# HOME IN : A Real Estate Website

Project report submitted in partial fulfilment of the requirement for the degree of

Bachelor of Technology

in

**Computer Science and Engineering/Information Technology**

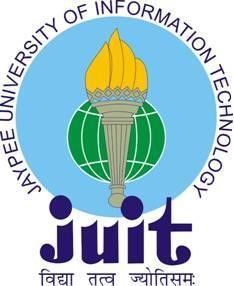
By

Ashwani Tandon (151465)

Under the supervision of

Manikant Kumar

to



Department of Computer Science & Engineering and Information Technology

**Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh**

.

i

# Candidate’s Declaration

I hereby declare that the work presented in this report entitled **“Home In :A Real Estate Website”** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Information Technology** submitted in the Department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology, Waknaghat is an authentic record of my own work carried out over a period from February 2019 to May 2019 under the supervision of  **Mr. Manikant Kumar.**

The work done embodied in the report has not been appeased for the award of any other degree or diploma.

|  |  |
| --- | --- |
|  | (Student Signature) |
|  | Ashwani Tandon  (151465) |

This is to certify that the above affirmation made by the candidate is true to the best of my knowledge.

Mr. Manikant Kumar

Systems Engineer, ETA Department

Infosys Pvt. Ltd, Mysore, Karnataka

ii

# Acknowledgement

I would like to take the opportunity to thank and express my deep sense of gratitude to my mentor and project guide **Mr. Manikant Kumar** for his immense support and valuable guidance without which it would not have been possible to reach at this stage of our final year project.

I am also obliged to all my faculty members for their valuable support in their respective fields which helped me in reaching at this stage of my project. My thanks and appreciations also go to my colleagues who have helped me out with their abilities in developing the project.

Date: Ashwani Tandon

iii

# Table of Contents

## 1. INTRODUCTION

1.1 From Small to Big Social Data Analysis……………………………………01

1.2 Project overview………………………………………………………………......01

1.3 Objectives ………………………………………………….………………....…..01

1.4 Methodology …………………………………………………….………………..02

1.5 Organization of Thesis…....……………………………………………………….03

1.6 Genesis of Problem……………………………………………………………......03

1. **LITERATURE SURVEY**

* 1. The KDD Process for Extracting Useful Knowledge from Volumes of Data…….04
  2. Data Mining with Big Data……………………………………………………….06
  3. Sentiment Analysis of Twitter Data……………………………………………....09
  4. Data Classification Using Support Vector Machine………………………………12

## 3. SYSTEM DEVELOPMENT

3.1 System Requirements ………..…………………………………………………..14

3.1.1 Software Requirements……………………………………….14

3.1.2 Minimum Hardware Requirements …………………………..14

3.2 Modelling …………………………………………………....……………..........14

3.3 Designing …………………………………………………………….………….15

## 4. ALGORTIHMS

4.1 Algorithms for Supervised Learning…………………………………………...18

4.1.1 Support Vector Machine………………………………………....…18

4.1.2 C4.5………………………………………………………...............18

4.1.3 Naïve Bayes………………………………………………...............18

4.1.4 Adaboost…………………………………………………................19

4.1.5 K-nearest neighbour………………………………………………..19

4.2 Algorithms for Unsupervised Learning………………………………………..19

4.2.1 K-means……………………………………………………………19

4.2.2 Apriori……………………………………………………………...20

4.2.3 Page Rank…………………………………………………………..20

4.3 Comparison Tables……………………………………………………..............21

## 5. IMPLEMENTATION

5.1 Tools and Techniques……………………………………………………...…...23

5.2 Implementation………………………………………………………………...24

5.3 Explanation…………………………………………………………………….25

5.4 Results………………………………………………………………………….26

## 6. CONCLUSION AND FUTURE PERSPECTIVE

6.1 Conclusion……………………………………………………………………..30

6.2 Future Scope…………………………………………………………………...31

**REFERENCES** ………………………………………………………………………….32

**List of Figures:**

1. Figure 1:
2. Figure 2:
3. Figure 3:
4. Figure 4:
5. Figure5:
6. Figure 6:
7. Figure 7:
8. Figure 8:
9. Figure 9:
10. Figure 10:

**List of Tables:**

1. Table 1:
2. Table 2:
3. Table 3:
4. Table 4:
5. Table 5:

# ABSTRACT

* Our project “**HOME IN: A Real Estate Agency**” is related to online application of estate.
* Real estate management system is advanced solution for every user around in need of house.
* Seller shows all the properties to Buyer that are secure & verify & send them to the Admin.
* Buyers buy the property according to their need of the house with the available facilities from the available houses.
* Registration of Buyer and Seller is free of cost. So anyone can register by using this system then Buy or Sell the property.
* The software is so reliable and optimal for every user .Our main concept is give best & quick result to user and make it better for every person in need of the house.

# CHAPTER 1: INTRODUCTION

## Introduction

## 1.2 Project Overview

In an era of social media and connectivity, web users are becoming increasingly enthusiastic about working together through online media. With a growing impact on everyday life, such as in education, health, commerce and tourism, leading to an exponential growth in the size of the social web. However, to find a house to own is a big task in this real world due to fake agents and wrong persons. We will hence focus on to connect the buyers and sellers under one roof to facilitate them with the approved houses and with ease to find the homes with the provided facilities according to the need of the buyers. To do this challenging task, we first need to build a cross functional platform and find the users who are in the need of the houses and then we have to find the sellers or real estate managers who will help the buyers to buy their houses.

## 1.3 Objective

## 1.4 Methodology

**Agile Methodology**

Agile is a software development approach where a self-sufficient and cross-functional team works on making continuous deliveries through iterations and evolves throughout the process by gathering feedback from the end users.Agile is one of the world's most broadly utilized and perceived programming advancement system. A large portion of the organizations have adopted in some structure or the other yet there is as yet far to go in the development of their reception programs. The sole point of this arrangement of instructional exercises is to installed innovation and non-innovation experts into the Agile World.

The Agile procedure imagines change and thinks about fundamentally more versatility than traditional techniques. Clients can take off little target enhancements without tremendous changes to as far as possible or schedule. The technique incorporates isolating each errand into sorted out requirements, and passing on each only inside an iterative cycle. An accentuation is the every day time table of developing little sections of an endeavor at some random minute. Each accentuation is assessed by the headway gathering and client. The encounters got from the examination are used to choose the resulting stage being created. Clients come to standard social affairs to review the work completed in the past accentuation, and to configuration work for the best in class cycle. Ordered targets are set in each accentuation meeting, for instance, anticipated changes, time evaluations, needs and spending plans.

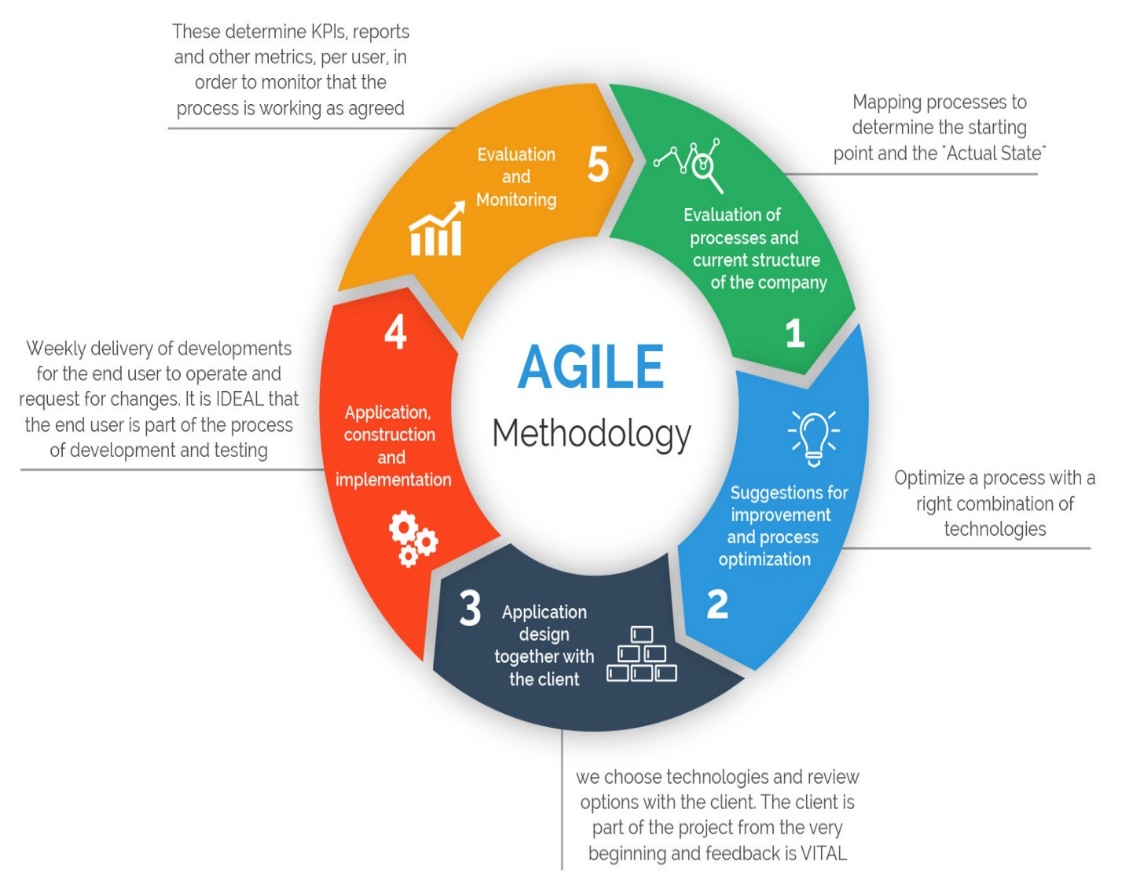


Figure 1. The Agile Methodology

There are many Agile procedures used across the world. The one used in this project is Scrum.

**Scrum**

Scrum is an agile advancement system which centers unequivocally around the most capable strategy to manage different tasks inside a group. It involves both Iterative and Incremental strategy for item improvement. The scrum in IT takes confidence in connected independently directed improvement bunches with three express and indisputably described occupations. These employments join – Product Owner (PO), Scrum Master (SM) and the headway bunch involving the software engineers and testers. They coordinate in iterative time boxed ranges called sprints. The initial step is the making of the product backlog by the PO. It's an arrangement for the day of stuff to be done by the scrum group. By then the scrum group picks the top required things and endeavors to finish them inside the time box called a sprint.

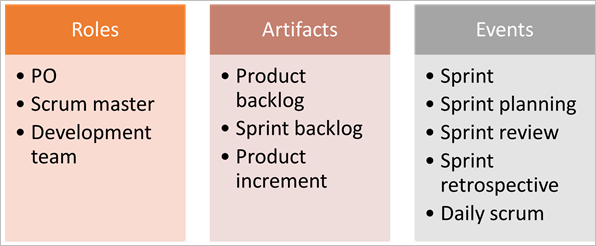


Figure 2. The Scrum Framework

**Scrum Terminologies:**

1. Scrum Team

It is a group incorporating 6 with + or – two people. These people acquire a mix of capacities and incorporate designers, database individuals, developers, testers, and so forth just as the scrum ace and PO. All of these people coordinate in close joint exertion for a recursive and obvious between time, to make and execute the said features. SCRUM bunch sitting arrangement expect a noteworthy activity in their collaboration, they never sit in work area territories or hotels, yet a major table. The group manages its own one of a kind work and deals with the work to complete the run.

1. Product Owner

He/she is the key accomplice or the lead customer of the application to be made. The PO is the person who addresses the customer side. He/she has the last say and is should be there for the group.

He/she should be reachable when anyone has any inquiries that need illustration. It is huge for the PO to understand and not to designate any new essential in the midst of the run or when the dash has started.

1. Scrum Master

He is the facilitator of the group. He/she guarantees that the scrum group is dynamic and productive. In case of any obstacles, scrum master settles them for the group meetings. Scrum Master is the middle person between the PO and the group.

He/she keeps the PO informed about the improvement of the Sprint. In case there are any problems or stresses for the group, looks at with the PO and gets them settled. Like the group's Daily meetings, a meeting of the SCRUM Master with the PO happens every day.

1. Sprint

It is a predefined between time or time distribution in which the work must be done and made arranged for review or arranged for product deployment. This time box customarily lies between around fourteen days to multi month.

In our regular day to day existence when we express that we seek after one month Sprint cycle, it basically suggests that we labor for one month on the endeavors and make it arranged for survey before that month's finished.

1. Product Backlog

It is where all the user stories are kept. This is kept up by the Product Owner. It tends to be imagined as a rundown of things wanted by the PO who sorts out it as per the business needs.

At the time of the group meeting, a user story is looked over the product backlog, by then the group does the conceptualizing, gets it and refines it and all things considered picks which user stories to take, with the assistance of the thing client.

1. User Story

These are the requirements of the user that have to be fulfilled. In the scrum, we don't have those colossal necessities archives, rather the prerequisites are characterized in a solitary passage, regularly having the arrangement as:

As a <User/Client/Admin>

I need to <Some reachable objective/target>

To accomplish <some result or explanation behind doing the thing>

For instance, if a client enters the wrong password multiple times, the administrator ought to have a password lock to prevent from the unauthorized access.

There are a couple of properties of user stories which should be followed. They should be short, viable, could be assessed, completed, debatable to be proven wrong and testable. A user story is never balanced or changed in the midst of the Sprint.

It is the duty of the SCRUM Master to guarantee that the PO has drafted the User Stories precisely with an authentic course of action of the Acceptance Criteria. If any change which will influence the arrival of sprint are to be made, by then such stories are pulled out of the run or they are finished by the available hours.

Every user story has an affirmation standard which should be particularly portrayed and appreciated by the group.

1. Sprint Backlog

In perspective on the need, user stories are taken from the Product Backlog one by one. The Scrum leader conceptualizes on it and chooses the attainability and settles on the accounts to take up at a particular sprint. The total rundown of the considerable number of stories which the whole group deals with in a sprint is called Sprint backlog.

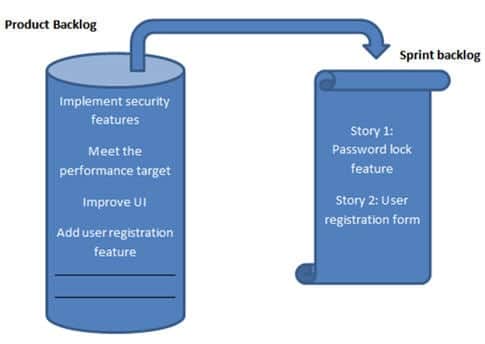


Figure 3. Product and Sprint Backlog

1. Burn Down Chart

It is a graph which evaluates the estimated effort v/s actual effort of the decided tasks. For a decided sprint, it tracks the daily divided work to check whether the project is going as decided and towards completion or not.

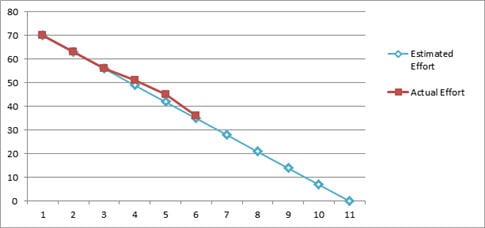


Figure 4. Burn Down Chart

**Flow of process:**

• Sprint is each step of a Scrum.

• List where all tasks are decided and composed to get a final result is called Product Backlog.

• Top things of Product Backlog are given to Sprint Backlog.

• Team completes and delivers product functionalities toward the end of the Sprint.

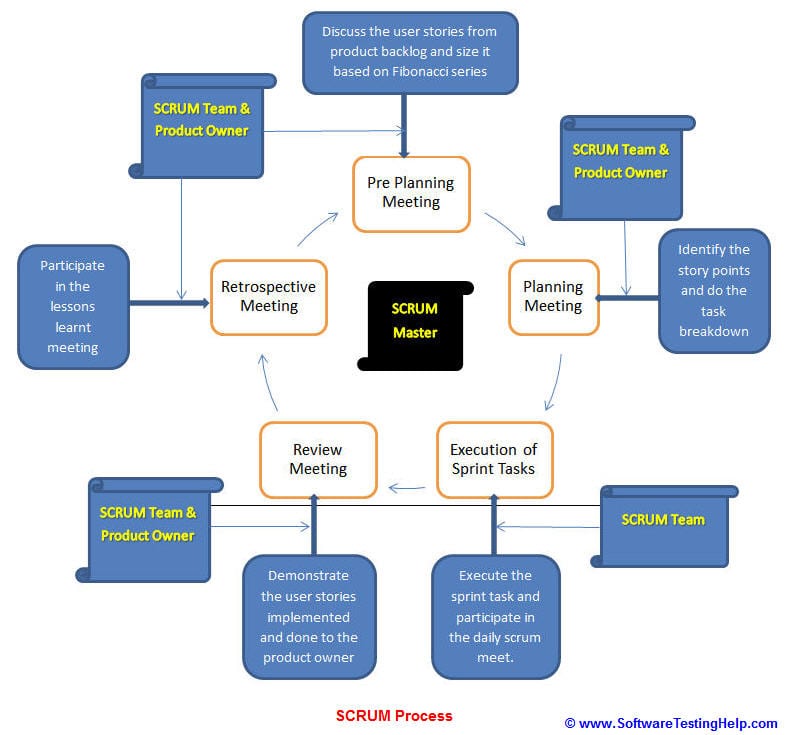


Figure 5. Scrum Process

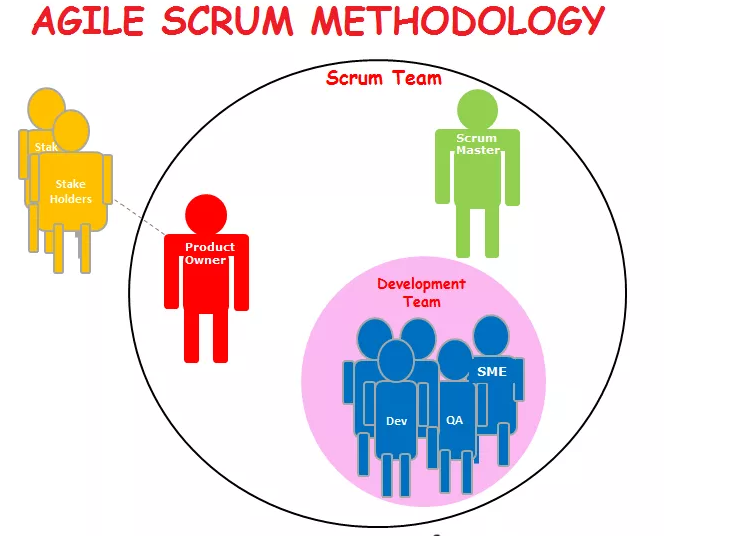


Figure 6. Agile Scrum Methodology

## 1.5 Organization of Thesis

In Chapter 2, we have discussed the literature survey. We have looked upon the various existing approaches that have been used by the real estate managers. Each one of them has its own unique features which distinguishes it from the rest.

In Chapter 3 we have developed the system design and explained various diagrams that have been used to explain our system.

In Chapter 4 we have discussed about various algorithms that are needed to make the work of buyer and seller easy and fast.

In Chapter 5 we have done feasibility study and requirement analysis.

In Chapter 6 we have thereby concluded our report.

## 1.6 Genesis of Problem

In this modern digital age there are many fake real estate mangers that are present at every corner of the city and are taking money from the poor people and then misguiding them in providing the houses.

# CHAPTER 2: LITERATURE SURVEY

**Paper 1:** *Usama Fayyad, Gregory Piatetsky Shapiro and Padhraic Smyth:*

*“*The KDD Process for Extracting Useful Knowledge from Volumes of Data” Communication of ACM, Volume 39, Nov. 2015, Page 29-34

**Overview:**

As we march into the age of of digital information, the problem of data overload looms ominously ahead. Our ability to analyse and understand massive datasets lags far behind our ability to gather and store the data. A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining.

**The KDD Process**

Here we present our (necessarily subjective) point of view of a bringing together process-driven structure for KDD. The objective is to give an outline of the variety of activities this multidisciplinary field and how they fit together. We define the KDD process as:

*The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*

## Data Mining

Data mining involves fitting models to or determining patterns from observed data. The fitted models play the role of inferred knowledge. Deciding whether or not the models reflect useful knowledge is a part of the overall interactive KDD process for which subjective human judgment is usually required. A wide variety and number of data mining algorithms are described in the literature—from the fields of statistics, pattern recognition, machine learning, and databases. Thus, an overview discussion can often consist of long lists of seemingly unrelated, and highly specific algorithms. Here we take a somewhat reductionist viewpoint. Most data mining algorithms can be viewed as compositions of a few basic techniques and principles.

**Conclusion:**

In spite of its fast development, the KDD field is still in its earliest stages. There are numerous difficulties to survive, yet a few triumphs have been accomplished (see the articles by Brachman on business applications and by Fayyad on science applications in this extraordinary segment). Since the potential settlements of KDD applications are high, there has been a race to offer items and administrations in the market. An extraordinary test confronting the field is the means by which to stay away from the sort of false desires tormenting other incipient (and related) advancements (e.g., manmade reasoning what's more, neural systems). It is the duty of analysts and experts in this field to guarantee that the potential commitments of KDD are not exaggerated what's more, that clients comprehend the genuine idea of the commitments alongside their confinements. Major issues at the core of the field stay unsolved. For instance, the essential issues of measurable surmising and disclosure stay as troublesome what's more, testing as they generally have been. Catching the specialty of examination and the capacity of the human mind to integrate new information from information is as yet incredible by any machine. Be that as it may, the volumes of information to be broke down make machines a need. This specialty for utilizing machines as a guide to examination and the expectation that the monstrous datasets contain chunks of profitable information drive intrigue and research in the field. Bringing together an arrangement of shifted fields, KDD makes ripe ground for the development of new devices for overseeing, examining, furthermore, in the long run picking up the high ground over the surge of information confronting present day society. The way that the field is driven by solid social and monetary needs is the stimulus to its proceeded with development. The rude awakening of genuine applications will go about as a channel to filter the great hypotheses what’s more, procedures from those less helpful.

**Paper 2:** *Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding:*

“Data Mining with Big Data” IEEE Transactions on Knowledge and Data

Engineering, Volume26 Issue 1, January 2014 Pages 97-107

**Overview:**

Enormous Data concern vast volume, mind boggling, developing informational collections with numerous, self-ruling sources. With the quick improvement of systems administration, information stockpiling, and the information gathering limit, Big Data are presently quickly growing in all science and building spaces, including physical, natural and biomedical sciences. This paper displays a HACE hypothesis that describes the highlights of the Big Data upheaval, and proposes a Big Data preparing model, from the information mining point of view. This information driven model includes request driven accumulation of data sources, mining and examination, client enthusiasm displaying, and security and protection contemplations. We investigate the testing issues in the information driven model and furthermore in the Big Data upheaval.

## Big data characteristics: Hace Theorem

Enormous Data concern vast volume, mind boggling, developing informational collections with numerous, self-ruling sources. With the quick improvement of systems administration, information stockpiling, and the information gathering limit, Big Data are presently quickly growing in all science and building spaces, including physical, natural and biomedical sciences. This paper displays a HACE hypothesis that describes the highlights of the Big Data upheaval, and proposes a Big Data preparing model, from the information mining point of view.

This information driven model includes request driven accumulation of data sources, mining and examination, client enthusiasm displaying, and security and protection contemplations. We investigate the testing issues in the information driven model and furthermore in the Big Data upheaval.

## Data Mining Challenges with Big Data

For an astute learning database framework to deal with Enormous Data, the fundamental key is proportional up to the astoundingly expansive volume of information and give medications to the qualities included by the previously mentioned HACE hypothesis. The difficulties at Tier I centre around information getting to and number-crunching registering strategies. Since Big Data are frequently put away at various areas and information volumes may constantly grow, a successful registering stage will need to take conveyed extensive scale information stockpiling into thought for figuring.

For instance, regular information mining calculations require all information to be stacked into the fundamental memory, this, notwithstanding, is turning into a reasonable specialized boundary for Big Data in light of the fact that moving information crosswise over various areas is costly (e.g., subject to escalated organize correspondence and other IO costs), regardless of whether we do have a super huge principle memory to hold all information for figuring. The difficulties at Tier II base on semantics and space learning for various Big Data applications. Such data can give extra advantages to the mining process, and in addition add specialized boundaries to the Big Data get to (Tier I) and mining calculations (Tier III).

For instance, contingent upon various space applications, the information security and data sharing instruments between information makers and information buyers can be fundamentally unique. Sharing sensor organize information for applications like water quality checking may not be debilitated, though discharging and sharing portable clients' area data is obviously not adequate for greater part, if not all, applications. In expansion to the above security issues, the application areas can likewise give extra data to profit or then again control Big Data mining calculation plans. For instance, in market bin exchanges information, every exchange is considered free and the found learning is commonly spoken to by finding profoundly connected things, potentially regarding distinctive transient or potentially spatial limitations.

In an informal community, then again, clients are connected and share reliance structures. The information is at that point spoken to by client networks, pioneers in each gathering, and social impact displaying, et cetera. In this way, understanding semantics and application learning is imperative for both low-level information get to and for abnormal state mining calculation structures.

At Tier III, the information mining difficulties focus on calculation plans in handling the challenges raised by the Enormous Data volumes, circulated information disseminations, and by mind boggling and dynamic information attributes. The hover at Level III contains three phases. To start with, scanty, heterogeneous, dubious, fragmented, and multisource information are pre-processed by information combination procedures. Second, complex and dynamic information are mined in the wake of pre-processing.

Third, the worldwide information gotten by nearby learning and model combination is tried and significant data is feedback to the pre-processing stage. At that point, the model and parameters are balanced by the criticism. In the entire procedure, data sharing isn't just a guarantee of smooth advancement of each stage, yet in addition a reason for Big Data handling.

## Conclusions

Driven by true applications and key modern partners and introduced by national financing offices, overseeing and mining Big Data have appeared to be a testing yet exceptionally convincing undertaking. While the term Big Information truly worries about information volumes, our HACE hypothesis recommends that the key qualities of the Big Data are:

1. immense with heterogeneous and assorted information sources
2. self-sufficient with conveyed and decentralized control, furthermore 3) complex and developing in information and learning affiliations.

Such joined qualities recommend that Big Information require a "major personality" to unite information for most extreme values. To investigate Big Data, we have examined a few difficulties at the information, model, and framework levels. To help Big Information mining, superior figuring stages are required, which force orderly structures to release the full intensity of the Big Data. At the information level, the independent data sources and the assortment of the information accumulation conditions, regularly result in information with confounded conditions, for example, missing/questionable qualities.

In different circumstances, protection concerns, commotion, and blunders can be brought into the information, to create changed information duplicates. Building up a sheltered and sound data sharing convention is a noteworthy test. At the model dimension, the key test is to create worldwide models by consolidating privately found examples to frame a bringing together view**.** This requires precisely structured calculations to break down model relationships between disseminated destinations, and breaker choices from various sources to pick up a best model out of the Big Data.

At the framework level, the fundamental test is that a Big Data mining system requirements to think about complex connections between tests, models, and information sources, alongside their advancing changes with time and other conceivable variables. A framework should be deliberately structured so unstructured information can be connected through their mind boggling connections to shape helpful examples, what's more, the development of information volumes and thing connections should enable frame to authentic examples to foresee the pattern also, future.

**Paper 3:** *Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca*

*Passonneau:* “Sentiment Analysis of Twitter Data”LSM '11 Proceedings of the Workshop on Languages in Social Media, June 23 - 23, 2011, Pages 30-38

**Data Description**

Twitter is an informal communication and microblogging administration that enables clients to post continuous messages, called tweets. Tweets are short messages, confined to 140 characters long. Because of the idea of this microblogging administration (speedy and short messages), individuals utilize acronyms, commit spelling errors, utilize emoji’s and different characters that express exceptional implications. Following is a concise phrasing related with tweets. Emoji’s: These are outward appearances pictorially spoken to utilizing accentuation and letters; they express the client's state of mind. Target: Users of Twitter utilize the "@" image to allude to different clients on the microblog. Alluding to different clients as such consequently alarms them. Hashtags: Users typically utilize hashtags to stamp themes. This is basically done to build the deceivability of their tweets.

## Resources and Pre-processing of data

In this paper we introduce two new resources for pre-processing twitter data:

1. an emoticon dictionary
2. an acronym dictionary.

1. We set up the emoji lexicon by naming 170 emoji’s recorded on Wikipedia with their passionate state. For instance, ":)" is marked as positive though “: = (" is named as negative. We appoint every emoji a name from the accompanying arrangement of marks: Extremely-positive, Extremely-negative, Positive, Negative, and Neutral. We aggregate an acronym word reference from an online asset.

1. The word reference has interpretations for 5,184 acronyms. For instance, lol is meant roaring with laughter. We pre-process every one of the tweets as pursues: a) supplant every one of the emoji’s with their estimation extremity by looking into the emoji lexicon, b) supplant all URLs with a label juju, c) supplant targets (e.g. "@John") with label jjTjj, d) supplant all refutations (e.g. not, no, never, n't, can't) by tag "NOT", and e) supplant an arrangement of rehashed characters by three characters, for instance, convert coooooooool to coool. We don't supplant the grouping by just two characters since we need to separate between the customary use and underscored utilization of the word.

|  |  |
| --- | --- |
| **Acronym** | **English Expansion** |
| gr8, gr8t | Great |
| Lol | laughing out loud |
| Rotf | rolling on the floor |
| Bff | best friend forever |

Table 1: Example acronym and their expansion in the acronym dictionary.

## Prior polarity scoring

Some of our highlights depend on earlier extremity of words. For getting the earlier extremity of words, we take inspiration from work by Agarwal et al. (2009). We utilize Dictionary of Affect in Language (DAL) (Whissel, 1989) and broaden it utilizing WordNet. This lexicon of around 8000 English dialect words doles out each word an enjoyableness score (2 R) between 1 (Negative) - 3 (Positive). We initially standardize the scores by plunging each score the scale (which is equivalent to 3).

We consider words with extremity under 0.5 as negative, higher than 0.8 as positive and the rest as unbiased. In the event that a word isn't straightforwardly found in the lexicon, we recover all equivalent words from Wordnet. We at that point search for every one of the equivalent words in DAL. On the off chance that any equivalent word is found in DAL, we dole out the first word a similar agreeableness score as its equivalent word. In the event that none of the equivalent words is available in DAL, the word isn't related with any earlier extremity. For the given information we straightforwardly found earlier extremity of 81.1% of the words. We discover extremity of other 7.8% of the words by utilizing WordNet. So we find earlier extremity of about 88.9% of English dialect words.

## Design of Tree Kernel

We plan a tree portrayal of tweets to consolidate numerous classes of highlights in one brief helpful portrayal. For ascertaining the similitude between two trees we utilize a Partial Tree (PT) bit previously proposed by Moschitti (2006). A PT bit computes the comparability between two trees by looking at all conceivable sub-trees. This tree piece is a case of a general class of convolution bits. Convolution Kernels, first presented by Haussler (1999), can be utilized to analyse unique items, similar to strings, rather than highlight vectors. This is on the grounds that these bits include a recursive estimation over the "parts" of unique protest. This estimation is made computationally effective by utilizing Dynamic Programming systems. By considering every conceivable mix of sections, tree bits catch any conceivable connection amongst highlights and classifications of highlights. Figure 2.1 demonstrates a case of the tree structure we plan. This tree is for an orchestrated tweet: "@Fernando this is anything but an extraordinary day for playing the HARP! :)". We utilize the accompanying technique to change over a tweet into a tree portrayal: Initialize the fundamental tree to be "ROOT". At that point tokenize each tweet and for every token: an) if the token is an objective, emoji, outcry stamp, other accentuation check, or a nullification word, add a leaf hub to the "ROOT" with the comparing tag.June

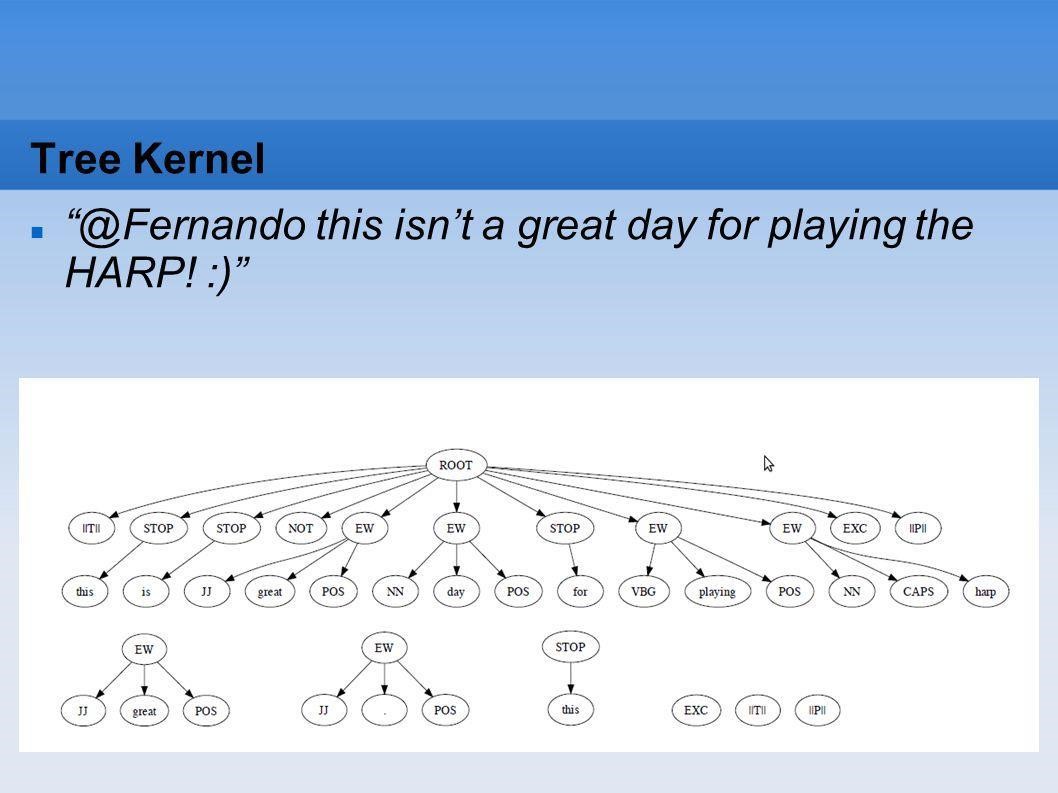


Fig 1: Tree kernel

## Results



|  |  |
| --- | --- |
| **Model** | **Accuracy percentage** |
| Unigram | 56.58 |
| Senti-features | 56.31 |
| Kernel | 60.60 |
| Unigram + Senti features | 60.50 |
| Kernel + Senti features | 60.83 |

Table 2 : Average and standard deviation for test accuracy for the 3-way classification task using different models

**Paper 4:** *Durgesh K. Srivastava, Lekha Bhambhu* “Data Classification Using

Support Vector Machine” Journal of Theoretical and Applied Information

Technology, June 2012, Pages 1-7

Grouping is a standout amongst the most imperative assignments for various application, for example, content arrangement, tone acknowledgment, picture order, smaller scale exhibit quality articulation, proteins structure expectations, information Classification and so on. The vast majority of the current regulated order techniques depend on customary insights, which can give perfect outcomes when test measure is keeping an eye on limitlessness. Be that as it may, just limited examples can be gained by and by. In this paper, a novel learning strategy, Support Vector Machine (SVM), is connected on various information (Diabetes information, Heart Data, Satellite Data and Shuttle information) which have two or multi class.

SVM, an amazing machine strategy created from factual learning and has made critical accomplishment in some field. Presented in the mid 90's, they prompted a blast of enthusiasm for machine learning. The establishments of SVM have been produced by Vapnik and are picking up fame in field of machine learning because of numerous alluring highlights and promising observational execution. SVM strategy does not endure the constraints of information dimensionality and restricted examples.

In our investigation, the help vectors, which are basic for grouping, are gotten by gaining from the preparation tests. In this paper we have demonstrated the relative outcomes utilizing diverse piece capacities for all information tests.

## Overview of SVM

SVMs are set of related directed learning strategies utilized for order and relapse. They have a place with a group of summed up direct characterization. An extraordinary property of SVM is, SVM at the same time limit the observational order mistake and amplify the geometric edge. So SVM called Maximum Margin Classifiers. SVM depends on the Structural hazard. Minimization (SRM). SVM outline vector to a higher dimensional space where a maximal isolating hyperplane is built. Two parallel hyperplanes are developed on each side of the hyperplane that different the information. The isolating hyperplane is the hyperplane that amplify the separation between the two parallel hyperplanes. A supposition is made that the bigger the edge or separation between these parallel hyperplanes the better the speculation mistake of the classifier will be.

## Model Selection of SVM

Display choice is additionally a critical issue in SVM. As of late, SVM have indicated great

execution in information characterization. Its prosperity relies upon the tuning of a few parameters which influence the speculation blunder. We frequently call this parameter tuning method as the model determination. In the event that you utilize the straight SVM, you just need to tune the cost parameter. Shockingly, direct SVM are frequently connected to straightly detachable issues. Numerous issues are non-directly distinct. For instance, Satellite information and Shuttle information are not sprightly distinct. In this way, we regularly apply nonlinear part to take care of grouping issues, so we have to choose the cost parameter and portion parameters. We as a rule utilize the lattice look technique in cross approval to choose the best parameter set. At that point apply this parameter set to the preparation dataset and afterward get the classifier. From that point forward, utilize the classifier to arrange the testing dataset to get the speculation precision.

## Results of Experiments

The order tests are directed on various information like Heart information, Diabetes information, Satellite information and Shuttle information. RSES Tool set is utilized for information characterization with all informational index utilizing diverse classifier system as Rule Based classifier, Rule Based classifier with Discretization, K-NN classifier and LTF (Local Transfer Function) Classifier. The equipment stage utilized in the analyses is a workstation with Pentium-IV-1GHz CPU, 256MB RAM, and the Windows XP(using MS-DOS Prompt). The accompanying three tables speak to the diverse analyses results. Table 2 demonstrates the Total execution time for all information to anticipate the exactness in a flash.

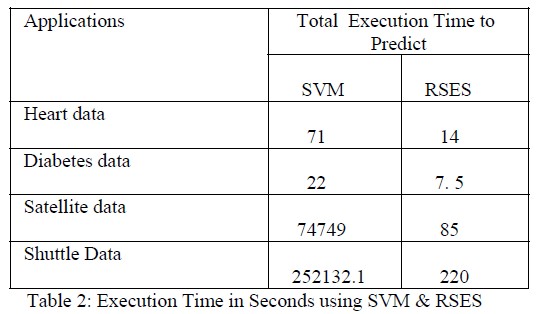


Table 3: Execution Time in Seconds using SVM & RES

# CHAPTER 3: SYSTEM DEVELOPMENT

**3.1 System Requirements:**

**3.1 1 Software Requirements:**

1. Windows 10 (64-bit) or Windows 7 (x64 or x86) or Windows Vista (x64 or x86) or Linux (tested on Ubuntu Linux**)**

1. [**Jupyter Notebook:**](http://jupyter.org/)

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.Jupyter Notebooks are a powerful way to write and iterate on your Python code for data analysis.

**Why Jupyter NoteBook?**

They turn into latex reports really easy. They also turn into slideshows really easy. They also let you run blocks of code really easily. **Some other alternatives:**  PyCharm, RStudio

**3.1.2 Minimum Hardware Requirement:**

* 2GB RAM
* Pentium Dual Core 4
* Memory Required 20 MB

## 3.2 Modelling

We are using incremental model in our project.

**What is Incremental Model?**

Incremental Model is a procedure of software development where requirements are broken into different independent modules of software development cycle.

Incremental development is done in ventures from analysis design, implementation, testing/verification, maintenance.

Each iteration will go through the requirements, design, coding and testing phases. Each subsequent release of the system adds function to the last release until all designed functionality has been implemented.

The system is put into production when the first increment is delivered. The first increment is often a core product where the basic requirements are addressed, and supplementary features are added in the following increments. Once the core product is analysed by the client, there is other plan development for the next increment.

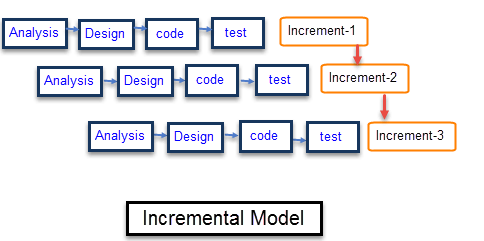


Fig 2 - Pictorial Representation of Incremental Model

## 3.3 Designing

Volume of information is increasing every day that we can handle from business transactions, scientific data, pictures, videos and many others. So, we need a system that will be capable of extracting the information available and that can automatically generate reports, views or summary of data for better decision-making.

There are three phases in designing a model:

1. **Training Data**: The actual dataset that we use to train the model. The model sees and learns from this data. The sample of data used to for the model.

1. **Validation Dataset**: The hyper-parameters of a classifier. It is sometimes also called the development set.

1. **Test Data**: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.



**Fig 3: 3 phases of training data**

To train data, we have to follow a process. This process is generally referred to as KDD Process.

Knowledge discovery in databases (**KDD**) is the process of discovering useful knowledge from a collection of data.

Steps followed in KDD Process are as follows:

1. **Data Cleaning** − In this step, the noise and inconsistent data is removed.
2. **Data Integration** − In this step, various data sources are combined.
3. **Data Selection** − In this step, data relevant for the analysis are extracted from the database.
4. **Data Transformation** − In this step, data is transformed into appropriate forms for mining by performing aggregation operations.
5. **Data Mining** − In this step, intelligent methods are applied in order to extract data patterns.
6. **Pattern Evaluation** − In this step, data patterns are evaluated.
7. **Knowledge Presentation** − In this step, knowledge is represented.

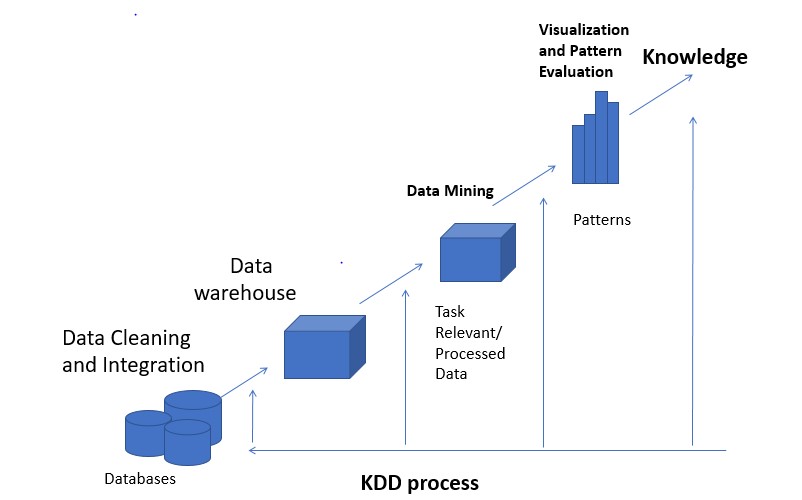


Fig 4: KDD Process

# CHAPTER 4: ALGORITHMS

**4.1 Algorithms for Supervised Learning:**

1. **Support Vector Machines (SVM):**

SVM takes in a hyperplane (work) to group information. This capacity finds augments separate between 2 classes. It is like C4.5, yet without decision tree. Analyses data for classification and regression analysis.

**Tools**-MATLAB

1. **C4.5:**

Builds a classifier as a decision tree Classifier is a tool that attempts to predict which class the new data belongs to. C4.5 utilizes a solitary pass pruning procedure to moderate over-fitting. Pruning results in numerous upgrades.

C4.5 can work with both constant and discrete information. The comprehension is it does this by indicating reaches or edges for ceaseless information in this way transforming nonstop information into discrete information.

**Example**: We have a patient of whom we need to predict if he/she has cancer or not. We can do so but classifying some of the aspects like tumour size, heredity, symptoms etc. C4.5 will help to judge which of these parameters can help us to predict.

**Tools**: Weka

1. **Naïve Bayes:**

Not a single algorithm, but a family of classification algorithms, that share one common assumption. Every feature of data being classified is independent (means when the value of one feature has no effect on the value of another feature) of all other features. In nutshell, it allows us to predict the class given a set of features using probability.

**Example**: Suppose we have a collection of 100 fruits. Now these fruits can exist in one or more classes, say, Banana, Orange, etc. and each fruit has some features associated with it like long, sweet, yellow etc.

1. **Adaboost:**

Constructs a classifier. It is a boosting algorithm, which means it takes multiple learning algorithm and combines them to create single strong leaner.

**Example**- A simple example might be classifying a person as male or female based on their height. You could say anyone over 5’ 9” is a male and anyone under that is a female. You’ll misclassify a lot of people that way, but your accuracy will still be greater than 50%.

**Tools**- scikit-learn, ICSI Boost

1. **k-nearest neighbour (knn):**

A classification algorithm. It is different from others as it is a lazy learner (which means it only stores data while training). An eager learner builds a classification model during training. Now the question arises how knn figures out what is closer?

* + - Euclidean Distance
    - Hamming Distance

**Example**- It helps banks in giving credit ratings to customers and predict if bank should give loan to that person or not. Another example is that is used on Image/Video Recognition.

**Tools**: MATLAB, scikit-learn

**4.1 Algorithms for Unsupervised Learning:**

1. **k-means:**

k-means creates *k* groups from a set of objects so that the members of a group are more similar. It’s a popular cluster analysis technique for exploring a dataset. Cluster analysis is a family of algorithms designed to form groups such that the group members are more similar versus non-group members. Clusters and groups are synonymous in the world of cluster analysis.

**Example:**  Suppose we have a dataset of patients. In cluster analysis, these would be called observations. We know various things about each patient like age, pulse, blood pressure, cholesterol, etc. You tell k-means how many clusters you want. K-means takes care of the rest.

**Tools:** Weka, MATLAB

1. **Apriori:**

[Apriori](https://www.techleer.com/articles/155-apriori-algorithm-classical-algorithm-for-data-mining/) is an algorithm that is used for frequent item-set mining and association rule learning overall transactional databases. The algorithm is proceeded by the identification of the individual items that are frequent in the database and then extending them to larger item-sets as long as sufficiently those item sets appear often enough in the database. These frequent item-sets that are determined by Apriori can be used for the determination of association rules which then highlight general trends.

**Example:** There was a survey done by Walmart in which it was deduced that a man who buys diaper is most likely to buy a beer too. This is a very unexpected yet popular prediction and thus after this analysis supermarkets used to keep beer and diaper next to each other. Similarly, another such example is bread-butter.

**Tools:** Weka, AR Tool

1. **Page Rank:**

[PageRank (PR)](https://en.wikipedia.org/wiki/PageRank) that was named after Larry Page who is one of the founders of Google is an algorithm that is used by Google Search to rank the websites in their search engine results. PageRank, that is the first algorithm that was used by the company is not the only algorithm that is being used by Google to order search engine results, but it is the best-known way of measuring the importance of website pages.

**Example:** The most prevalent example of PageRank is Google’s search engine. Although their search engine doesn’t solely rely on PageRank, it’s one of the measures Google uses to determine a web page’s importance. PageRank is really just a super effective way to do link analysis.

**Tools:** Python, C++

-

**Comparison Tables:**

**Table 4: Comparison Table for Supervised Learning Algorithms**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithms** | **Description** | **Example** | **Tool** |
| C4.5 | Builds a classifier in form of a decision tree. | Patient->  Cancer->  Tumour Size | Weka |
| SVM | Hyperplane is used to classify the data.  Regression is used to predict the  future outcomes. | Bunch of red and blue balls.  Use Hyperplane if required. | MATLAB |
| Naïve Bayes | Family of class algorithms.  Used to predict the classes. | 1000 Fruits **Classes:** Banana, Apple  Features: Long,  Sweet,  Yellow | Weka |
| Adaboost | Boosting  Algorithm  Uses  Classification  Strong  Learner | Face Detection | Scikitlearn |
| K nearest neighbour | Uses  Classification Lazy Learner Determines  how close data is related. | Credit  Rating  Video/Image  Recognition | MATLAB,  Scikit  Learn |

**Table 5: Comparison Table for Unsupervised Learning Algorithms**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithms** | **Description** | **Example** | **Tools** |
| K-Means | It creates clusters of elements with similar characteristics. | Patients  going to  respective doctors in a hospital. | Weka,  MATLAB |
| Apriori | An association algorithm in which 3 parameters are kept in mind-  Item-size,  Minimum Support Count,  Probability. | BeerDiaper problem, BreadButter example. | Weka, AR Tool |
| PageRank | Link analysis algorithm. In this algorithm rank is calculated on the basis on links. | Google, Bing search engines. | Python, C++ |

# CHAPTER 5: IMPLEMENTATIONS AND RESULTS

## 5.1 Tools and Techniques

We have Python Language for implementing our project.

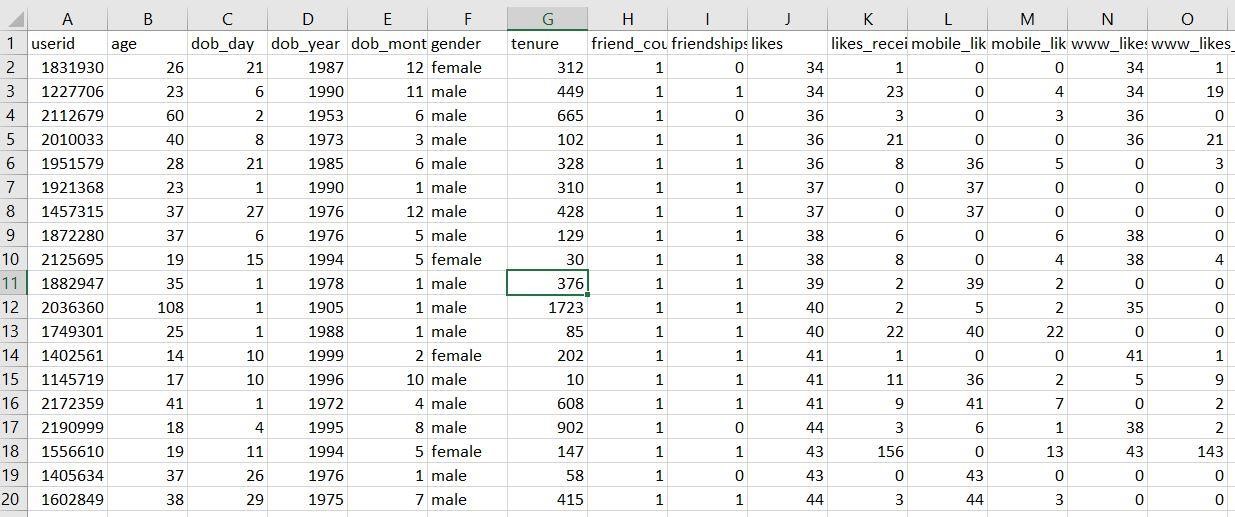
**Why Python?**

Python does contain special libraries for machine learning namely scipy and numpy which great for linear algebra and getting to know kernel methods of machine learning. The language is great to use when working with machine learning algorithms and has easy syntax relatively.

**Dataset Used:**

Facebook Structured data with records of 100000 users with columns like userid, age, dob\_day, dob\_year, dob\_month, gender, tenure, friend\_count, friendships\_initiated and likes.

Fig 5 : Facebook Dataset Snapshot from www.kaggle.com(pseudo\_facebook)



**5.2 Implementation:**

**Implemented Algorithm:**

**C4.5**

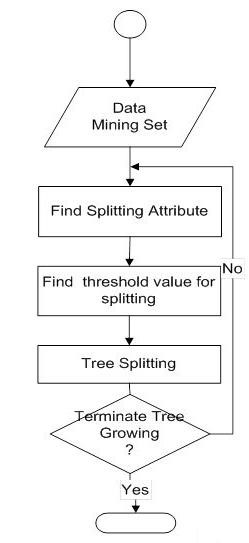
**What does it do?**

C4.5 constructs a classifier in the form of a decision tree. In order to do this, C4.5 is given a set of data representing things that are already classified. This is supervised learning, since the training dataset is labelled with classes.

**Why C4.5**

* Firstly, C4.5 uses [information gain](http://en.wikipedia.org/wiki/Entropy_(information_theory)) when generating the decision tree instead of using gini index or gain index. These two indices, though generate a decision tree but the decision tree is not as precise.
* Secondly, ease of interpretation and explanation. It is quite fast, popular and the output is [human readable.](https://www.google.com/search?q=j4.8+output&tbm=isch)
* Thirdly, C4.5 can work with both continuous and discrete data.

**Fig 6 -Flow Chart of C4.5**



**Explanation:**

C4.5 algorithm used to generate a decision tree that is developed by Ross Quinlan. C4.5 algorithm is an extension of ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. Weka machine learning software authors described C4.5 algorithms as "a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date".

The general algorithm for building decision trees is :

1. Check for the above base cases that are necessary for building a decision tree.
2. We will find the normalized information gain from splitting on each attribute of *b*.
3. Consider *b\_best* be the attribute that has the highest normalized information gain.
4. Now we create a decision *node* that splits on *b\_best*.
5. Recur on the sublists that has been obtained by splitting on *b\_best*, and add those nodes as children of that *node*.

**Libraries Used:**

We have used following libraries in the implementation code:

Pandas: It offers operations for manipulating data.

matplotlib.pyplot : Provides pie-charts for plotting data. seaborn: Provides bar-graphs for plotting data.

Numpy: For adding support for multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions for the operations of arrays.

**5.3 Results:**

## *(i)*

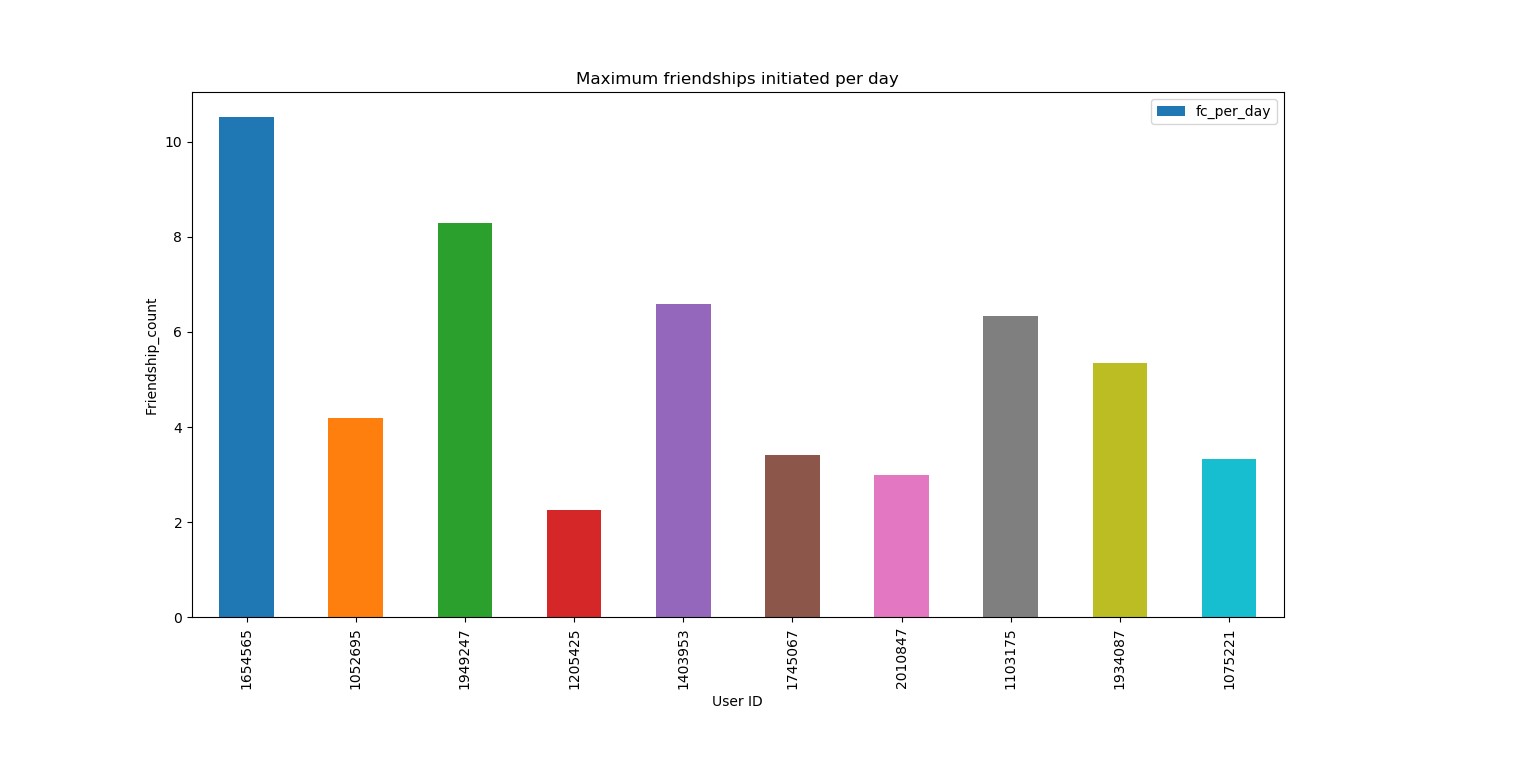


Fig 7- Maximum friendships initiated per day

This graph concludes that

1. We can find those people who are most interested in sending friend requests by calculating the average friendships initiated per day.

1. Hence, we can target those people and if a company launches a product or service, they would know that they have to target those users who have more friend count. By doing this, the knowledge of the products/services will be expanded to more users.

Organisations can hereby invest and plan accordingly when it comes to promotion of any service/product. Also this is not the only criteria that has to be kept in mind. There are multiple criterions which have to be simultaneously kept in mind.

## *(ii)*

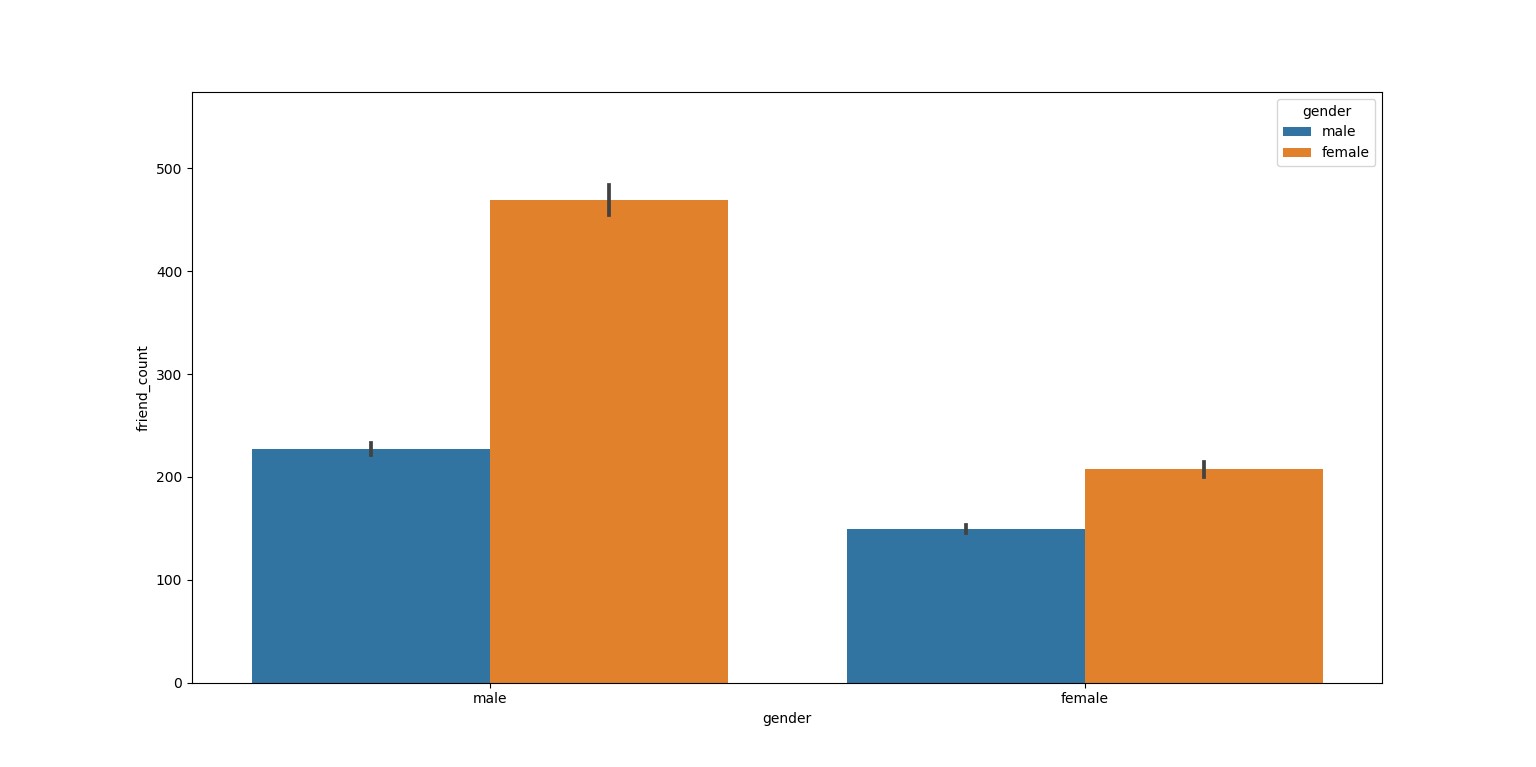


Fig 8- Friendship count according to gender

This bar graph helps us to get mainly two conclusions:

1. In case of Males, they have more number of female friends than male friends.

1. In case of Females, they have more number of females friends than male friends.

Hence, for business analytics, it can be inferred that females can play a major role in expanding business and for office-related work. Males can rather be deployed at on-site projects.

## *(iii)*

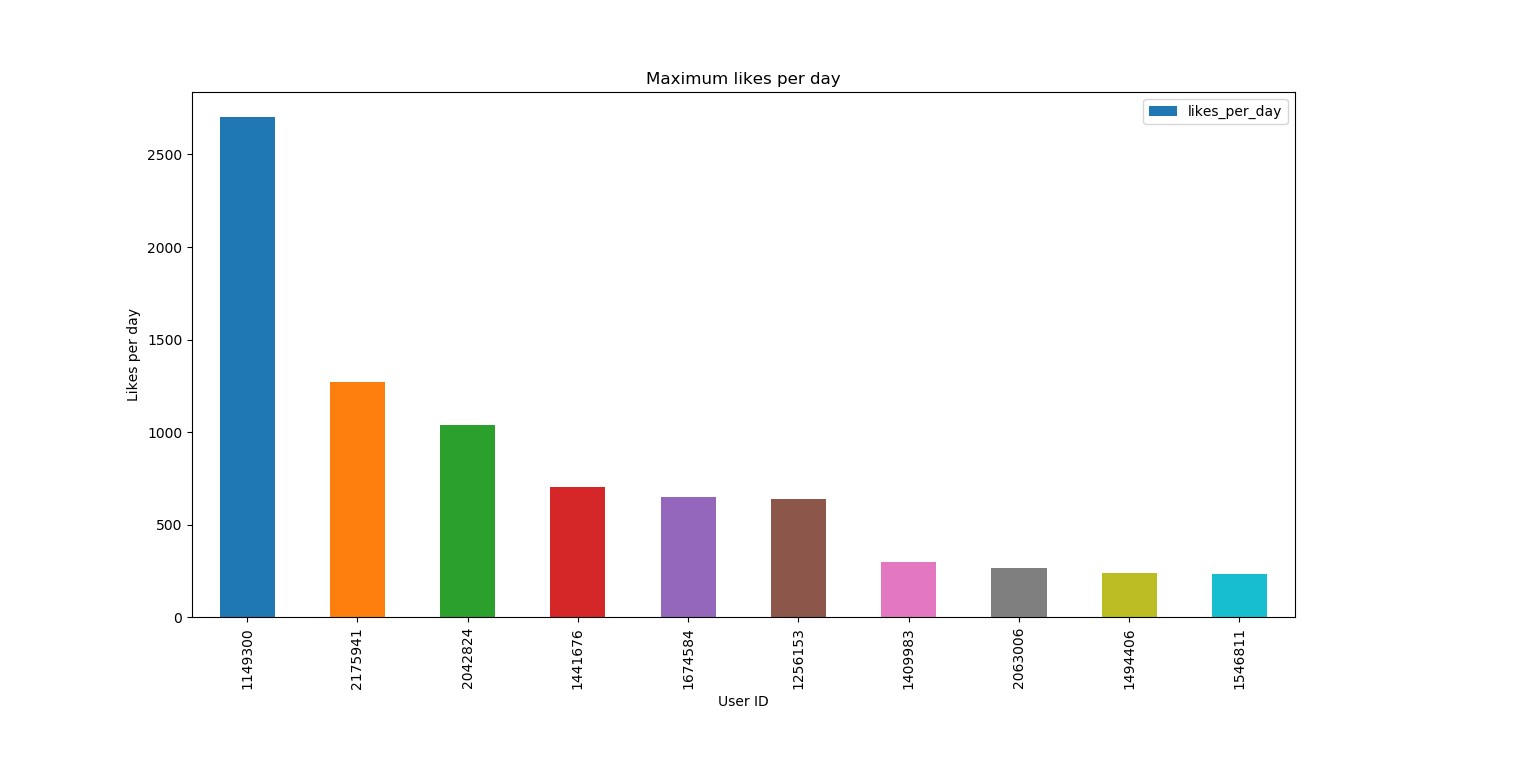


Fig 9- : Maximum likes per day

.

The graph concludes the following points:

1. We can target those users who like frequently and identify the genres of media they like.

1. Using this we can calculate what users are looking for and hence an organization can launch a product/service based on the demand of that entity.

## *(iv)*

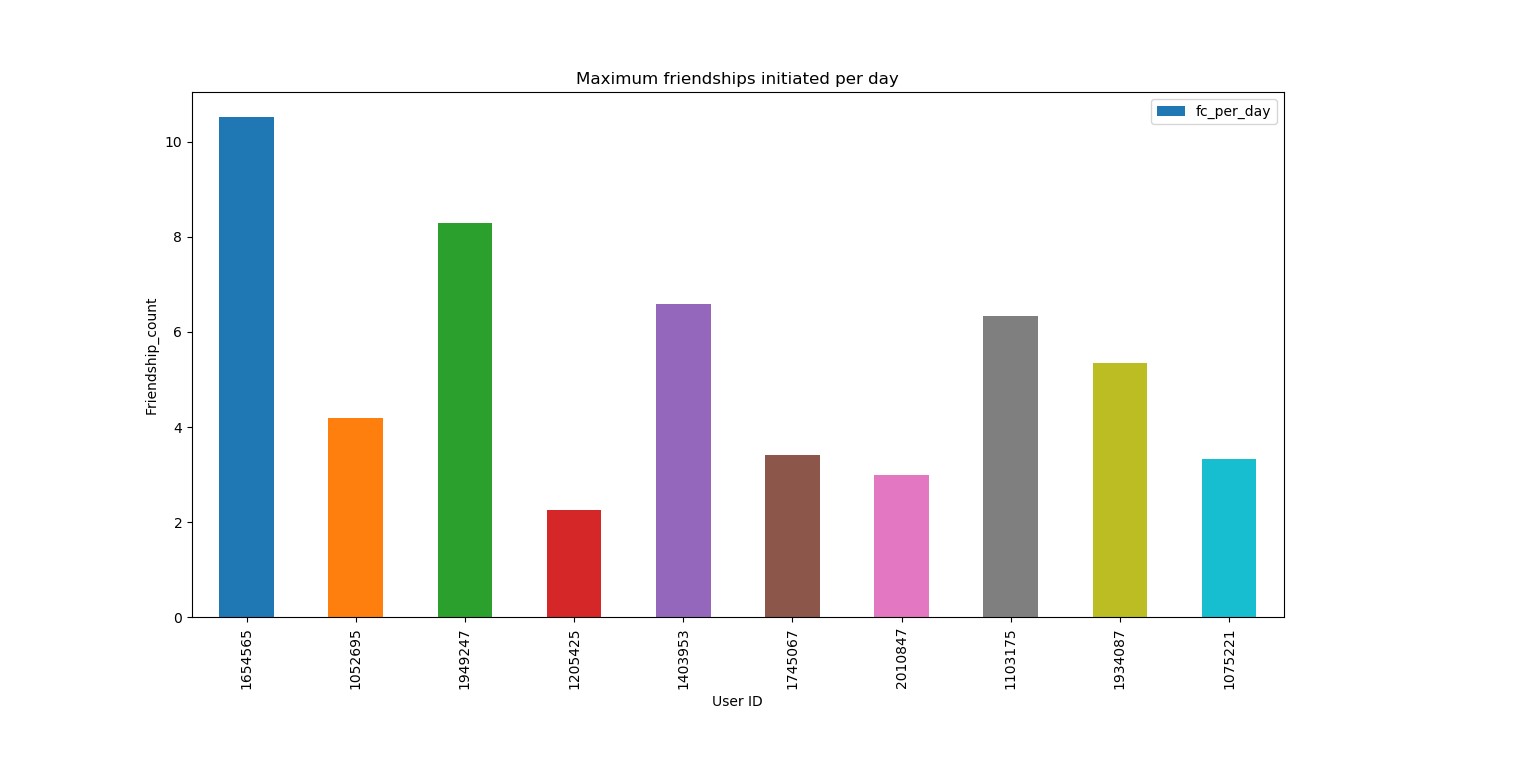


Fig 10- Maximum friendships initiated in total

This graph concludes that

1. We can find those people who are most interested in sending friend requests since their account has been active.

1. Hence, we can target those people and if a company launches a product or service, they would know that they have to target those users who have more friend count. By doing this, the knowledge of the products/services will be expanded to more users.

We can hence identify the users with maximum number of friends (and hence those with maximum network chain), meaning that by targeting a lesser number of users, we can promote efficiently.

# CHAPTER 6: CONCLUSION AND FUTURE SCOPE

## 6.1 Conclusion

As society advances at a similar pace with the Web, online social data are winding up progressively critical for the two people and organizations. The Web, in any case, has uncovered a time of online interest that is causing client produced substance to develop exponentially and, subsequently to end up ceaselessly bigger and more mind boggling. In order to reduce such complexity, we need to adopt different approach to big social data analysis, which combines different perspectives by means of social media analytics, trend discovery, multimedia management, social network analysis, opinion mining, and more. In particular, with the end goal to make an interpretation of negligible gathered insight into genuine aggregate knowledge, we have to go for an idea level examination of online social information that can empower the total of the semantics related with content.

We have learnt how machine learning algorithm help us to fabricate big-social data. Though there are many algorithms such as C4.5, Adaboost, KNN, yet we have specifically used C4.5 in this project as it provides a flexible environment, improved performance than others with greater efficiency and less complexity. As our project revolves around classification, hence C4.5 is the best choice to implement it as C4.5 creates a decision tree which performs seamless classification.

**What we have done till now?**

We have gone through literature and information from research papers. We have reviewed most popular algorithms which could have assisted us in this project and selected only one of them based on the flexibility or resources available. We have used Python for basic implementation of this project and using built in libraries we have come up with some useful results. We have done analysis of the data using Python and portrayed outputs in pictorial graphs for easy judgement and analysis.

**6.2 Future Scope:**

Social Media applications allow users to share comments, opinions, ideas, and media with friends, family, businesses, and organizations. The data contained in these comments, ideas, and media are valuable to many types of organizations.

This allows professionals to apply this data to solve different problems, in different organizations. We plan to make more classifiers using the one which we have made, that would help us to predict user trends or behaviour and could help us to target types of users.

We will further recommend friends, pages, communities by keeping a track of unique user reactions to different scenarios and genres users like to spends time on.

In the future this system can be implemented to automate most of the system and it can be designed for cross platform use. This would greatly improve the level of penetration and transparency that the several organisations currently have.

We could offer a good level of safety structure for analysts such that the data extracted cannot be breached which could result in privacy issues for many us

# References

1. *Usama Fayyad, Gregory Piatetsky Shapiro and Padhraic Smyth:*

*“*The KDD Process for Extracting Useful Knowledge from Volumes of Data” Communication of ACM, Volume 39, November 2015, Page 29-34.

1. *Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding:*

“Data Mining with Big Data” IEEE Transactions on Knowledge and Data Engineering, Volume26 Issue 1, January 2014 Pages 97-107.

1. *Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau:*

“Sentiment Analysis of Twitter Data”LSM '11 Proceedings of the Workshop on Languages in Social Media, June 23, 2011, Pages 30-38.

1. *Durgesh K. Srivastava, Lekha Bhambhu* “Data Classification Using Support Vector Machine” Journal of Theoretical and Applied Information Technology, June 2012, Pages 1-7.

1. *Vigneswara Ilavarasan, Arpan Kumar Kar* “Social media and business practices in emerging markets: still unexplored” Journal of Advances in Management Research , May 2018, Pages 110-114.

1. *Arpan Kumar Kar, “* *Applications of Analytics in Social Media*”, Intelligent Systems, IEEE,· January 2015.

32