# pure logo copy

# Janardan Bhagat Shikshan Prasarak Sanstha’s

# CHANGU KANA THAKUR ARTS, COMMERCE, SCIENCE COLLEGE

# Re-accredited ‘A+’ Grade by NAAC

# College with Potential for Excellence Status Awarded by

# University Grants Commission

# ‘Best College Award’ by University of Mumbai

# New Panvel (W)

PROJECT PROPOSAL ON

**“Stock Market Prediction using Data Mining Techniques”**

## By

## MR. RAHUL BALIRAM PATIL

(M.Sc.-II (SEM-III) COMPUTER SCIENCE)

Under the Guidance of

Prof. Ms.

Department of Computer Science

**CERTIFICATE**

This is to certify that the Project Proposal entitled

**“Stock Market Prediction using Data Mining Techniques”**

Is successfully completed by **Rahul Baliram Patil**, Examination Seat Number under the guidance of **Prof. Ms.,** during the academic period of 10th June, 2017 to 28th Dec, 2017as per the Syllabus, and the fulfilment for the completion of the M.Sc.-II (Semester-III) in the Computer Science of **University of Mumbai**. It is also to certify that this is original work of the candidate done during academic year 2017-2018.

**Place:**

**Date:**

**Internal Examiner Head of Department**

**External Examiner**

**ACKNOWLEDGEMENT**

No project is ever completed without the guidance of those expert have already traded this past before and hence become and master of it and as a result, our leader. So I would like to take this opportunity to take all those individuals how have helped me in visualizing the project.

It is indeed a matter of great pleasure and proud privilege to be able to present this project proposal on “**Stock Market Prediction using Data Mining Techniques**”.

The completion of the project work is a milestone in student life and its execution is inevitable in the hands of guide. I express my deep gratitude to my project guide **Prof.**  for providing timely assist to my query and guidance that they give their experience in this field past many year. They had indeed been a lighthouse for us in this journey.

I would like to tender our sincere thanks to the H.O.D. **Prof. Mrs. P. M. Jadhav** and all the teachers for their co-operation.

I would also like to express our deep regards and gratitude to the Principal **Dr. .**

I would wish to thank the non – teaching staff and my friends who have helped me all the time in one way or the other.

Really it is highly impossible to repay the depth of all the people who have directly or indirectly helped me for performing the project.

**“STOCK MARKET PREDICTION USING DATA MINING TECHNIQUES”**

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Title** | **Page No.** |
| 1 | Introduction | 6-8 |
| 2 | Related Work | 9-10 |
| 3 | Objective | 11 |
| 4 | Methodology | 12 |
|  | 1. Frequent Itemset Using Datamining |
| * 1. Apriori Algorithm |
|  | 1. Classication | 13 |
|  | 1. Clustering | 13 |
| 3.1 K-Means | 14 |
|  | 1. Dataset | 15 |
|  | 1. System Configuration | 16 |
| 5 | References | 17-18 |

**INTRODUCTION**

Today we live and breathe data. Forecasting the stock data is an important financial subject which involves an assumption that the fundamental information publicly available to the past predictive relationship for future stock return. Stock market prediction contain uncovering market trends, planning investment tactics, identifying the best time to purchase the stock and which stock to purchase. A stock exchange or business sector is a non-direct, non-parametric framework that difficult to model. It is mix speculators which need purchase or offer of hold a share at a specific time. Prediction will continue to be an exciting locale of research, making scientists in the analytics field always desiring to enhance the existing forecasting models. The motivation is that companies and individuals are empowered to make investment decision to develop viable system about their future endeavours.

Stock trend forecasting is considered as one of the most difficult tasks to achieve in money related gauging because of the difficulty to achieve in money related gauging because of the difficulty in the multifaceted world of stock market. Many investor in the stock market are finding a technique that could guarantee easy profiting by forecasting the stock trends and minimize the risk of investing. This motivates the researchers in the domain field to delve and develop new forecasting models. Time series data analysis techniques use verifiable information as the premise for evaluating future results. Time series data can be defined as numerical data collected in a particular sequence over a period of time at regular intervals. The time series data can include the values collected at the end of every week, month, quarter, or year etc. The intention is to find if there is any link between the data collected so far and in what way does the data changes. In order to reduce the risk of investment, exchange the securities between the seller and buyer are facilitated by the stock exchanges. A stock exchange is an organization or a place where the stock traders or investors can deal with stock. Some example of stock market organization S&P, NSE, BSE, NASDAQ etc.

R is a programming language and environment for statistical processing and graphics. The   
R is generally utilized among analysts the data and data excavators for statistical programming and data analysis. R language was created by Ross Ihka and Robbert Gentleman at the University of Auckland, New Zeland and is presently maintained by the R development core team. Amid the most recent decade, the energy originating from both scholarly world and industry has lifted the R programming dialect to turn into the most essential tool for computational insights, perception and data science. Around the world, a large number of analysts and information researchers use R language to take care of their most difficult issues in the fields going from computational science to quantitative promoting.

**RELATED WORK**

There have been several studies conducted on YouTube due to the face is one of the most popular video sharing website. The studies focused on different characteristics of videos. In [1], Zhou studied the impact of YouTube recommendation system on video views. In [2], M. Cha analyzed the popularity life-cycle of videos, the intrinsic statistical properties of requests and their relationship with video.

Several previous works studied the impact of YouTube recommendation system and uploaders on total view count of videos. There are some other works focusing on the impact of videos categories on the size of YouTube. For [3] Filippova studied the video categories to YouTube and considered the task of assigning categories to YouTube video based on the text information related video title, user tag, description and viewers, comments.

YouTube uploaders are the central agents in the YouTube phenomenon. We conduct extensive measurement and analysis of YouTube uploaders. YouTube uploaders and demonstrated the positive reinforcement between on-line social behaviour and uploading behaviour Ding studied [4].

As the world’s largest video sharing website, YouTube hosts a large number of mostly user-generated videos that are viewed by millions of user each day is based on count of YouTube videos via random prefix sampling. They designed an unbiased estimator of total number of YouTube videos [5].

Video recommendation system that YouTube uses and it role in increasing the total number of views for video and the system aim to predict items that may be interest o users. There is no need of explicit request for information the system learns about user and generates personalised suggestions [6]. Diversity of scenarios and domain make the task of finding relevant item.

Asur provided a theoretical basis for analyzing the formation, persistence and decay of trends for the trending topic on Twitter [7]. However, to the best of our knowledge, YouTube trending videos have not studied thoroughly.

**OBJECTIVE**

1. How many views do our trending videos have? Do most of them have a large number of views? Is having a large number of views required for a video to become trending videos.
2. To analyze how much likes and comment count by instead of views.
3. Which video in given dataset remain trending videos contain fully-capitalized word in their titles.
4. To analyze how are views, likes, dislike, comment, title, and other attributes correlates with each other? How are they connected?
5. what are the length of trending videos titles, is this length related to the videos become popular or trending
6. Which YouTube channels have the largest number of trending videos and which video category (e.g. Entertainment, Gaming, Comedy, etc. ) has the largest number of trending videos.
7. Which video remained the most on the trending videos list

**METHODOLOGY**

1. **FREQUENT ITEMSET USING DATAMINING**

Much research has been focused on finding efficient algorithm for mining large Itemset. In this work we will try to analyse the frequent Itemset using the algorithms of datamining like the Apriori algorithm, K-Apriori algorithm, and these algorithms are explained in detail below.

* 1. **APRIORI ALGORITHM**

Apriori [4] is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

Apriori algorithm for Frequent Itemset Mining

Cdn: Candidate itemset of size n

Ln: frequent itemset of size n

L1 = {frequent items};

For (n=1; Ln!= ; n++)

Do begin

Cdn+1 = candidates generated from Ln;

For each transaction T in database do

Increment the count of all candidates in Cdn+1 that are

contained in T

Ln+1= candidates in Cdn+1 with min\_support

End

Return Ln

1. **CLASSIFICATION**

In [classification](https://en.wikipedia.org/wiki/Statistical_classification), inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more ([multi-label classification](https://en.wikipedia.org/wiki/Multi-label_classification)) of these classes. This is typically tackled in a supervised way.

Classification is used to classify each item in a set of data into one of predefined set of classes or groups. The data analysis task classification is where a model or classifier is constructed to predict categorical labels (the class label attributes). Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

Classification is a model finding process that is used for partioning the data into different classes according to some constrains. In other words we can say that classification is process of generalizing the data according to different instances. Several major kinds of classification algorithms including C4.5, k-nearest neighbor classifier, Naive Bayes, SVM, Apriori, and AdaBoost, etc.

**Our** Analysis determines the Optimal Resource Organization to facilitate the Construction Services for Different Area or Sector including public and private or government for identification and evaluation of construction capability, applicability, and level of the uptake within the contract.

1. **CLUSTERING**

Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding to data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis of E-Tendering.

The personal data combined with shopping, location, interest, actions and an infinite number of indicators, can be analysed with this methodology, providing very important information and trends. Examples of this are the market research, marketing strategies, web analytics, and a lot of others.

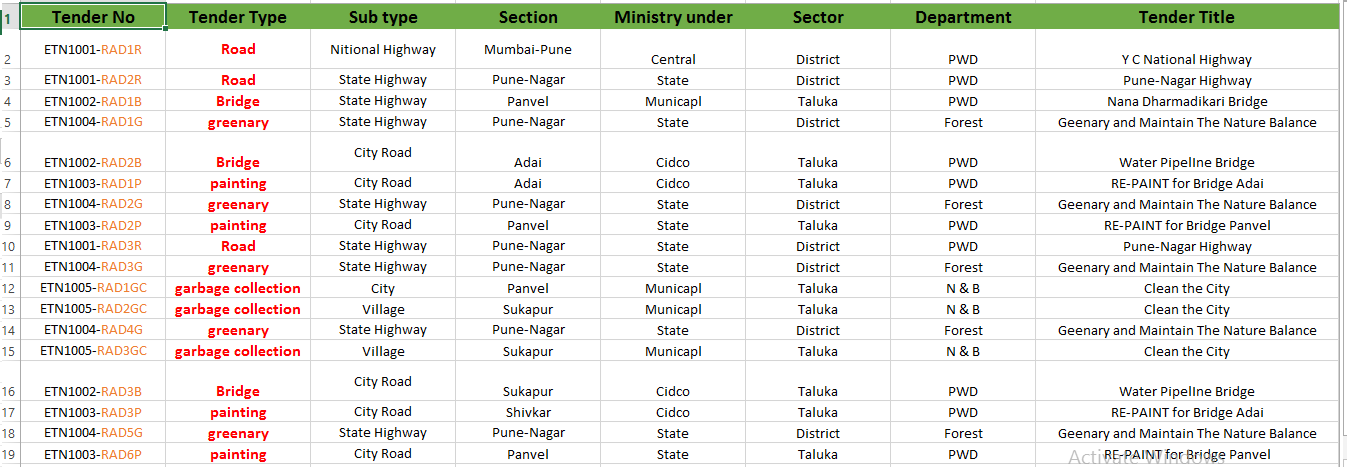
**3.1 K-means Clustering**

K-means is an iterative clustering algorithm in which items are moved among setsof clusters until the desired set is reached. This technique is used to classify thedata which have no previous knowledge about the data or the training set. Theparameter K denotes the amount of clusters required to partition the data. The ideaof this clustering technique is, given K number of clusters we can define Kcentres, one for each cluster based on all samples belonging to a cluster. Thesecentres must be placed far away from each other and then associate each sampleto the cluster that has the closest centroid.

**In our**E-Tendering Analysis**K-means Clustering**can be usedtoexamine the Organizations that provide Constructions Services in Particular area that enables in Public or Government Section Tenders of Construction Service.

1. **DATASET:**

The dataset used for this work can make us understand the algorithms more clearly.

The effects algorithm can be understood in much depth

1. **SYSTEM CONFIGURATION**

**H/W System Configuration:**

* System - Core-2-Due 2.4Ghz
* Speed - 2.4Ghz
* RAM - 4GB
* Hard Disk - 500GB
* Keyboard - Standard Windows Keyboard(Neosoft)
* Mouse - Standard Intex
* Monitor - 19” LED

**S/W System Configuration:**

* Operating system - Win 10
* Database - MySql/Excel
* Language - R/Weka/Python

**REFERENCES**

**Journal Reference**

**[1]** Agrawal, R.; Imieliński, T.; Swami, A. (1993), Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207. doi:10.1145/170035.170072. ISBN 0897915925, "Mining association rules between sets of items in large databases".

**[2]** https://en.wikipedia.org/wiki/Apriori\_algorithm, Rakesh Agrawal and RamakrishnanSrikant Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

**[3]** Sergey Brin, Rajeev Motwani, Jeffery D. Ullman Department of Computer Science Stanford University {Serjey, rajeev, ullman}@cs.stanford.edu, Shalom Tsur, R&D Divison, Hitachi America Ltd. [tsur@hitachi.com](mailto:tsur@hitachi.com), “Dynamic Itemset Counting and Implication Rule for Market Basket Data[1997]”.

**[4]** ShahriyarMohammadiIT group, Faculty of industrial engineeringK.N.Toosi University of technologyTehran, Iran,HediyJahanshahiIT group, Faculty of industrial engineeringK.N.Toosi University of technologyTehran, Iran,“A Secure E-Tendering system”.

**[5]** Eric ChoenWeng Lou Research Institute for the Built and Human Environment (BuHu), University of Salford, UK.,Mustafa Alshawi Research Institute for the Built and Human Environment (BuHu), University of Salford, UK, “Critical Success Factors For E-Tendering Implementation In Construction Collaborative Environments: People And Process Issues”

**[6]** Rakesh K. Arora Krishna Engineering College Ghaziabad, UP, India,Manoj K. Gupta Rukmini Devi Institute of Advanced Studies Rohini, Delhi, India, “e-Governance using Data Warehousing and Data Mining”

**[7]** Tejas C. Patil,Post-Graduation Student Civil Engineering Department, Savetribai Phule Pune University, Dr. D Y Patil School of Engineering & Technology, Lohegaon, Pune, Maharashtra 412105, India, Ashish P. Waghmare, Assistant Professor Civil Engineering Departments, Savetribai Phule Pune UniversityDr. D Y Patil School of Engineering & Technology, Lohegaon, Pune, Maharashtra 412105, India.P.S.Gawande, Assistant Professor Civil Engineering Departments, RashtrasantTukadojiMaharaj Nagpur University TJawaharlal Darda Institute of Engineering and Technology,Yavatmal, Maharashtra 445002, India.”Tender and Bidding Process in Construction Projects”

* **URL Reference**

1. [http://web.fhnw.ch/personenseiten/taoufik.nouri/Data%20Mining/Course/Case%20Study/ PA-Tutorial/mba.html](http://web.fhnw.ch/personenseiten/taoufik.nouri/Data%20Mining/Course/Case%20Study/%20PA-Tutorial/mba.html)
2. <https://en.wikipedia.org/wiki/Association_rule_learning>
3. <https://etenders.gov.in/eprocure/app>
4. <https://maharashtra.etenders.in/mah/index.asp>