade (230343025014)

Kunal Barapatre (230343025030)

Have successfully completed their project on

School Reviews Analysis And Recommender System

Under the guidance of Anay Tamhanakar sir and Prasad Deshmukh sir

ACKNOWLEDGEMENT

This project " SCHOOL REVIEWS ANALYSIS AND RECOMMENDER SYSTEM " was a great learning experience for us and we are submitting this work to CDAC Know-IT (Pune).

We all are very glad to mention the name of Anay Tamhanakar sir and Prasad Deshmukh sir for his valuable guidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to Mr. Vaibhav Inamdar Manager (Know-IT), C-DAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartfelt thanks goes to Mrs. Dhanashree (Course Coordinator,PG-DBDA) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC Know-IT, Pune.

From:

Nikhil Datar (230343025011)

Vaibhav Gangurde (230343025015)

Rushikesh Gade (230343025014)

Kunal Barapatre (230343025030)

TABLE OF CONTENTS

Abstract

Our Project **School Reviews Analysis and Recommender System** focuses on utilizing the power of data analytics and machine learning to enhance the quality of schools through comprehensive review analysis and personalized recommendations. In today's digital age, where information is easily accessible and user-generated reviews play a crucial role in decision-making. our project addresses the need for an intelligent system that assists parents in making informed choices about schools.

The primary objective of this project is to develop a robust framework that collects, analyzes, and visualizes school reviews from various online sources. Natural language processing techniques are employed to extract sentiment, key themes, and opinions from textual reviews. The analysis is further extended to understand overall sentiment trends, identifying strengths and weaknesses of schools, and recognizing patterns that contribute to positive or negative feedback.

Additionally, our project uses the Recommender System to that leverages the insights gained from the review analysis. By employing content-based recommendation the system can provide tailored suggestions for schools that align with the preferences and priorities of users. Using this we achieve the freedom for the parents to select the school and to see the positive and negative overall average ratings of the school according to the various parameters.

Last but not the least we have created various charts and graphs which will help the user to get visuals of the data in the highest informative and understandable format. Hence this project fills the gap between having the data and taking the informed decision for selecting the appropriate school comparing on the basis of various parameters.

INTRODUCTION

School Reviews Analysis and Recommender System focusses on creating the model based upon various machine learning algorithms and data preprocessed using PySpark. IT involves the unsupervised machine learning algorithm like k-means Clustering, AI based field like Natural Language Processing, recommender technique like content based recommender system and visualization created using Tableau.
Comparing and finding the appropriate school is cumbersome task. This commonly seen problem is tried to solve to some extent is the main motive and the foundation of the idea of this project.

Datasets and features

Data used in this project scraped using the Octoparse Software which provides the interface to scrape the data from the webpages and the data is collected in the different csv files of according to the schools which then combined to form a raw dataset which is used for the further processing and evaluation.
Raw Dataset contains 4 columns:

name (String),
school name(String),
content(String),
ratings(String).

total 82469 record are there in the raw dataset.

SYSTEM REQUIREMENTS

Hardware Requirements:

- Platform – Windows
- RAM – 8 GB of RAM (Recommended)
- Peripheral Devices – Mouse, Keyboard, Monitor
- A network connection for data recovering over network.

Software Requirements:

- Octoparse
- Python 3
- PySpark
- Google Colab and Jupyter
- Tableau
- OS – Windows

FUNCTIONAL REQUIREMENTS

❖ **Python 3:**

- Python is a general purpose and high level programming language.
- It is use for developing desktop GUI applications, websites and web applications.
- Python allows to focus on core functionality of the application by taking care of common programming tasks.
- Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, Small Talk, and Unix shell and other scripting languages

❖ **PySpark:**

- PySpark is the Python API for Apache Spark.
- Used to perform real-time, large-scale data processing in a distributed environment using Python
- PySpark provides a PySpark shell for interactively analyzing the data.
- PySpark supports all of Spark's features such as Spark SQL, DataFrames, Structured Streaming, Machine Learning (MLlib) and Spark Core.

❖ **Tableau:**

- Data visualization is the graphical representation of information and data.
- It helps create interactive elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- Tableau is widely used for Business Intelligence but is not limited to it.
- It helps create interactive graphs and charts in the form of dashboards and worksheets to gain business insights.
- All of this is made possible with gestures as simple as drag and drop.

DATA PREPROSSING

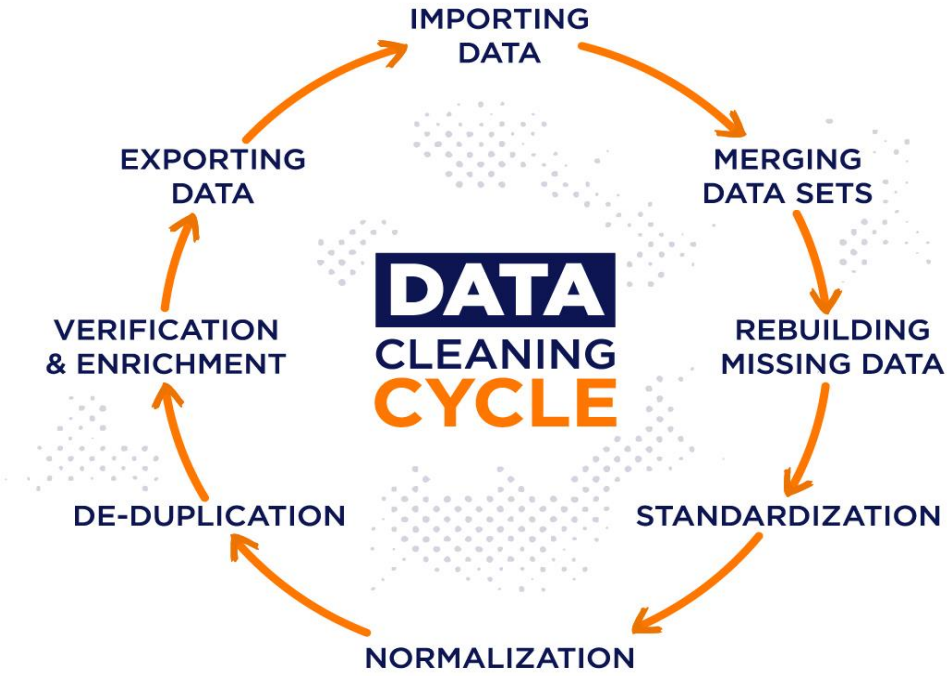


Fig: Data Cleaning Process

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting.

After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

Here in this project after collecting the data from octoparse we have preprocessed the data using PySpark. We droped the null values present in the dataset and also the irrelevant columns like “names” so that the final dataset created which is used for further operations.

SYSTEM ARCHITECTURE

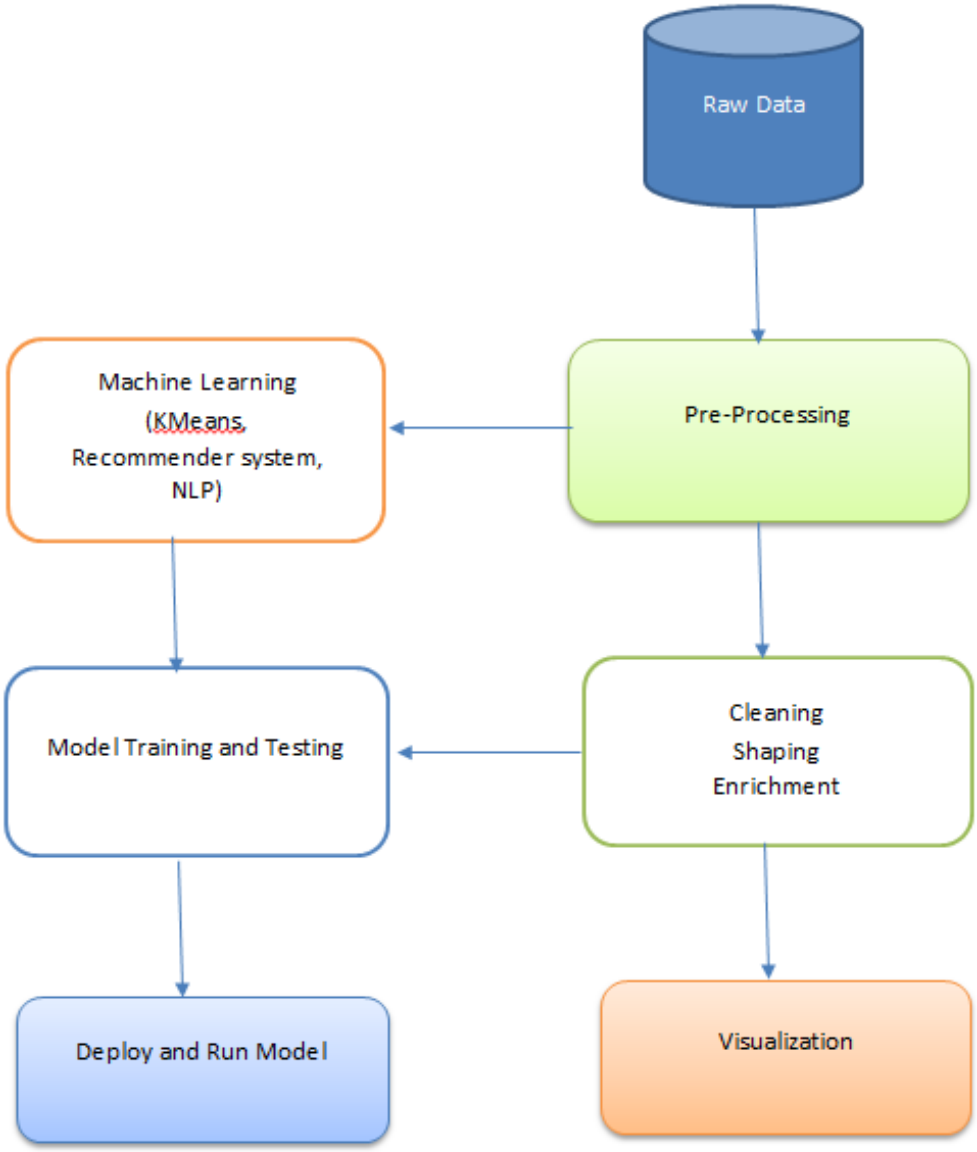
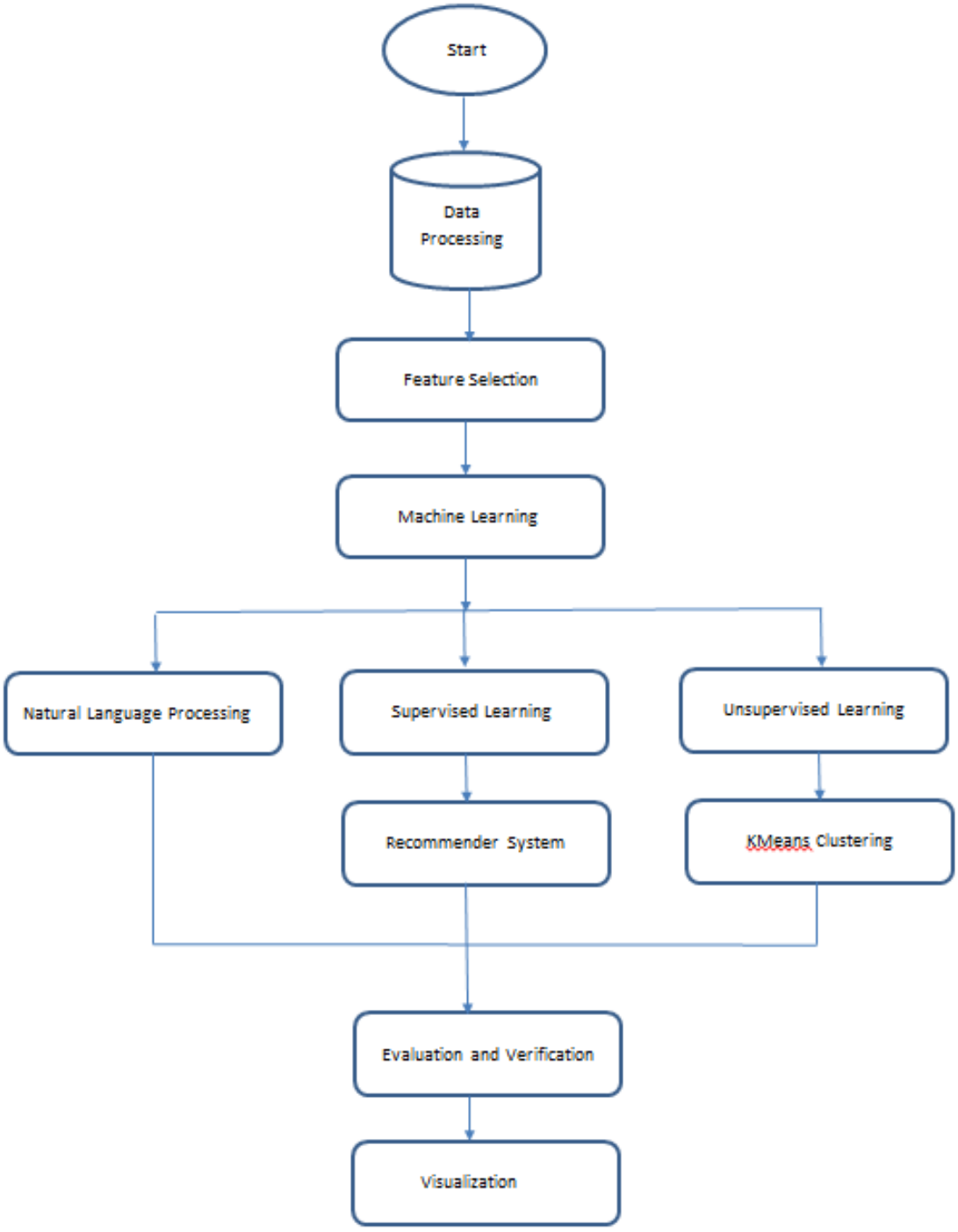


Fig: System Architecture Of Crime Prediction Project

METHODOLOGY



Methodology of School Reviews Analysis And Recommender System

MACHINE LEARNING ALGORITHMS

In this project we have used various machine learning algorithms like K-Means Clustering, Natural Language Processing , Recommender System and variour tools for data preprocessing.

Machine learning is the research that explores the development of algorithms that can learn from data and provide predictions based on it. They were mainly used for classification and prediction. In this project we use various machine learning algorithms which are as follows:

K-Means Clustering

- K-means clustering is a popular unsupervised machine learning technique used for partitioning a dataset into distinct groups, or clusters, based on their similarity.
- It aims to group data points that are close to each other while being far from points in other clusters.
- The algorithm operates iteratively and is widely used in various fields, including data analysis, image processing, and customer segmentation.
- Algorithm Steps :-

Initialization

- ❖ Assignment
- ❖ Update Centroids
- ❖ Repeat
- ❖ Final Clustering

Advantages:

- Simple and computationally efficient.
- Scalable to large datasets.
- Can be applied to various types of data, assuming a reasonable notion of distance can be defined.
- Useful for data exploration, pattern recognition, and preprocessing.

Limitations:

- Requires a predefined number of clusters (K).
- Sensitive to initial centroids, which can lead to different outcomes.
- Assumes clusters are spherical and equally sized, which might not hold for all types of data.
- Can struggle with non-linear or complex data distributions.

Natural Language Processing

- Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written referred to as natural language. It is a component of artificial intelligence.
- NLP enables computers to understand natural language as humans do. Whether the language is spoken or written.
- Natural language processing uses artificial intelligence to take real-world input, process it, and make sense of it in a way a computer can understand. Just as humans have different sensors such as ears to hear and eyes to see, computers have programs to read and microphones to collect audio. And just as humans have a brain to process that input, computers have a program to process their respective inputs.
- At some point in processing, the input is converted to code that the computer can understand.

Advantages

- Ability to automatically make a readable summary of a larger, more complex original text.
- useful for personal assistants such as Alexa, by enabling it to understand spoken word.
- easier to perform sentiment analysis.
- improved accuracy and efficiency of documentation

Limitation

- Requires clarification dialogue.
- NLP system doesn’t have a user interface that lacks features that allow users to further interact with the system.
- In complex query language, the system may not be able to provide the correct answer it a question that is poorly worded or ambiguous.

Recommender System

- A recommendation system is an artificial intelligence or AI algorithm, usually associated with machine learning that uses Big Data to suggest or recommend additional products to consumers.

- These can be based on various criteria, including past purchases, search history, demographic information, and other factors. Recommender systems are highly useful as they help users discover products and services they might otherwise have not found on their own.

Types of Recommender systems

- Content based filtering
- Collaborative filtering
    - User
    - Item

Advantages

- Easy recommendation make less search and some time end up on good deals
- Speed up the process of decision
- Cost-Effectiveness
- Help to take decision

Limitation

- If system recommends with bias the client will be landing into wrong deal
- Chances are that some websites or data resources may suggest wrong information on basis of that model will give wrong information
- Cold start problem (user provide input that totally new for model)

DATA VISUALIZATION AND REPRESENTATION

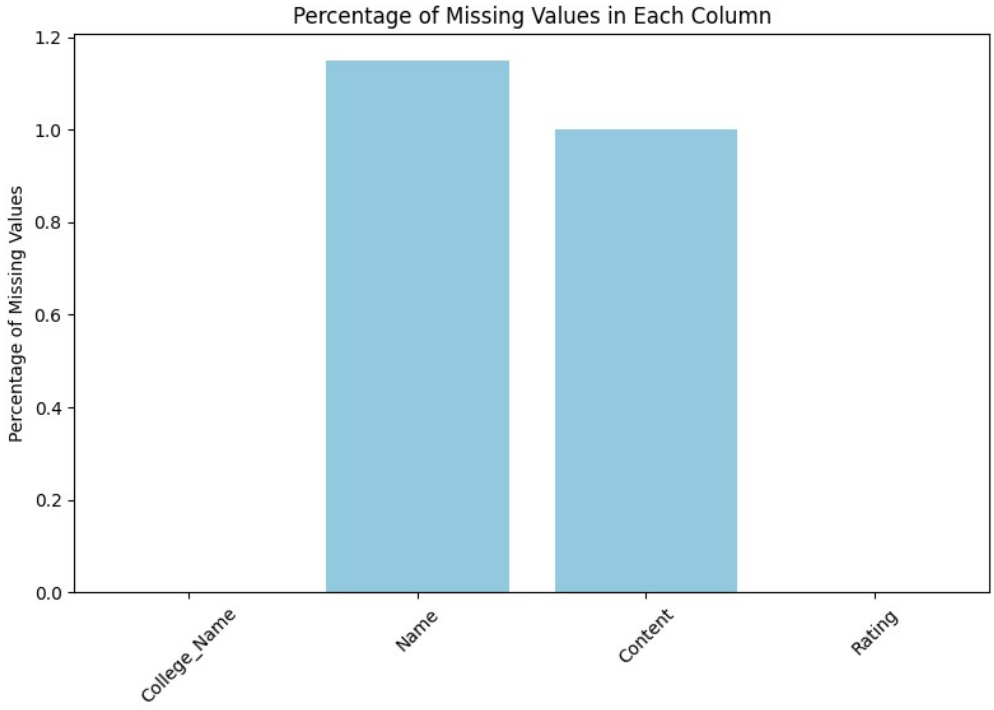


Fig- Percentage of null values per categories

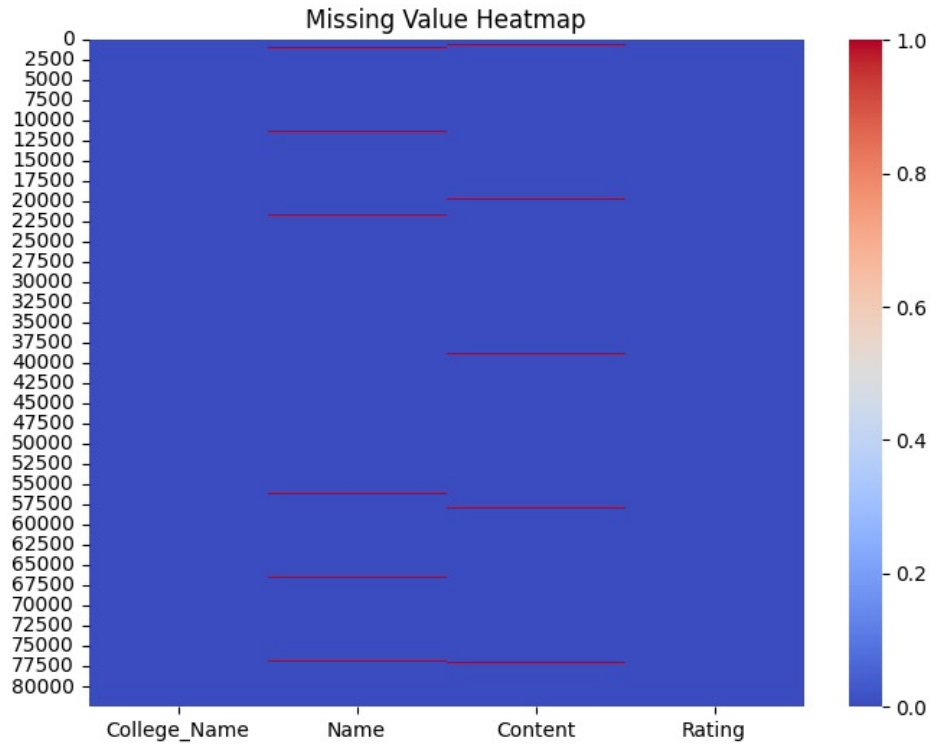


Fig- Null Values Heat Map

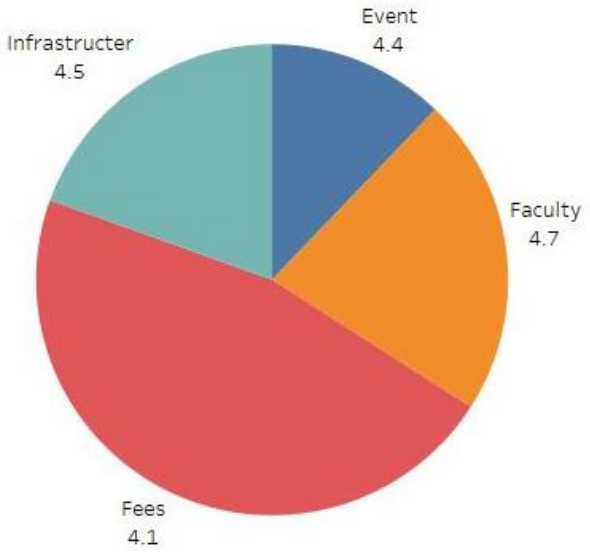


Fig:- Pie Chart of Average Rating per Categories

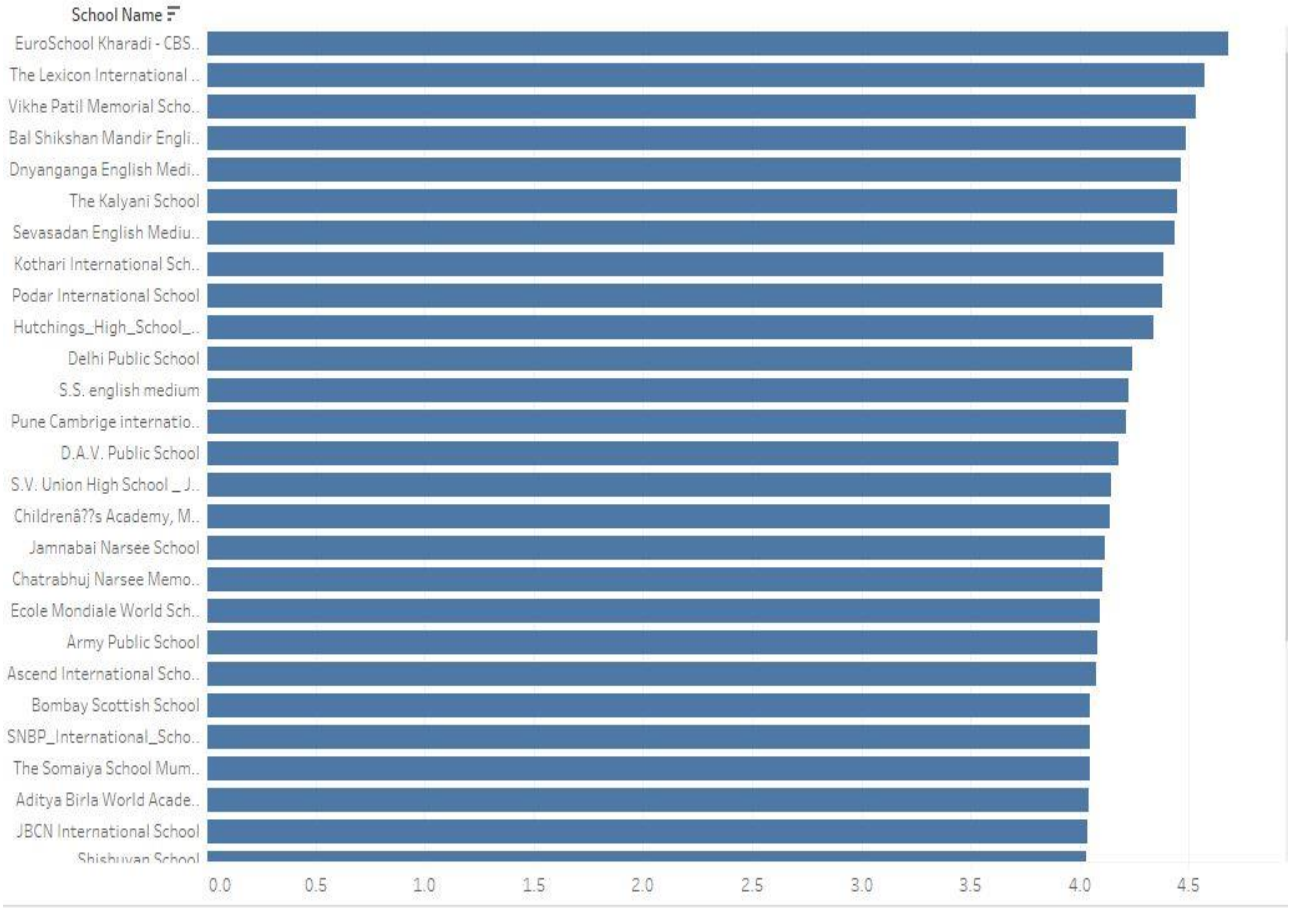


Fig:- Line chart of Schools in descending order of avg rating of category selected.

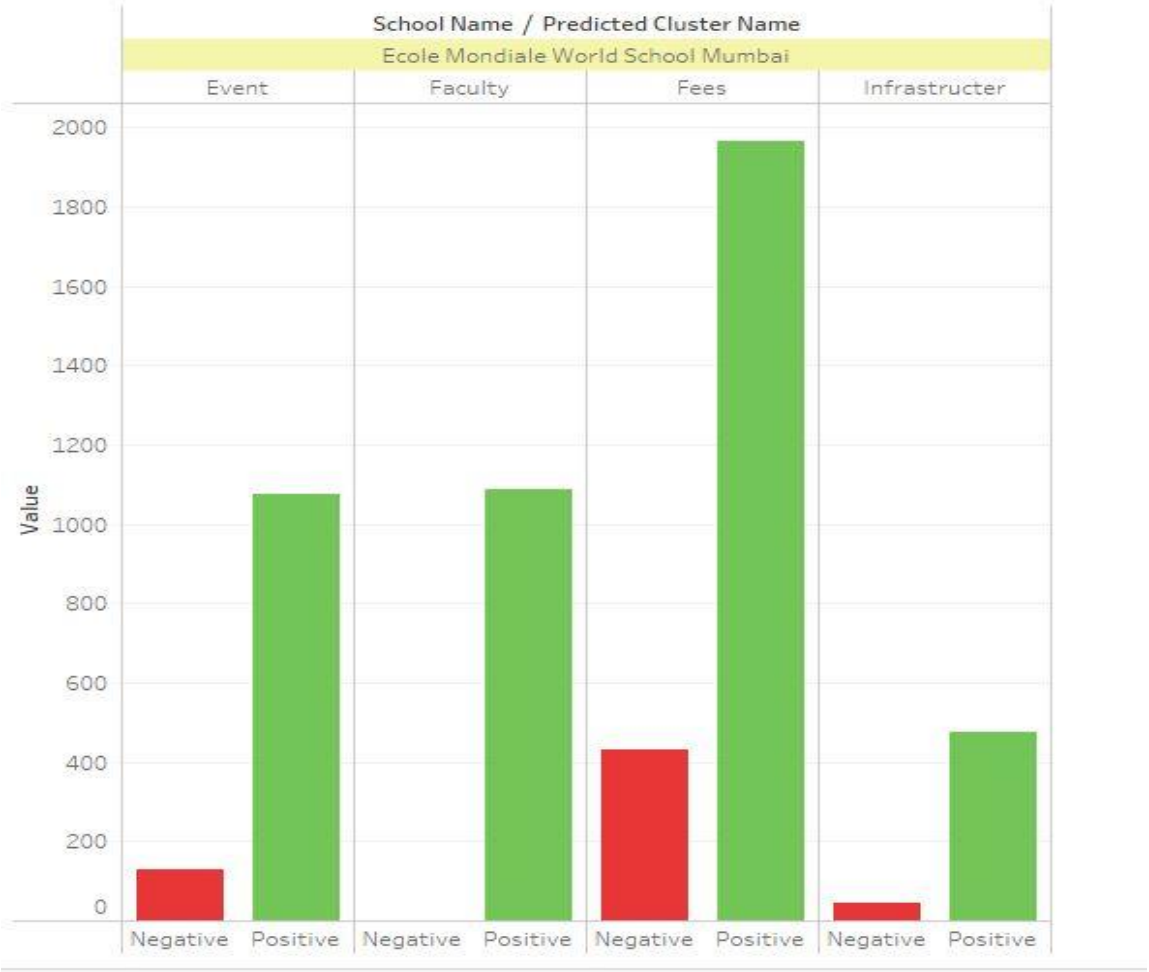


Fig:- Distribution of positive and negative reviews per category per selected school

FUTURE SCOPE

The "School Reviews Analysis and Recommender System" project holds a promising future.

The scope of this project can be expanded in different directions:

Refined Recommendation Algorithms:

As your system gathers more data, you can implement more advanced recommendation algorithms. Collaborative filtering, matrix factorization, and deep learning techniques can enhance the accuracy of school recommendations based on user preferences and historical data

.

Personalized User Profiles:

Develop a system that allows users to create detailed profiles, including their interests, location, educational goals, and preferences. This will enable the system to provide more tailored and relevant recommendations.

Geographical Expansion:

If your system is currently focused on a specific region, consider expanding its coverage to include schools from different cities, states, or even countries, making it a comprehensive tool for users seeking education options worldwide.

Integration with Educational Data Sources:

Collaborate with educational institutions and official databases to access accurate and up-to-date information about schools. This could include academic performance, extracurricular activities, teacher qualifications, and more.

Real-time Updates:

Implement a mechanism to provide real-time updates about schools, including news, events, changes in faculty, and any noteworthy developments.

Mobile App Development:

Create a mobile app to make your platform more accessible to users on smartphones and tablets. A mobile app can also leverage device features like location services for location-based recommendations.

CONCLUSION

If incorporated on the large scale we can create such a big recommendations systems and through websites and applications it will really help people get recommendations about whatever they want that too appropriate with the factores they are interested in. In conclusion, the "School Reviews Analysis and Recommender System" project presents a comprehensive solution to the challenges faced by students, parents, and educators in making informed decisions about educational institutions. Through the systematic collection and analysis of user-generated reviews, this project provides valuable insights into the strengths and weaknesses of different schools. The integration of AI-driven recommendation algorithms further enhances the user experience by offering personalized suggestions based on individual preferences and requirements.

.

REFERENCES

https://spark.apache.org/docs/latest/api/python/index.html

https://openapi.octoparse.com/en-US/

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

https://www.nltk.org/

https://en.wikipedia.org/wiki/Recommender_system

https://www.tableau.com/support/help