Intelligent Most Popular Location Prediction in Cloud Environment through Facebook Check-ins using Multi-Model Ensembling Approach

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Shobhana Kashyap

(Roll No. 801532048)

Under the supervision of:

Dr. Maninder Kaur

Assistant Professor

CSED, Thapar University, Patiala

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

June 2017

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Intelligent Most Popular Location Prediction in Cloud Environment through Facebook Check-ins Using Multi-Model Ensembling Approach*", in partial fulfillment of the requirements for the award of degree of Master of Technology in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Maninder Kaur* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Shobhana Kashyap

(Shobhana Kashyap)

801532048

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

Maninder Kaur

(Dr. Maninder Kaur)

Assistant Professor

Computer Science and Engineering Department

ACKNOWLEDGEMENT

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life.

This work would not have been possible without the encouragement and able guidance of my supervisor **Dr. Maninder Kaur**. I thank my supervisor for their time, patience, discussions and valuable comments. Their enthusiasm and optimism made this experience both rewarding and enjoyable.

I am equally grateful to **Dr. Maninder Singh**, Associate Professor and Head, Computer Science & Engineering Department, a nice person, an excellent teacher and a well – credited researcher, who always encouraged me to keep going with work and always advised me with his invaluable suggestions.

I am thankful to **Dr. Ashutosh Mishra**, P.G. Coordinator, and Associate Professor of Computer Science & Engineering Department of Thapar University for the motivation and inspiration for the completion of thesis.

I will be failing in my duty if I don't express my gratitude to **Dr. S.S. Bhatia,** Senior Professor and Dean of Academic Affairs, Thapar University, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection, which made my stay at Thapar University memorable.

Last but not least, I would like to thank my family whom I dearly miss and without whose blessings none of this would have been possible. To my parents, I own thanks for their wonderful love and encouragement. I would also like to thank my brother, since he insisted that I should do so. I would also like to thank my close friends for their constant support.

Shobhana Kashyap

(Shobhana Kashyap)

(801532048)

Date: June, 2017

Place: Thapar University, Patiala

ABSTRACT

With the advent of check-in functionality in Facebook, people are able to share more information with the world. Almost every person is using social networking sites nowadays, but the amounts of information they share are appreciated by few.

In this research, a new model has been designed for identification of Facebook check-ins dataset for predicting most popular places for the user that he/she would like to check-in. Two different machine learning environments, Apache Mahout and R Tool, have been used for predicting most popular places. Each platform has its different classification algorithms. These two machine learning platforms through Ensembling technique have been compared and their analysis has been listed out. In both environments, unique multilevel ensemble model is generated for prediction of Facebook more popular places.

In the first module, Facebook check-ins dataset has been used on R tool on a standalone machine, machine learning algorithms have been executed on the given dataset to foresee accuracy for the most famous area. Support Vector Machines model has been chosen as a powerful model since it gives the most astounding accuracy of 77.03% after Conditional Inference Tree model and k-Nearest Neighbors Machine model. Further, these 3 models are ensembled leading to 82.12% accuracy. After that k-fold method is applied, this gives the highest accuracy of 88.18%. In the second module, the Mahout Classification machine learning algorithm has been implemented. For Ensembling technique, the top three models have been chosen; afterward these three models are ensembled to get the highest accuracy. The ensemble model of Facebook check-ins accomplishes 91.66% of accuracy.

The experimental outcomes have likewise been assessed utilizing 9768 instances that distinctly support the most extreme accuracy through Ensembling and utilize less execution time in machine learning environment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SNAPSHOTS

CHAPTER-1

INTRODUCTION

Today's reviews of Convivial Media, Facebook has set up itself as the chief informal communication site. In a statement by a technology review website "social media today", Facebook is the most utilized interpersonal interaction site these days. A number of issues that are originated from social networking websites have been denounced by critics in general, Facebook present a ton of benefits for Business users. A social networking service is an online platform that is used by people to build social networks or social relations with other people who share similar personal or career interests, activities, backgrounds or real-life connections.

Web-based social networking is the group of online interchanges channels committed to group based info, association, content-sharing and joint effort. Web-based social networking and interpersonal interaction appear to have a basic impact on individuals' lives the world over. There are some who wrangle about whether it is enhancing or devastating relational abilities. Sitting behind a PC speaking with digital companions can be simple and fun, however, can debilitate a man's verbal relational abilities. Online networking is a rising pattern on the planet today. Relational abilities are exemplified by the utilization of web-based social networking organizing. Web-based social networking organizing takes into account a correspondence outlet. Online networking is being used by understudies, guardians, organizations, and religious associations. It is being utilized as a part of many structures by a wide range of stages for some reasons.

The utilization of Social Network Sites is expanding these days particularly by the more youthful eras. It enables groups to share their interests and encounters. Clients would access be able to their Social Network Sites from their cell phones, PCs or portable workstations at whenever and anyplace.

Top most popular Social Networking Sites are:

- Facebook
- Youtube
- Instagram
- Twitter
- Google-Plus

- LinkedIn

Fig. 1.1 represents the most popular social networking sites. Almost all teenagers today have used social media.

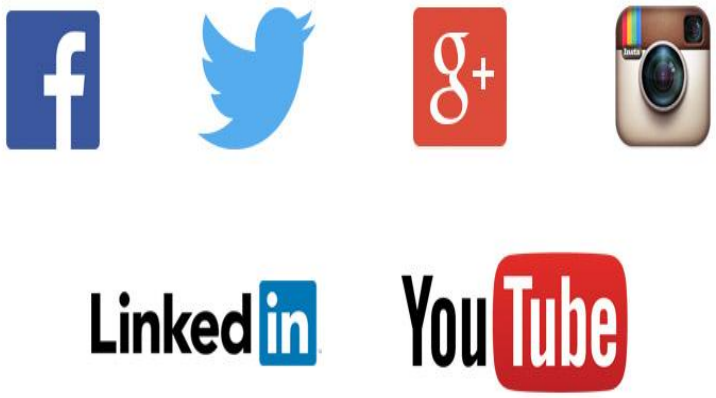


Fig. 1. 1 Social Networking Sites (SNS)

Nine out of 10 (90%) 13-to 17-year-olds have utilized some type of online networking. Three out of four (75%) young people right now have a profile on a long range informal communication site, and one in five (22%) has a present Twitter account (27% have ever utilized Twitter). Facebook completely commands long range informal communication use among youngsters: 40% of all high scholars say Facebook is their primary interpersonal interaction site, contrasted with 20% for Twitter, 8% for Youtube, 6% for Twitter, 1% for Google-Plus, and 1% for LinkedIn (24% don't have a person to person communication site).For by far most of the adolescents, social and other advanced interchanges media are a day by day part of life. Online networking is an exceptionally dynamic and quick moving space. The pie chart of young person usage of informal communication destinations has appeared in Fig 1.2. The accessibility of Social Network Sites enables clients to express their interests, sentiments and offer their day by day schedule. Gathering User produced content from any Social systems administration destinations could be utilized as a part of wellbeing related human practices.

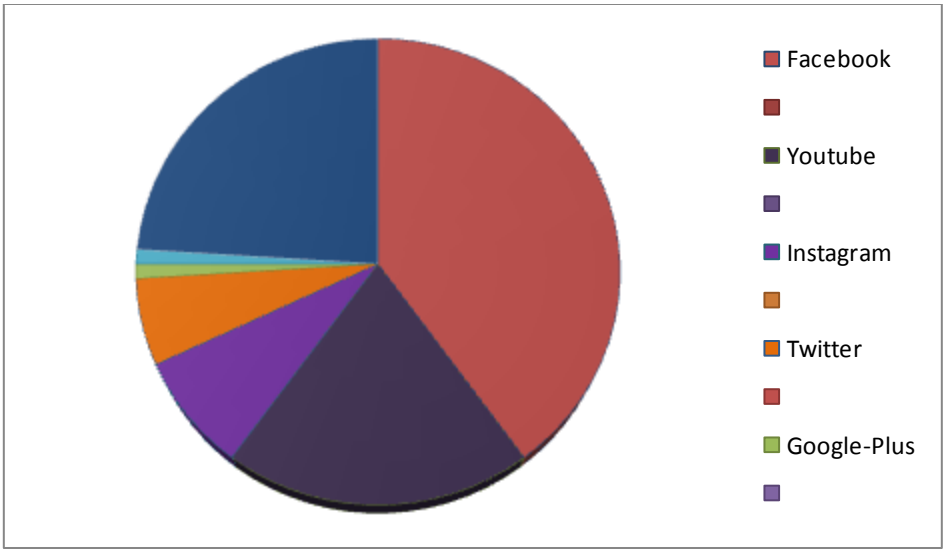


Fig. 1. 2 Social networking percentage among teenagers

From the review [1] it is discovered that Facebook is the most well known long range informal communication site among all. The review directed over online networking is given in Table1.1:

Table 1. 1 Monthly visitors of social networking sites

S.No.	SOCIAL NETWORKING SITES	MONTHLY VISITORS
1	FACEBOOK	1,790,000,000
2	YOUTUBE	1,000,000,000
3	INSTAGRAM	500,000,000
4	TWITTER	313,000,000
5	GOOGLE+	111,000,000
6	LINKEDIN	106,000,000

In Fig1.3 appears, the popularity chart of long range informal communication destinations has been plotted. This is person to person communication destinations versus various month to month visitors diagram. From the given figure it has been discovered that one of the long range informal communication locales i.e. Facebook have a huge number of month to month guests. After that Youtube guests are roughly 1000 million. The ubiquity of interpersonal interaction locales expands step by step,

these destinations likewise valuable client's day to the everyday schedule. Facebook is a person to person communication benefit that propelled in February 2004. In February 2012, Facebook has more than 845 million dynamic clients. A large number of individuals utilize Facebook consistently to stay aware of companions transfer a boundless number of photographs, share connections and recordings, and take in more about the general population they meet. The client may join normal intrigue client gatherings, composed by working environment, sharing your area, appearing out their own status and texting to every one of your companions. Furthermore, Facebook gives a little stage to all clients that to play some applications like recreations and test to engage their clients. The Facebook client can likewise share their video clips and photographs to the greater part of your companions.

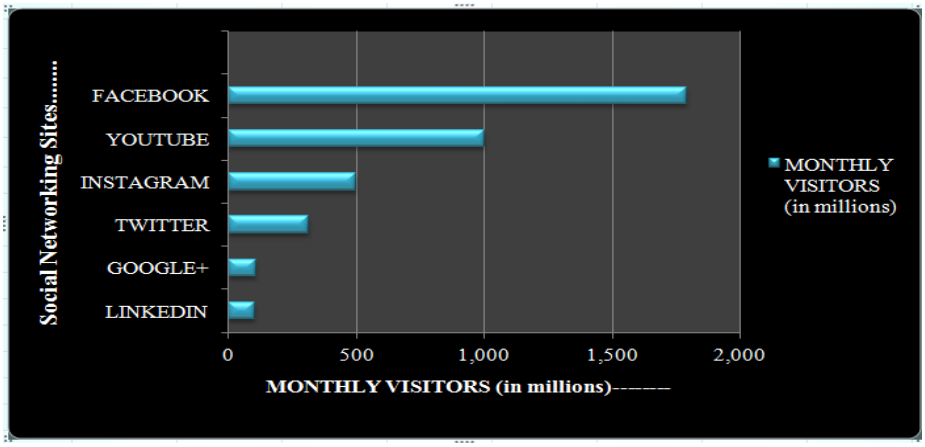


Fig. 1. 3 Popularity Graph of Social Networking Sites

Web-based social networking ought not to be seen essentially as the stages whereupon individuals post, but instead as the substance that is posted on these stages. These substances change significantly from district to area, which is the reason a relative report is important. The route in which depict online networking in one place ought not to be comprehended as a general portrayal of web-based social networking: it is somewhat a provincial case. Web-based social networking is today a place inside which mingle, not only methods for correspondence. Preceding web-based social networking, there were predominantly either private conversational media or open telecom media.

1.1 Facebook Check-ins

In this day and age, Smartphone gives a considerable measure of components including a guide, where you would find be able to your area. Facebook places are

practically same as this usefulness, where you can associate with your companions close to your area. Fig 1.4 demonstrates the logo of Facebook registration. Facebook places work along these lines: when a client visits some place and needs to share the area on Facebook, and after that registration highlight of Facebook comes into the photo. A client seeks the place in the wake of tapping the registration catch and if the pursuit coordinates the area is set, which is then appeared on your course of events with the photo or the post. Registration enables a client to label companions or includes a short portrayal. Client registration has picked up Vogue via web-based networking media stages off late and significantly related work is that of prescribing registration areas to clients. This inspired one to apply the psyche into the 'Facebook Vs: Predicting Check-Ins' dataset on Kaggle. This opposition was a piece of a renewing occasion held by Facebook. While in most recommended frameworks, the client profile is worked, here the assignment is to suggest checking completely in light of the area and time spent by the client. The dataset contains around 30millions individual registration for 108389 novel spots.

Fig. 1. 4 Facebook Checkins

Customers who wish to pronounce their zone to their buddies on Facebook would tap a "check in" catch to see a summary of spots contiguous, and subsequently, pick the place that matches where they are. In the wake of checking in, your enlistment will make a story in your sidekicks' News Feeds on Facebook and show up in the Recent Activity region on the page for that place.

Registration administrations are an extension of casual association structures stretching out from the desktop (PCs) to versatile stages. The various enlistment organizations, for instance, Facebook and Foursquare, engage customers to make

records and exhort their colleagues and have influenced exactly when and how people spread information. Differentiated and ordinary PCs and site organizations, registration engage customers in checking in at specific zones and talking with people in their framework, likewise, don't purpose of imprisonment customers' disclosure of constant region information. That is, registration empower customers to dissipate and get information in a flash and to change the progress and methodology of information exchange.

Registration information, which appears on long range interpersonal communication destinations put pages, customer's friend's newsfeeds, and customer's timetables when customers share their experiences and use information, and likewise can affect the essential administration of other framework people. Enlistment records have in like manner showed up on customers' courses of occasions and customers' partners' newsfeeds, growing correspondence, and association inside SNS. Such enrollment direct is seen as a social association, information disclosure, and social, yet the information spread through enlistment fluctuates transformation process and progression from standard social correspondence.

For example, people eating in a diner would now be able to in a blaze take a photograph of a dish, make a review out of the restaurant, and Mark sidekicks on Facebook by using registration organizations, however, beforehand, people could simply spread messages in the wake of leaving the diner Individuals would check be able to in while or even before they start their activities in a space. Toward the day's end, the present change of mobile phones has engaged the diffusing of information at any range and at whatever point, even before usage instead of traditional information dispersal. To improve promoting works out, it is useful to investigate the motivation driving current enlistment works on using mobile phones.

1.2 Social media for big data using machine learning

At this age of digital information, technology is growing day by day and with the use of technology, the size of data is also hiking up, which is named as Big Data.The digital data is sizing up enormously with every entry in all fields, which makes currently used storage device outdated since the storage capacity of this generation of computers are low. With large dataset, it is very difficult to handle to work on the standalone machine. To cope with this problem a new solution is introduced called Cloud Computing.

There are various applications in the territory of Social Media, Education and Medical. However, the advantage of Cloud for social media purposes is consistent, especially because of the huge information produced by the social networking sites. This bulk data can be overseen through huge information analytics and hidden patterns can be removed utilizing machine learning techniques. Specifically, the issue in the web-based social networking area is the prediction of social networking sites which can be settled through the summit of Cloud Computing and Machine Learning.

In this way, an attempt has been settled on to propose an intelligent decision support model that can help Facebook specializes in foreseeing most mainstream area in view of the verifiable information of Facebook check-ins. There are different cloud stages for machine learning applications.Cloud computing is the conveyance of on demand registering assets. With the help of computing, transfer of data over the server and set a virtual environment. This gives the ease of entering and storing a big amount of data in the system.

In this work, the proposed framework utilizes cloud innovation. The main reason is the monster size of the dataset and the second one is the precision of the dataset. At the point when the model is run, at that point, it rattled off the outcomes in parts of seconds which are snappier than established model. Distinctive machine learning estimations have been realized on the coronary disease dataset to anticipate precision for coronary sickness. In this system, analysts assemble bona fide data of nature of that territory and endeavor to make a model in light of it to foresee how the atmosphere will progress over a time period.

In machine getting the hang of, Ensembling is the computerization of the essential leadership prepares that additions from instances of the past and duplicate those decisions actually. Emulating the decisions subsequently is a focal thought in the farsighted examination. The word Ensembling reliably recollects science class, discovering the portrayal of animals. In case anybody can recall how these classes are portrayed and comprehend that there were certain properties that specialize found in existing animals, and in light of these properties, they arranged another animal. Other honest to goodness instances of course of action could be, for instance, when a man visits the pro. He/she advances certain request, and in light of their answers, he/she can recognize whether they have a particular infection or not.

1.3 Thesis Organization

Organization of thesis is structured into various chapters. Outline of each chapter structured as below:

Chapter 1(Introduction) gives the basic review of the thesis topic i.e. starting with aim and objectives, followed by the introduction of Facebook check-ins, its history, and applications. It gives the brief description of machine learning and its types. With social media as background chapter includes a brief description of prediction most popular location using machine learning concern with this research work.

Chapter 2 (Literature Survey) summarizes the work done by many researchers in which they describe how to predict the given dataset using machine learning approach on the standalone machine as well as a cloud environment. Research's that have been made in both fields have been considered and methods that have been used from the past under these fields have been described in details. Ensembling techniques are the popular methods available to predict the given data set and have been described in detail along with their feature, objectives, and shortcomings.

Chapter 3(Problem Statement and Objectives) includes the description of problem statement that has been chosen for the research. At last objectives are defined that are needed to be attained for successful completion of research work.

Chapter 4(Tools and Methodology) provides a new methodology that has been proposed with a motive to address the major issue. The methodology aims to describe how data flow in various steps. The main feature of proposed algorithm is to attain objectives and results with large data set thereby making the algorithm more optimal.

Chapter 5(Implementation and Results) includes a comparative study that is performed between proposed methodologies in terms of objectives that are achieved and shortcoming of various existing approaches. Comparative analysis shows that proposed method provides a better result in terms of already existing method, quality of data. Results have been described in tabular form at last. The method has been verified by applying the algorithm on large data set gathered from Facebook check-ins data set. The method is analyzed with different real-time data, situations, and scenario to examine its effectiveness.

Chapter 6(Conclusion and Future Scope) includes a brief description of what type of algorithm and platforms has been used for better result and it also describes how much the result obtained by proposed model gives the highest accuracy. The scope for further research has been stated clearly under the future scope section.

CHAPTER-2

LITERATURE SURVEY

This chapter summarizes the work done by many researchers to predict the most popular location using machine learning techniques on the standalone machine as well as a cloud environment. Research that has been made in the field of machine learning, Ensembling technique, Facebook has been described in detail. Research gaps after survey has been listed out. Ensembling techniques predict the result have been described in detail along with features, objectives, and shortcomings.

2.1 Related Work

J. Vernack [8] proposed the examination to anticipate whether a client will check in a given area later on in light of his Facebook information, to survey the additional benefit of utilizing the whole Facebook dataset far beyond information utilized as a part of past research and to decide the most critical variable gatherings and single factors. The outcomes demonstrate that it is practical to make a model with a high prescient execution as proposed model has an average Area Under the receiver operating Curve (AUC) of 0.74 and a normal Accuracy of 0.80. Besides, the outcomes demonstrate the additional, benefit of utilizing the whole Facebook dataset far beyond the restricted dataset that was utilized as a part of past research as this could build the AUC by 13.62% and the Accuracy of 2.44%. The investigations demonstrate that client area particular factors are the most critical factors and that the most imperative indicators were rule of registration, amongst Club and a client's closest development zone, how drawn in a client's companions are with the area, how connected with a client is with the area, and how often a client checked in the area some time recently. To the best of this research learning, this is the principal consider that assesses the additional benefit of utilizing the whole Facebook dataset in the forecast of future registration. Also, this investigation utilizes a somewhat unique research question contrasted with past research in view of Facebook information. This exploration predicts whether a client will check in a given area later on as opposed to foreseeing a client's next registration. Organizations could utilize these expectations keeping in mind the end goal to tweak their promoting arrangement.

J. Lin, R. Oentaryo, E. P. Lim, C. Vu, A. Vu & A. Kwee [10] proposed to address the issue by researching the utilization of openly accessible Facebook Pages information

which incorporates client "registration", sorts of business, and business areas—to assess a client chose physical area concerning a kind of business. Utilizing a dataset of 20,877 examples for sustenance organizations in Singapore, creators direct an investigation of a few key elements including business classes, areas, and neighboring organizations. From these elements, the proposed work extricates an arrangement of significant components and builds up a strong prescient model to appraise the fame of a business area. the proposed tests have demonstrated that the notoriety of neighboring business contributes the key elements to perform precise forecast. At last, delineate the pragmatic utilization of their proposed approach by means of an intelligent web application framework. In this work, utilization of information gathered from Facebook Pages to answer essential research inquiries, for example, "Where should a proprietor set up his physical retail location at, in order to upgrade the store's ubiquity?", "What are the imperative components affecting a store's prevalence?", and "Is there a "nearby" impact, whereby organizations would benefit be able to from the nearness of more prevalent/set up neighbors?" To this end, this exploration proposes another area examination system that works on top of Facebook Pages information. The centerpiece of the present structure is the accompanying forecast undertaking: Given an objective area that a business/property proprietor needs to theoretically set his/her store at, in what manner would extract be able to the pertinent information of organizations inside the region of the objective area and utilize them to appraise the target of the objective area.

S. Kumar, D. Kulkarni, B. Padmanabha & N. Y. Murali [11] traces the results of an exact investigation of the Facebook registration information and depicts a model planned in light of the examination to foresee registration areas. The components accessible are the area directions, time and precision for each registration. An expansive number of conceivable registration areas (38,000). Assessment is performed utilizing the mean exactness at 3 measurements. This examination utilizes a generative way to deal with demonstrate the conveyance of the elements. The model is then used to yield the main three positioned forecasts for each test point. It is demonstrated experimentally that this technique beats discriminative grouping strategies in wording mean exactness at 3. This examination additionally watched that there are some freedom designs among the components and this is utilized to calculate the generative model and learn singular models for autonomous arrangements of factors. There are a few actually noteworthy perceptions about the information which

are utilized to build highlights as needs are. The subsequent model is aggressive with beat arrangements on the Kaggle leader board with a substantially littler measure of highlight building.

J. D. Zhang, C. Y. Chow & Y. Li, [12] proposed a customized and productive geological area suggestion structure called iGeoRec to take the full favorable position of the land impact on area proposals. In iGeoRec, there are mostly two difficulties: first is customizing the geological impact to precisely anticipate the likelihood of a client going to another area, and second is productively registering the likelihood of every client to every new area. To address these two difficulties, creators proposed a probabilistic way to deal with customize the land impact as an individual circulation for every client and foresee the likelihood of a client going to any new area utilizing client individual conveyance. Besides, this work build up a productive guess strategy to figure the likelihood of any client to every single new area; the proposed technique decreases the computational intricacy of the correct calculation technique from O(|L|n3) to O(|L|n) (where L is the aggregate number of areas in an LBSN and n shows quantity of registration areas of a client). At long last, this exploration direct broad tests to assess the suggestion precision and effectiveness of iGeoRec utilizing two huge scale genuine informational collections gathered from the two of the most well-known LBSNs: Foursquare and Gowalla. The test comes about demonstrate that iGeoRec gives essentially better execution looked at than other cutting edge geological suggestion systems.

P. Luarn, J. C. Yang & Y. P. Chiu,[13] the essential goal of this examination is to create and refine a reasonable structure from social verbal inspirations and the portable point of view to give a hypothetical comprehension of the inspirations that actuate purchasers to take part within proper limits in conduct. The outcomes demonstrate that the social condition (e.g., tie quality, subjective standards, expressiveness, social support, and data sharing) assume the most basic part of inspiring individuals to take part under control in conduct. Furthermore, the perceptual (e.g., saw a social advantage, saw pleasure, and saw esteem) and utilization based conditions, (for example, consumer loyalty and communicator contribution) additionally persuade individuals to take part in the registration conduct and to spread their utilization encounters by utilizing cell phones. The outcomes give certain hypothetical and pragmatic ramifications for advertising professionals in their

arranging of new promoting techniques to draw in buyer consideration and will add to a superior comprehension of registration conduct.

J. B. Gomes, C. Phua, & S. Krishnaswamy [27] proposed the area expectation through portable information mining use such huge information in applications, for example, activity arranging, area based publicizing, and shrewd asset designation; and additionally in recommender administrations including the exceptionally well known Apple Siri or Google Now. This work concentrates on the testing issue of foreseeing the following area of a versatile client given information on his or her present area. In this proposed work, Next Location - a customized versatile information mining structure - that utilizations spatial and transient information as well as other logical information, for example, an accelerometer, Bluetooth, and call/SMS log. Likewise, the proposed structure speaks to another worldview for security protecting next place forecast as the cell phone information is not shared without client authorization. Trials have been performed utilizing information from the Nokia Mobile Data Challenge (MDC). The outcomes on MDC information indicate extraordinary inconstancy in the prescient precision of around 17% of clients. For instance, sporadic clients are an extremely hard to anticipate while for more general clients it is conceivable to accomplish over 80% precision. At the point when analyzed against existing outcomes, this approach accomplishes the most elevated prescient exactness. At long last, proposed work an option plan of action for portable publicizing that utilizations Next Location structure.

2.2 Machine Learning Approach for Predicting Facebook check-ins

Machine learning is the investigation of propelling PCs to act without being unequivocally altered. In the earlier decade, machine learning has given us self-driving cars, reasonable talk affirmation, convincing web looks for, and a boundlessly improved appreciation of the human genome. Machine learning is so unavoidable today that one can in all probability use it commonly every day without knowing it. Various investigators also think it is a perfect way to deal with makes progress towards human-level AI. In this class, the client will get some answers concerning the best machine learning frameworks and get deal with completing them and motivating them to work for them. More critically, the client will get some answers concerning the speculative underpinnings of learning and also get the valuable know-how anticipated that would quickly and proficient apply these methodology to new issues.

Finally, the client will get some answers concerning some of Silicon Valley's acknowledged methodology in progression as per machine learning and AI.

2.2.1 Supervised Machine Learning

The dominant part of pragmatic machine learning utilizes supervised learning. Supervised learning is the place where a user has input factors (x) and a yield variable (Y) and utilizes a calculation to take in the mapping capacity from the contribution to the yield.

$$Y = f(X)$$

The objective is too rough the mapping capacity so well that when the user have new info information (x) that can foresee the yield factors (Y) for that information. It is called supervised learning in light of the fact that the procedure of a calculation gaining from the training dataset can be thought of as an instructor overseeing the learning procedure. Everyone knows the right answers; the calculation iteratively makes expectations on the preparation information and is amended by the instructor. Learning stops when the calculation accomplishes a worthy level of execution.

Supervised learning issues can be additionally gathered into classification and regression issues.

- **Classification:** A characterization issue is a point at which the yield variable is a class, for example, “blue" or “red" or "sickness" and "no infection".
- **Regression:** A regression issue is a point at which the yield variable is a genuine esteem, for example, “weight" or "dollars".

Some basic sorts of issues based on top of classification and regression incorporate suggestion and time arrangement expectation separately. Some prevalent cases of directed machine learning calculations are:

- Random forest for classification and regression problems.
- Linear regression for regression problems.
- Support vector machines for classification problems.

2.2.2 Unsupervised Machine Learning

Unsupervised learning is the place where a user just has input information (X) and no relating yield factors. The objective of unsupervised learning is to show the hidden structure or dissemination in the information keeping in mind the end goal to take in more about the information. These are called unsupervised learning in light of the fact

that dissimilar to regulate learning above there are no right answers and there is no educator. Calculations are left to their own devices to find and present the intriguing structure in the information. Unsupervised learning issues can be additionally assembled into association problems and clustering.

- **Association:** An association learning issue is the place a user need to find decides that depict extensive parts of the information, for example, individuals that purchase X likewise tend to purchase Y.
- **Clustering**: A clustering issue is a place where a person needs to find the intrinsic groupings in the information, for example, gathering clients by acquiring conduct. Some well-known cases of unsupervised learning calculations are:
- Apriori algorithm for association rule learning problems.
- K-means for clustering problems.

2.2.3 Ensemble Machine Learning

Ensemble model is the combination of no less than two machine learning and data mining computations which fill in as a joined unit to give out a perfect result from either count. This strategy is by and large used these days in look into in different fields.

Commonly, this methodology grows the general capability of the results because of the solidarity variable of the two models ending up noticeably potentially the most vital element. Regardless, this technique may in like manner reduce the capability in circumstances where the data is enormous and one model is better than other with an exceptional edge. This system moreover expects a basic part of growing the region of datasets for one topic of research. A wide variety of educational accumulations might be hurled at the model and this will, regardless, give a reasonable result where one single model won't perform well by any extent of the creative ability.

Ensembling method incorporates uniting different numerous expectations dictated by different learning calculations in order to make a more grounded general forecast and get best results. Ensembling is one of the more mainstream managed learning techniques for machine learning as which can be prepared and also used for making forecasts. Ensembling tends to yield overwhelming results when there is a tremendous grouped quality among the models is so far used.

2.3 Classification algorithms (using R-programming)

2.3.1 Principal Component Analysis

Principal Component Analysis is a method of pattern identification in a given data and representing these data in which one can identify the same as well as the different pattern. For the large dimension of data it can be difficult to find a pattern, when a user don't have any information about data in graphical representation form then this method is very useful for data analysis. One of the biggest advantages of PCA is that when a user find a pattern in a given data and a user can compress data, it means the user can reduce the number of dimension in a given dataset, without any loss.

2.3.2 Least Angle Regression

Least Angle Regression (LARS), comes with a new feature, it introduces how one can select an algorithm for new model, very useful and less eager version of regular selection methods. This algorithm implements the Lasso, and this is very simple modification of this algorithm, ordinary least squares is an attractive version that constrains regression coefficient of sum of the absolute value; for a given problem, Lasso calculates and estimates all possible values using the LARS modification method, and it is observed that it will take less computer time than previous one method. One new different modification method of LARS implements linear regression method i.e. Forward Stagewise i.e. another new model selection; for Stagewise and the Lasso same type of numerical previously result are observed it helps people for understanding both methods and their properties, for LARS algorithm it can be seen as simpler constrained versions.

2.3.3 Bayesian Ridge Regression

Bayesian Ridge Regression is one of the linear regression methods in which statistical analysis is taken as inward the context of Bayesian inference. Sometimes the model that are under regression has errors, having a normal distribution, and one can be assumed a prior distribution of particular form is assumed, then model's parameter of explicit results are feasible for the method of posterior probability distributions.

2.3.4 Stochastic Gradient Boosting

Stochastic Gradient Boosting is a way for solving regression as well as classification problems of machine learning technique, which crops a model of prediction in the

ensemble form of models which has the very weak type of prediction, like decision trees. It outlines the model in a manner of stage-wise like different techniques for boosting do, and it closes them by permitting an addition of a subjective differentiable loss function.

2.3.5 k-Nearest Neighbors

In acknowledgment of pattern, a non-parametric technique utilized for regression and classification is a k-Nearest Neighbors calculation (k-NN). In regression and also in classification case, in the element space, the info contains the k nearest preparing illustrations. The yield bet on whether the k-NN technique is utilized for classification or regression:

- The yield of the k-NN grouping is a class investment. A thing is ordered by a more number of votes of its closer thing, with the thing being labeled to the class most normal in the midst of its k closest neighbors (k is a positive whole number, regularly little). In the event that the estimation of k = 1, then the thing is just labeled to the class of that solitary adjoining neighbor.
- In k-NN regression, the subsequent result is the property rate for the thing. This esteem is the normal rate of its k closest nearby neighbors

2.3.6 Non-Negative Linear

Non-Negative least Squares (NNLS).In an optimization, the problem of the least squares of a constrained version where the coefficients of the given problem are not sanctioned to become negative. That is if it is assumed that a matrix X and its a response variables(column) vector of b, the main aim is to find

$$\text{Arg min}[a] \text{ subject to } a \geq 0.$$

Here $a \geq 0$ means that each component of the vector 'a' should be non-negative and $\|\bullet\|$ represents the norm.

Non-negative least squares problems converted as sub problems in matrix decomposition form, for example. In an algorithms for non-negative matrix/tensor factorization and PARAFAC. The recent can be assumed as a generalization of NNLS. Another type of generalization of NNLS is bounded-variable least squares (BVLS), with the same type of upper and lower bounds like $\alpha_i \leq x_i \leq \beta_i$.

2.3.7 Bayesian Regularized Neural Networks

Bayesian Regularized Neural Networks (BRNNs) strategy that is more capable and vigorous than the technique for standard back-proliferation nets, one can decrease and take out the purpose for protracted cross-approval. Bayesian regularization is a procedure that identified with the numerical approach and changes over straight relapse to nonlinear relapse into a "very much postured" factual dilemma in the way of another edge relapse technique. The advantages of BRNNs is that the models are the approval and hearty process, which set apart as O (n^2) in everyday relapse strategies, for example, back spread, is unimportant.

2.3.8 Boosted Tree

Boosted Tree work executes the `classical' inclination boosting trees using regression as base-learners. Typically, the comparable sort of calculation is actualized in gbm bundle. The fundamental contrasts of this technique are that the improvement of subjective loss capacities and it can be assigned by means of the family contention to blackboost in the other hand gbm additionally utilizes hard-coded misfortune capacities. Additionally, the base-learners (conditional inference trees, visually perceive ctree) are insignificantly more adaptable.

2.3.9 kk-Nearest Neighbors

In acknowledgment of example, the kk-Nearest Neighbor's calculation (kk-NN) is an approach for non-parametric and is utilized for regression and classification. In both cases, the data involves the k closest planning cases in the part space. The yield depends on upon whether kk-NN is used for relapse and order:

- In kk-NN classification, the yield is a class investment. A question is requested by a bigger part vote of its neighbors, with the challenge being consigned to the class most fundamental among its k nearest neighbors (k is a positive entire number, typically little). In case k = 1, then the question is quite recently consigned to the class of that single nearest neighbor.
- In kk-NN regression, the yield is the property estimation for the question. This regard is typical of the estimations of its k nearest neighbors.

2.3.10 The Lasso

In machine learning and statistics, Lasso (least absolute shrinkage and selection operator) (additionally Lasso or LASSO) is a method of regression analysis that

performs both variable choice and regularization to upgrade the expectation exactness and interpretability of the actual model it produces Lasso was initially detailed for minimum squares models and this basic case uncovers a significant sum about the conduct of the estimator, including its relationship to edge relapse and best subset choice and the associations between lasso coefficient gauges thus called delicate thresholding. It additionally uncovers that the coefficient gauges require not be novel if covariates are collinear.

2.3.11 Conditional Inference Tree

Conditional Inference Tree is a method of the regression trees of non-parametric class that embedding regression model of tree-structured into a well-defined subject of procedures of conditional inference. This method is applicable to all types of regression quandaries, including numeric, nominal, censored, ordinal as well as arbitrary quantification scales of the covariates and multivariate replication variables. This scene encompasses a practical guide to exploiting the extensible and flexible computational implements in party kit for visualizing and fitting conditional inference trees.

2.3.12 Elastic Net

In statistics and machine learning, the Elastic Net is a method of regularized regression that linearly amalgamates the L1 and L2 penalties of the ridge and lasso methods in particular, in the fitting of logistic or linear models of regression.

2.3.13 Decision Tree

A Decision Tree Technique is a choice bolster that is used to implement tree-like model or chart of choices and their conceivable number of results, including chance occasion asset costs, results, and utility. It is one approach to displaying a calculation. Choice trees are normally used in solidly in the choice investigation, operations research, to benefit distinguish a methodology most at risk to achieve an objective, however, are also a prominent actualize in machine learning.

2.3.14 Support Vector Machines

In an algorithm of machine learning, Support Vector Machine is a part of supervised learning models with the algorithm of associated learning that are used for the analysis of classification and regression. In this approach, an optimal disuniting hyper plane between two or more classes that is obtained by increasing the margin between

the fortification vectors or the boundary points. To remove or reduce their influence rate, the points which are on the erroneous side of the discriminate margin are weighted down. This approach can be optically discerned as a quadratic optimization quandary which can be solved by the given techniques such as kernel estimator.

2.3.15 Multivariate Adaptive Regression Spline

Multivariate Versatile Regression Splines (MARS) are a kind of regression investigation. This is one of the non-parametric regression strategies and can be optically recognized as an augmentation of straight models that consequently models association's non-linearity and between factors. The term utilized here i.e. "MARS" is authorized and trademarked to Salford Systems. With a specific end goal to avoid trademark encroachments, numerous openly accessible source executions of MARS are called "Earth".

2.4 Classification algorithms (using Apache Mahout)

2.4.1 Stochastic Gradient Descent Algorithm

Gradient descent limits the cost work. For vast datasets, Gradient descent is an exceptionally costly system. Stochastic Gradient Descent (SGD) is an alteration of the Gradient descent calculation to deal with huge datasets. Inclination drop figures the slope utilizing the entire dataset, while SGD registers the gradient utilizing a solitary specimen. Along these lines, gradient descent stacks the full dataset and tries to discover the neighborhood least on the chart and after that rehash the full procedure, while SGD changes the cost work for each specimen, one by one. A noteworthy favorable position that SGD has over Gradient descent is that its speed of calculation is a ton quicker. Substantial data sets in RAM, for the most part, can't be held as the capacity is restricted. In SGD, the weight on the RAM is decreased, wherein each test or group of tests are stacked and worked with, the outcomes for which are put away, and so on.

2.4.2 Naïve Bayes Algorithm

For Naïve Bayes algorithm, one ought to have a full understanding of Bayes Rule and conditional probability. In extremely straightforward terms, the conditional probability is the likelihood that something will happen, given that something else has as of now happened. It is communicated as P (A/B), which can be perused as the

likelihood of A given B, and it finds the likelihood of the event of occurrence A once occasion B has as of now happened. Numerically, it is characterized as takes after:

$$P(A|B) = P(A \cap B) / P(B)$$

In Bayes' hypothesis, one can have seen that the result is constructed just with respect to one proof, however, on classification issues, and have various confirmations and need to foresee the result. In Naïve Bayes, one can uncouple different bits of confirmation and treat every one of them freely. It is characterized as takes after:

$$P(\text{result} \mid \text{numerous Proof})) = P(\text{Proof 1}|\text{result})^* P(\text{Proof 2}|\text{result})^* P(\text{Proof 3}|\text{result}) \ldots . / P(\text{Evidence})$$

Run this equation for every conceivable result. Since attempting to arrange, every result will be known as a class. The errand is to take a gander at the proof (elements) to consider how likely it is for it to be of a specific class and after that relegate it in like manner. The class that has the most elevated likelihood gets doled out to that blend of confirmations.

2.4.3 Hidden Markov Model

The Hidden Markov Model (HMM) is one of the most interesting topics of the classification method. It is a Markov model in which the states are covered up. This is the method of classification to predict the conditions of a framework by watching the results without approaching the real states themselves. Utilizing the Hidden Markov Model, three sorts of issues can be settled. The initial two are identified with the example acknowledgment issue and the third kind of issue produces a Hidden Markov Model, given a succession of perceptions. How about to take a gander at these three sorts of issues:

- **Decoding:** This is finding the most plausible grouping of concealed states from a few perceptions. It has been utilizing the Viterbi calculation to decide the most plausible arrangement of concealed states when one can have a grouping of perceptions and an HMM.
- **Evaluation:** This is discovering the likelihood of a watched succession, given an HMM. From various HMMs that depict diverse frameworks and a succession of perceptions, the objective will be to discover which HMM will most likely produce the required arrangement. It has been utilize the forward calculation to ascertain the likelihood of a perception grouping when a specific HMM is given and find out the most plausible HMM.

- **Learning:** Learning is producing the HMM from a succession of perceptions. Along these lines, if one can have a succession, it may ponder which is the in all probability model to produce this succession. The forward in reverse calculations are valuable in tackling this issue.

2.4.4 Random Forest

One of the most popular methods of classification is Random Forest. It begins with a machine learning procedure called decision tree. The Random Forest algorithm was produced by Adele Cutler and Leo Breiman. Random Forest develops numerous order trees. They are an ensemble learning technique for characterization and relapse that builds various decision trees at preparing time and likewise yields the class that is the method of the classes yielded by individual trees. Single decision trees demonstrate the bias–variance tradeoff. So they, for the most part, have high, change or, on the other hand, high inclination. The accompanying is the parameters in the calculation:

- **Variance:** This is a mistake that reaches from affect ability to little vacillations in the preparing set Random Forest to alleviate this issue by averaging to locate a characteristic adjust between two extremes. A Random Forest takes a shot at showing, which is to normal boisterous and fair models to make a model with low fluctuation. A Random Forest calculation fills in as a substantial gathering of de correlated choice trees.
- **Bias:** This is a mistake brought about by a wrong supposition in the learning calculation.

2.4.5 Multilayer Perceptron

An artificial neural network or neural network, by and large, alludes to an MLP organize. The characterized neuron as a usage in PCs in the past area. An MLP organize comprises of various layers of these neuron units. The primary layer of the MLP speaks to the information and has no other reason than steering the contribution to each associated unit in a sustain forward form. The second layer is called shrouded layers, and the last layer fills the extraordinary need of deciding the yield. The enactment of neurons in the shrouded layers can be characterized as the whole of the heaviness of all the info.

The MLP usage depends on a broader neural system class. It is actualized to keep running on a solitary machine utilizing Stochastic Gradient Descent, where the

weights are refreshed utilizing one information point at once. The quantity of layers and units per layer can be indicated physically and decides the entire topology with every unit being completely associated with the past layer. An inclination unit is consequently added to the contribution of each layer. An inclination unit is used for moving the enactment capacity to one side or right. It resembles adding a coefficient to the straight capacity.

2.5 Research Gaps

The following research gaps have been traced from the study of literature in prediction of most likely location using Facebook data.

- The work done so far has utilized only for small datasets for the problem in hand.
- No work has been done in the large check-ins datasets of Facebook using Apache Mahout.
- The literature work reveals that only three algorithms have been exploited for prediction of most likely location using Facebook dataset. No other classification algorithms have been explored yet.

CHAPTER 3

PROBLEM STATEMENT AND OBJECTIVES

This chapter states the problem in hand and lists out various objectives that are required to be met for solving the problem.

3.1 Problem Statement

User check-ins has gained trend on social media platforms off late and important strive is that of proposing check-in locations to users. The platform of Facebook other than offering a way to connect to your distant away friends provides a facility to user to share the site he or she visits, using check-in feature of Facebook. A user hunts the place after clicking the check-in button and if the search matches, the site is set which is then revealed on your timeline with the picture or the post. Check-ins permits a user to tag friends or attach a short description.

The problem in hand focuses on prediction of most popular places for the user that he/she would like to check-in using Facebook check-ins data set. The main objective of the current work is to design a model which aids in prediction of most likely locations.

3.2 Thesis Objectives

Various objectives that are needed to be fulfilled to solve the problem in hand are listed as below:

- To study various techniques and tools available for finding a most popular location using facebook.
- To develop an Ensembled approach using R programming to predict which place a person would like to check-in, using small chunks of Facebook check-ins dataset, based on their location accuracy and time sharing attributes.
- To develop more efficient machine learning Ensembled approach using and Apache Mahout using large Facebook check-ins dataset.
- Performance comparison of Facebook check-ins prediction models on a standalone machine(R Tool) and using Apache mahout cloud environment.

CHAPTER 4

TOOLS AND METHODOLOGY

4.1 Tools

Tools that are used for implementation of the problem solution are as follows:

- R Studio: Version 0.99.473 - © 2009-2015 R Studio, Inc.
- Microsoft Excel 2007
- mahout-0.9+cdh5.8.0.p0.56.el6.noarch

4.2 Methodology

The methodology used for building multiple classification models is shown in Fig 4.1. The model methods are categorized into 2 subparts. The first module has been used to predict the most likely location as per the user demand employing different machine learning approaches and the model with best accuracy has been chosen. The next module describes the selection of best algorithm in the previous module and implements Ensembling operation. For machine learning study this thesis cover up supervised and unsupervised learning techniques also cover up the methods used in classification techniques in Chapter 2. Various classification models have been studied.

4.2.1 Using R Tool on Standalone machine Environment

The R computer programs are an essential tool for progression in the numeric examination and machine learning spaces. R is a perfect way to deal with make reproducible, extraordinary examination. R is extensible and offers rich value for architects to manufacture their own specific gadgets and procedures for examining data. With machines winding up recognizably more basic as data generators, the noticeable quality of the dialects must be depended upon to create. When it at first turned out, the best-favored angle was that it was free programming. The vastness of package organic framework is irrefutably one of the R's most grounded qualities - if a true technique exists, odds are there's presently an R package out there for it. R's positive conditions fuse its package natural framework.

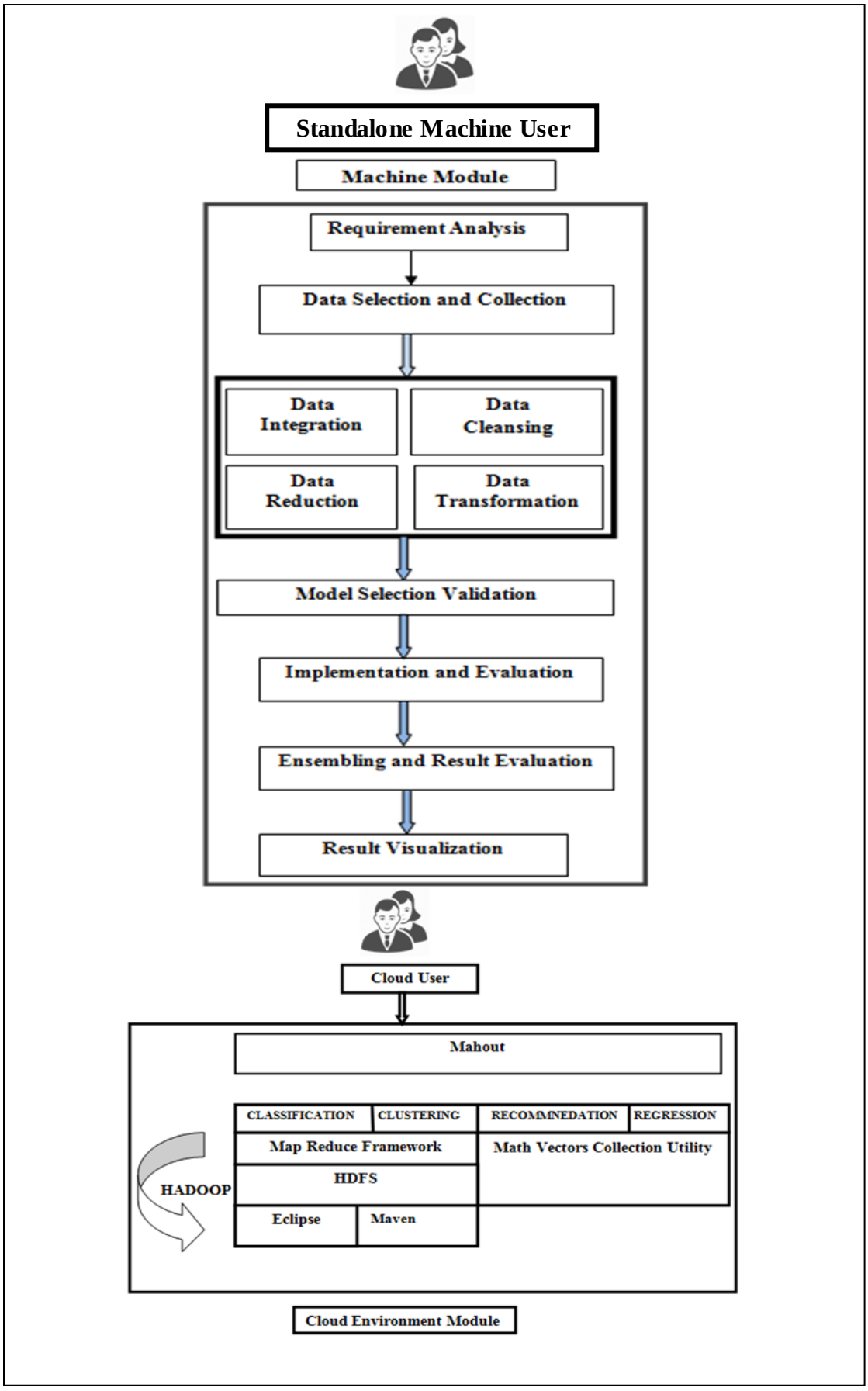


Fig.4. 1 Methodology

In this module, the accuracy of different machine learning algorithms has been explored using R Tool on the Standalone machine. Here initial analysis has been done using Microsoft excel. A csv file has been provided as an input for R-Studio. Analysis has been done using programming language R.

- The data is collected from web sources followed by Pre-processing of Data which includes Data cleaning, Integration of data, Data Reduction and Data Transformation.
- Data has been thoroughly cleaned, removing erroneous records. Interpolation is used to deal with missing values.
- Feature Selection and Correlation analysis is used for selecting most relevant predictor attributes.
- After this different machine learning approaches such as Principal Component Analysis, Least Angle Regression, Bayesian Ridge Regression, Stochastic Gradient Boosting, k-Nearest Neighbors, Non-Negative Linear, Bayesian Regularized Neural Networks, Boosted Tree, The Lasso, Conditional Inference Tree, Elastic Net, Decision Tree, Support Vector Machines, Multivariate Adaptive Regression Spline are implemented to predict the location id on the dataset cleaned in earlier steps.
- Different evaluation measures are used. Henceforth, the outcomes have been analyzed on the basis of accuracy. Further, top three models with the highest accuracies have been ensemble for improving the accuracies of the models by combining the efficiency of best three models.
- In any case, R has both upsides and downsides that designers ought to know. With enthusiasm for the programming developing, as appeared on language notoriety files, for example, TIobe, Redmond and PyPL, R initially showed up in the 1990s and has filled in as an execution of the S measurable programming languages.

- "R is the most mainstream dialect utilized as a part of the field of statistics."It has all the adaptability and power. R is in reality only accumulations of scripts that are sorted out into projects."
- Data purifying/cleaning is a term identified with getting the significant data from the crude information and noisy data removal (information not profitable

to us). This should be possible effectively in Microsoft Excel and is a generally utilized strategy for each information researcher.

4.2.2 Cloud Environment for Apache Mahout Platform

In this module, various classification machine learning methods which are scalable to mahout environment has been used.

- Stochastic Gradient Descent, Naïve Bayes, Hidden Markov Model, Random Forest and Multilayer Perceptron have been executed to foresee the event of most prominent location on the cleaned dataset.
- After this, three best algorithms have been selected based on their accuracy values and ensemble approach applied on it.
- The result originated from this ensemble method gives the best result as compare to existing model. A mahout is a man who rides and controls an elephant.

The vast majority of the classification in Apache Mahout is actualized on top of Hadoop, which is another Apache-authorized venture and has the image of an elephant. As Apache Mahout rides over Hadoop, this name is supported. Apache Mahout is a venture of Apache Software Foundation that has usage of machine learning calculations. Mahout was begun as subprojects of the Apache Lucene extend in 2008. After some time, an open source extends named Taste, which was produced for communitarian separating, and it was consumed into Mahout. Mahout is composed in Java and gives adaptable machine learning calculations. Mahout is the default decision for machine learning issues in which the information is too substantial to fit into a solitary machine. Mahout gives Java libraries and does not give any UI or server. It is a structure of devices to be utilized and adjusted by designers.

4.2.3 Algorithms supported in Mahout

The algorithm supported in mahout has been studied which helps to work on cloud environment. The execution of calculations in Mahout can be classified into two gatherings:

- **Sequential algorithms:** These calculations are executed successively and don't utilize Hadoop adaptable handling. They are normally the ones gotten from Taste. For example: logistic regression, user-based collaborative

filtering, singular value decomposition, , multi-layer Perceptron, Hidden Markov Model

- **Parallel algorithms:** These calculations can bolster peta bytes of information utilizing Hadoop outline consequently lessen parallel handling For example, k-means clustering, Naïve Bayes, canopy clustering, spectral clusters, Random Forest, and so on.

4.2.4 Contemplate reasons behind Mahout being a good choice for classification

In machine learning frameworks, the more information utilize, the more precise the framework constructed will be. Mahout, which utilizes Hadoop for versatility, is a path in front of others as far as taking care of gigantic datasets. On the off chance that the info estimate for preparing cases is from 1 million to 10 million, at that point Mahout is an incredible decision. For order issues, expanded information for preparing is alluring as it can enhance the precision of the model. By and large, as the quantity of datasets expands, memory prerequisite additionally increments and calculations turn out to be moderate, however, Mahout's adaptable and parallel calculations work better with respect to the time taken. Each new machine included declines the preparation time and gives higher execution.

4.3 Evaluation Criteria Used for Classification

4.3.1 Correlation Matrix

The confusion matrix is also called as Error matrix. It is a table that is often used to describe the performance of a classification method on a set of test data for which actual value are known. Each row of the matrix represents the instances in the actual class. Each column of the matrix represents the instances in a predicted class. the correlation matrix is represented as:

		Predicted	
		No	**Yes**
Actual	**No**	**TN**	**FP**
	Yes	**FN**	**TP**

True Positive: Case in which a user predicted “Yes” and the locations actually exist.

True Negative: Case in which a user predicted "No" and the locations actually don't exist.

False Positive: Case in which a user predicted "Yes" and the locations actually don't exist.

False Negative: Case in which a user predicted "No", but the locations actually exist.

4.3.2 Accuracy and Precision

In classification, accuracy and precision are two important evaluation parameters. Accuracy is defined as the sum of true positive and true negative instances divided by 100.And Precision is fraction of true positive and predicted yes instances. The formula of Accuracy and Precision are given below:

$$\text{Accuracy: } \frac{TP+TN}{100} \qquad \text{Precision: } \frac{TP}{\text{Predicted Yes}}$$

4.3.3 Recall and F-Square

Recall is defined as the fraction between True Positive instances and Actual yes instances whereas F-Square is the fraction between product of the recall and precision to the summation of recall and precision parameter of classification. The formula of recall and precision given below:

$$\text{Recall: } \frac{TP}{\text{Actual Yes}} \qquad \text{F-Square: } \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

4.3.4 Sensitivity, Specificity and ROC

Sensitivity is defined as the fraction of true positive and actual yes instances whereas specificity is the difference between one and false positive rate value.ROC is defined as the fraction between true positive rate and false positive rate.

$$\text{Sensitivity: } \frac{TP}{\text{Actual Yes}} \qquad \text{Specificity: } 1 - \frac{FP}{\text{Actual No}}$$

$$\text{ROC: } \frac{TPR}{FPR}$$

4.3.5 Significance and analysis of ensemble method in machine learning

Ensemble technique includes consolidating various multiple predictions determined by various learning algorithms so as to make a stronger overall prediction and get best outcomes. Ensembling is one of the more popular supervised learning methods of machine learning as which can be trained and additionally utilized for making predictions. Ensembling tends to yield predominant outcomes when there is a huge assorted quality among the models is so far utilized. Ensemble modeling is the most vital piece of this research in which consolidating at least two machine learning models. A brief description of ensemble models is done in the Chapter 5 (Implementation and Results).

CHAPTER-5

IMPLEMENTATION AND RESULTS

This chapter incorporates an investigation that is performed when proposed technique is applied on dataset regarding destinations that are accomplished and weaknesses of different existing methodologies. Relative examination demonstrates that proposed strategy gives better outcomes in term of the outcome produced by existing model. Experimental results have been described in tabular form. This chapter is divided into 3 sections:

1. Measuring performance using R Tool on standalone machine
2. Measuring performance using Apache Mahout
3. Comparison between the results

5.1 Measuring performance using R Tool on standalone machine

Implementation technique using R Tool on standalone machine can be illustrated in following flow of data:

- Dataset collection and description
- Data Cleansing Phase
- Training and Testing Data
- Models Applied for Predictions
- Ensembling Technique
- Result: Final Prediction

5.1.1 Dataset collection and description

For this research facebook check-ins dataset has been used, this dataset is having 6 attributes. Attribute 6 is set as target data and dataset is taken from Facebook for Kaggle Recruiting Event. Generally, numeric esteems are utilized as information. For the analysis dataset that is utilized as a part of this examination consider 1 by 1 km square lattice. This reduces dataset because the dataset is huge and the system can't handle too large dataset that's why small dataset has been used. The dataset that are used for research are described as follows:

- Record id: This feature specifies the number of records enter for any event in the facebook dataset.

- Cordinate_1: This feature about the first coordinate bounded between the given ranges.
- Cordinate_2: This feature about the second coordinate bounded between the given ranges
- Accuracy: This feature is about the accuracy on the basis of the check-in data is made. This component apparently speaks to the exactness of the coordinate_1 and coordinate_2 area dependant on natural elements, gadget attributes and so forth.
- Time: This attribute is said as being obscure by Facebook. After some investigation, and assume that from reference time, it is the passed away time in the form of minutes.
- Location_id: This is a remarkable identifier speaking to the place at which the registration was made.

5.1.2 Data Cleansing Phase

Data cleansing is most important string of data science. A genuine information researcher dependably discovers something out of the uproarious information by the specialty of information cleaning. . There are numerous procedures of doing data cleansing from which the propose system picked to get the clamor out of crude information utilizing Microsoft Excel. This information was in csv (comma separated values) design. So, there was a need of removing of the unwanted columns out of the raw data. For this correlation method has been used. Rather than looking at the independent correlations of each variable in the dataset, and look at the correlations of all variables together using 6 functions of the correlation in R returning a correlation matrix. This matrix shows the relationship of each attribute with all others. Correlation matrix provides an insight into the relationship strength between the attributes. Thumb is one way to determine the relationship which is as follows: The values in between 0 to 0.25 states weak relationship, 0.25 to 0.75 describe moderate relationship and values existing between 0.75 to 1.00 show strong relationships. Negative relationship is interpreted if the values contain negative symbol where one of the values increases and other falls, whereas, in case of direct or positive relationship both the values increases. If the variables have the value 1, it is called as perfect correlation. Table 5.1 shows a scale of possible correlations. Numeric values are a must for finding the correlations.

Table 5. 1 Correlations between Features

	row_id	X	y	accuracy	Time	place_id
row_id	1	0.972	0.694	0.972	0.0932	0.278
x	0.278	1	-0.726	0.158	0.12	0.178
y	-0.023	0.716	1	0.120	0.335	0.250
accuracy	-0.236	0.862	0.327	1	-0.273	0.030
time	0.866	0.987	-0.674	-0.347	1	-0.322
place_id	0.876	-0.075	-0.09	0.093	0.971	1

5.1.3 Training and Testing Data

This is the section of usage where the apportioning of this information into training and testing information. In this experimental analysis, [70,30] was the fraction and utilized for down to earth purposes since that is the standard dividing proportion.

[70, 30] means 70% of the informational collection is committed to training and 30% of information is devoted to testing the calculations in the event that they foresee the information being tried as precisely as could be expected under the circumstances, contrasted with their unique number. This expectation is done on the premise of training data information that have bolstered to the machine calculations.

5.1.4 Models Applied for Predictions

In this paragraph, Table 5.2 and Table 5.3 encapsulate the overall output of percentage split using multiple machine learning algorithms. All models have been built by using 70% of data for training and 30% of data for testing part. For a given dataset that have 6 attributes, 9768 rows. Various evaluation criteria like Correlation, R-Square, Root Mean Square Error (RMSE), Accuracy, values have been compared and presented in a tabular form given in Table 5.2 and Table 5.3.

Table 5. 2 Prediction of models based on r and r^2

S no.	Machine Learning Model	Method	r	r^2
1	Principal Component Analysis	pls	0.7	0.49
2	Least Angle Regression	lars	0.4	0.16
3	Bayesian Ridge Regression	bridge	0.79	0.62
4	Stochastic Gradient Boosting	gbm	-0.1	0
5	k-Nearest Neighbors	knn	0.78	0.61
6	Non-Negative Linear	nnls	0.5	0.25
7	Bayesian Regularized Neural Networks	brnn	0.71	0.5
8	Boosted Tree	blackboost	0.47	0.22
9	k-Nearest Neighbors	kknn	0.42	0.18
10	The Lasso	lasso	0.45	0.2
11	Conditional Inference Tree	ctree	0.25	0.06
12	Elastic Net	enet	0.51	0.26
13	Decision Tree	rpart	0.97	0.94
14	Support Vector Machines	svm	0.45	0.2
15	Multivariate Adaptive Regression Spline	earth	0.28	0.28

From Table 5.1 the correlation coefficient values for different models lies between 1 to -1.Histogram of comparison between correlation coefficient of different models has been shown in fig 5.1.

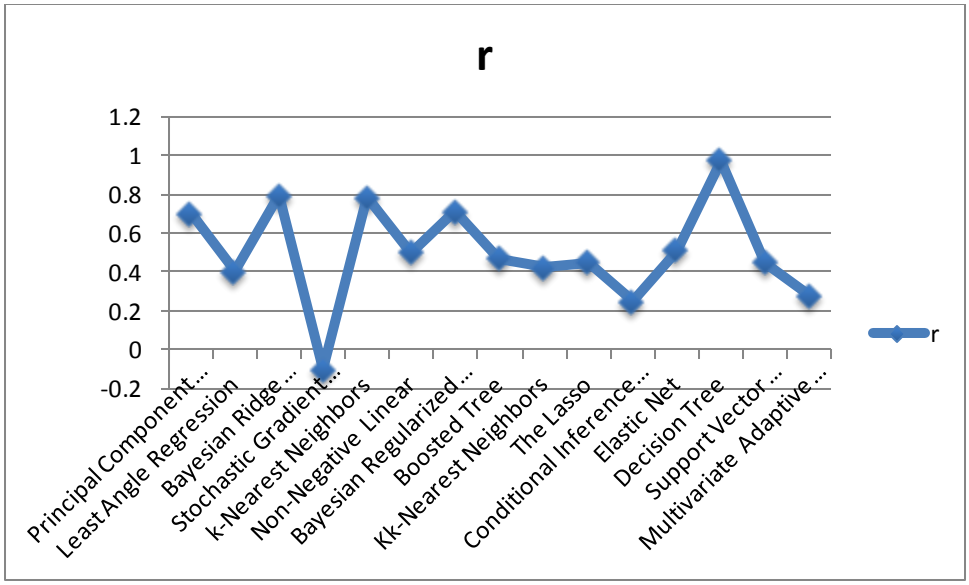


Fig. 5. 1 Comparison between correlation coefficient

Histogram of comparison between R-Squared values of different models has been shown in fig 5.2.The plot of different algorithm shows the values that has been taken from Table 5.1.

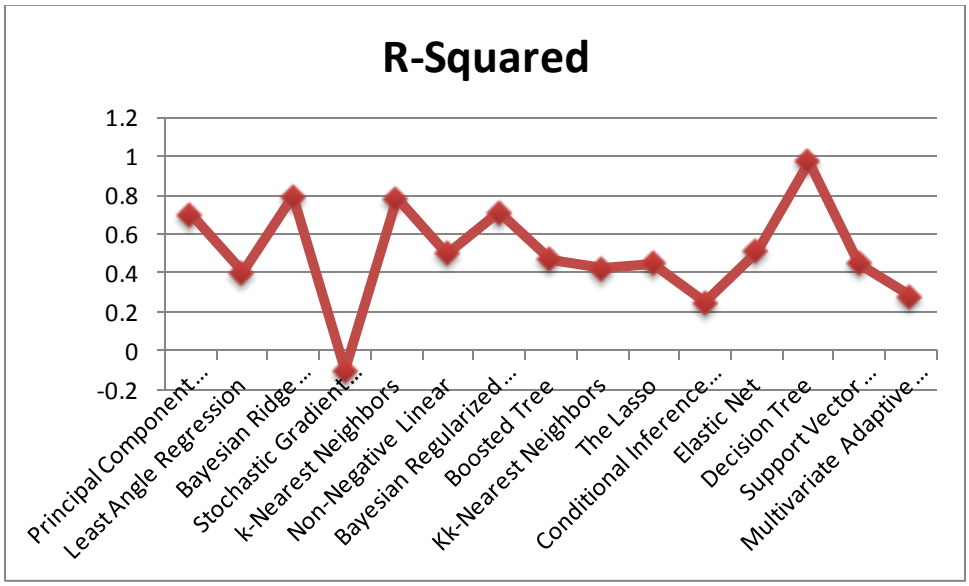


Fig. 5. 2 Comparison between r-squared

Table 5.3 shows that Root Mean Square Error and Accuracy values of the following models such as: Principal Component Analysis, Least Angle Regression, Bayesian Ridge Regression, Stochastic Gradient Boosting, k-Nearest Neighbors, Non-Negative Linear, Bayesian Regularized Neural Networks, Boosted Tree, Kk-Nearest Neighbors,

The Lasso, Conditional Inference Tree, Elastic Net, Decision Tree, Support Vector Machines and Multivariate Adaptive Regression Spline .

Table 5. 3 Prediction of models based on RMSE and Accuracy

S no.	Machine Learning Model	RMSE	Accuracy
1	Principal Component Analysis	0.09108	52.69
2	Least Angle Regression	0.12714	63.15
3	Bayesian Ridge Regression	0.07565	62.98
4	Stochastic Gradient Boosting	0.15175	53.56
5	k-Nearest Neighbors	0.07577	72.26
6	Non-Negative Linear	0.11939	66.38
7	Bayesian Regularized Neural Networks	0.08994	67.66
8	Boosted Tree	0.12214	48.25
9	Kk-Nearest Neighbors	3.8796	65.21
10	The Lasso	0.11671	63.64
11	Conditional Inference Tree	4.25646	75.84
12	Elastic Net	3.50721	62.14
13	Decision Tree	0.02626	65.41
14	Support Vector Machines	4.19288	77.03
15	Multivariate Adaptive Regression Spline	0.10913	41.24

The plot of comparison of models on the basis of Root mean square error values has been shown in Fig 5.3The accuracy plot shown in fig 5.4.this figure shows that Support Vector Machine model has highest accuracy of 77.03% .and and Multivariate Adaptive Regression Spline model has the worst accuracy of 41.24 %. Conditional Inference Tree has the second top highest accuracy and k-Nearest Neighbors model has the top third highest accuracy.

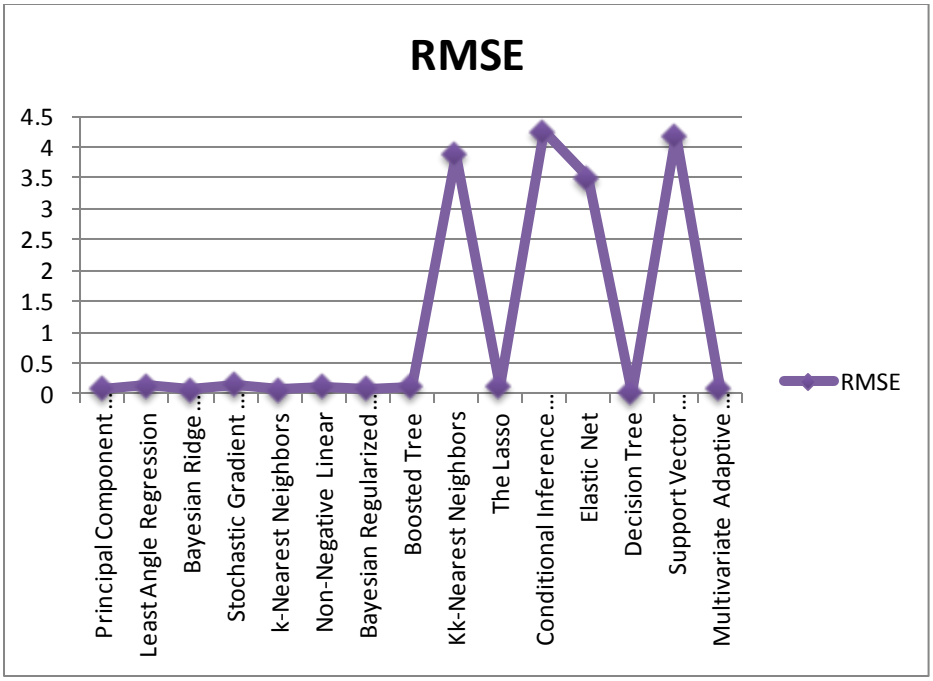


Fig. 5. 3 Comparison of Root Mean Square Error

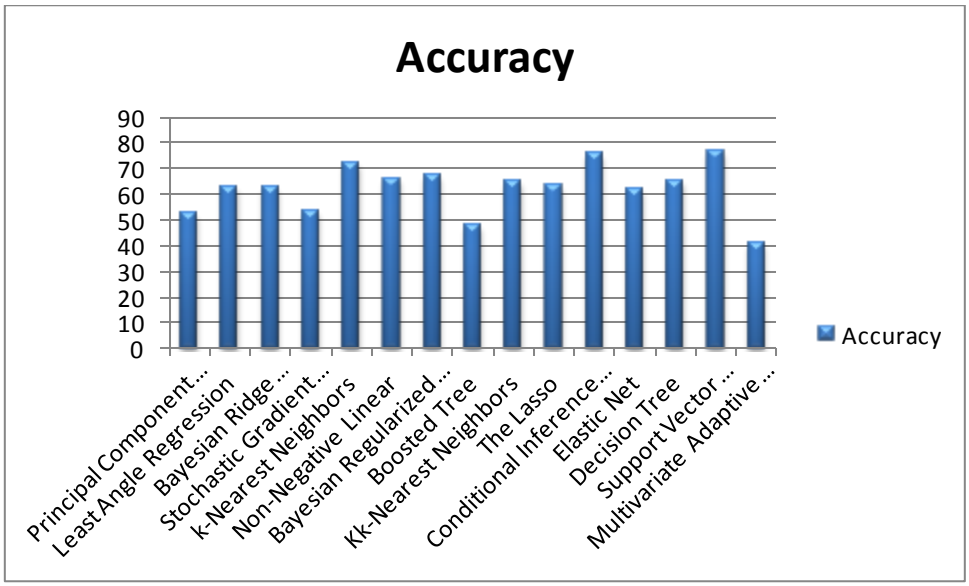


Fig. 5. 4 Comparison between Accuracy

5.1.5 Ensembling Technique and Result

It can require investment to discover well performing machine learning calculations for the dataset. This is a direct result of the experimentation way of connected machine learning. When a waitlist of exact models are held, then it can utilize calculation tuning to take full advantage of every calculation. Another approach that can use to build exactness for dataset is to consolidate the forecasts of various diverse

models together. In Table 5.4 top three models that attain highest accuracy value has been displayed with their method name, correlation coefficient value, coefficient of determination value, Root Mean Square Error value.

Table 5. 4 Top Three Models with Highest Accuracy

S no.	Machine Learning Model	Method	r	r-squared	RMSE	Accuracy
1	k-Nearest Neighbors	k-nn	0.78	0.61	7.577	72.26
2	Conditional Inference Tree	Ctree	0.25	0.06	4.25646	75.84
3	Support Vector Machines	Svm	0.45	0.20	4.19288	77.03

Table 5.5 represents result of ensemble model when the top 3 highest model has been gathered i.e. k-Nearest Neighbors, Conditional Inference Tree and Support Vector Machines model. The ensemble model has 82.12 % of accuracy.

Table 5. 5 Result of Ensemble Model

Machine Learning Model	r	r-square	RMSE	Accuracy
k-Nearest Neighbors, Conditional Inference Tree, Support Vector Machines	0.48	0.2304	3.9876	82.12

In the next step 10-fold cross validation method has been applied. Table 5.6 represents the result of 10-fold cross validation. In this k-fold method the proposed research take the value of k is 10. After 10-fold the highest value of accuracy that is achieved is 88.18%. The worst value of k-fold is 46.16%.

Table 5. 6 Representation of k-fold Cross Validation Result

Runs	Accuracy
1	54.38
2	87.57
3	88.18
4	46.16
5	78.42
6	54.58
7	51.73
8	87.38
9	54.44
10	73.95

5.1.6 Result: Final Prediction

Combination of k-Nearest Neighbors, Conditional Inference Tree and Support Vector Machines models are best for predicting most popular location in the given dataset. The plot of the accuracy of ensemble model has been shown in fig. 5.5.the plot show at run 3 it get highest accuracy of 88.18%.

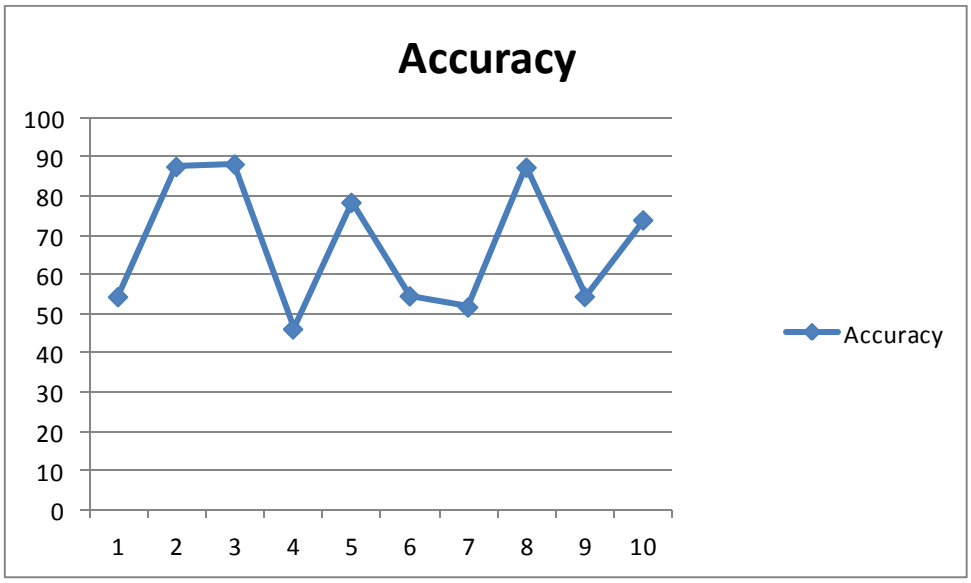


Fig. 5. 5 Accuracy of Ensemble Model

5.2 Measuring performance using Apache Mahout

5.2.1 Dataset Description

The dataset using for investigation of more than 100,000 spots situated in 10 by 10 km square matrix in the recreated world. The conspicuous explanation behind this dataset to be of significance is that the information utilized as a part of this dataset has been gathered more than two years and the example information speaks to the entire dataset which thus helps in building the model. Testing, as well as training information, are accessible in the .csv record with 6 qualities introduces the information. Table 5.7 describe the attribute description, feature name, number of unique values and number of missing values in details.

Table 5. 7 Dataset Description

S no.	Feature Name	Unique Values	Missing Values	Description
1	Row id	29118020	0	This feature indicates the number of record of the check in.
2	X	100000	0	The x-coordinate bounded between [0,10].
3	Y	100000	0	The y-coordinate bounded between [0,10].
4	Accuracy	1024	0	The accuracy with which the check-in was made. This presumably represents the accuracy of the (x,y) location dependent on environmental factors, device characteristics etc.
5	Time	786238	0	This feature is mentioned as being doubtful by Facebook. After some analysis, and surmise that it is the number of minutes passed since some reference time.
6	Place Id (Target Variable)	108389	0	This is a unique identifier representing the place at which the check-in was made.

5.2.2 Feature Selection Using Correlation Feature Selection Method

To extract the most important features from all dataset correlation feature selection method is used. After running the feature selection method Correlation between the features has been calculated and it has been shown in fig 5.6:

row_id	x	y	accuracy	time	place_id
1	-0.00077	0.002773	0.000441	0.00569	0.003108
-0.00077	1	-0.00712	-0.006306	-0.005048	-0.006677
0.002773	-0.00712	1	-0.007207	0.013127	-0.015594
0.000441	-0.006306	-0.007207	1	0.09455	0.004263
0.00569	-0.005048	0.013127	0.09455	1	-0.004348
0.003108	-0.006677	-0.015594	0.004263	-0.004348	1

Fig. 5. 6Correlation between all attributes

5.2.3 Training and testing

The model has been built by using 70% of data for training and 30% of data for testing. For dataset contains 29118020 total instances, and feature 6 set as a target data.

5.2.4 Predictions of Classification Parameters for Models

Confusion-matrix contains parameter such as True Positive, False Negative, False Positive ,True Negative, True Positive Rate and False Positive Rate .After that various evaluation parameter like Classification Precision, Recall, F-Square, Sensitivity, Specificity and Accuracy values have been compared and presented in a tabular form. Table 5.8 shows confusion matrix parameters such as True Positive, False Negative, False Positive, True Negative calculated values.

Table 5. 8 Confusion Matrix parameters

S.No.	Algorithm Name	TP	FN	FP	TN
1	Stochastic Gradient Descent	12410354	15279224	843218	1628224
2	Naïve Bayes	1234359	16279226	924218	1653268
3	Hidden Markov Model	15410355	15327922	689749	1667539
4	Random Forest	13410352	11279222	503238	1676590
5	Multilayer Perceptron	17410354	16279228	807618	1632151

The diagrammatical representation of confusion matrix parameter True Positive, False Positive, True Negative and False Negative values has been shown in Fig 5.7.

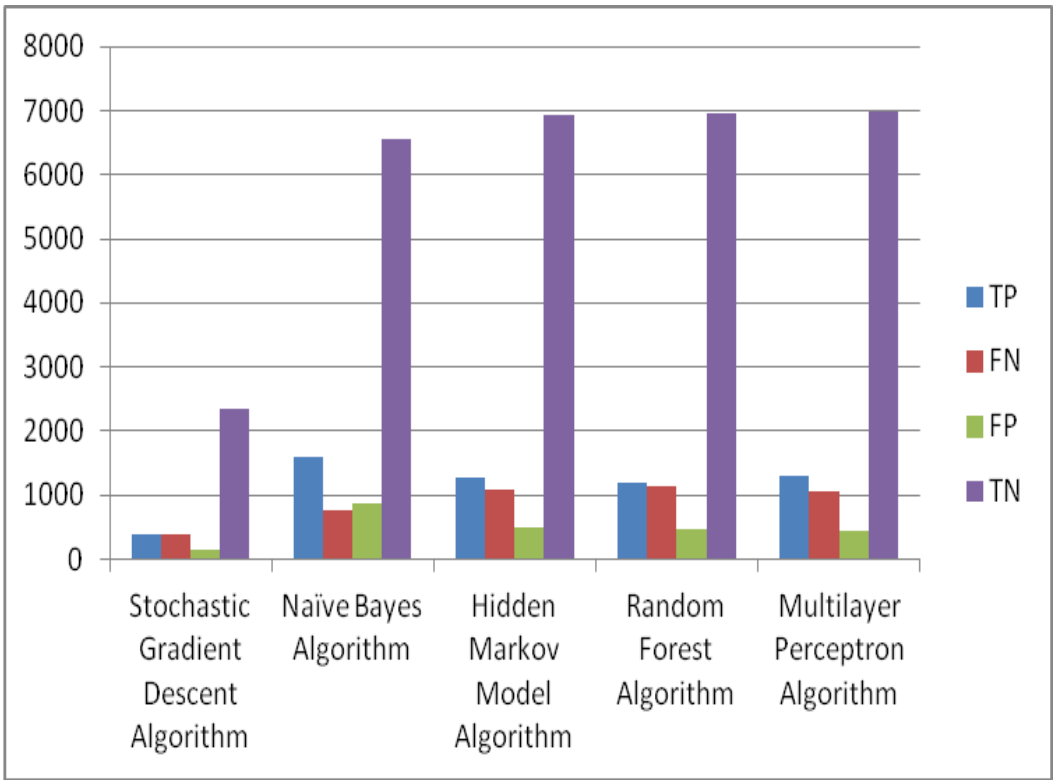


Fig. 5. 7 Confusion matrix parameter

The classification evaluation parameter True Positive Rate and False Positive Rate value of different algorithm has been shown in Table 5.9

Table 5. 9 TPR vs. FPR

S.No.	Algorithm Name	TPR	FPR
1	Stochastic Gradient Descent	0.4903	0.0057
2	Naïve Bayes	0.6774	0.1186
3	Hidden Markov Model	0.5418	0.0661
4	Random Forest	0.5127	0.0625
5	Multilayer Perceptron	0.5494	0.0608

Diagrammatical representation of Table 5.9 has been given in Fig 5.8

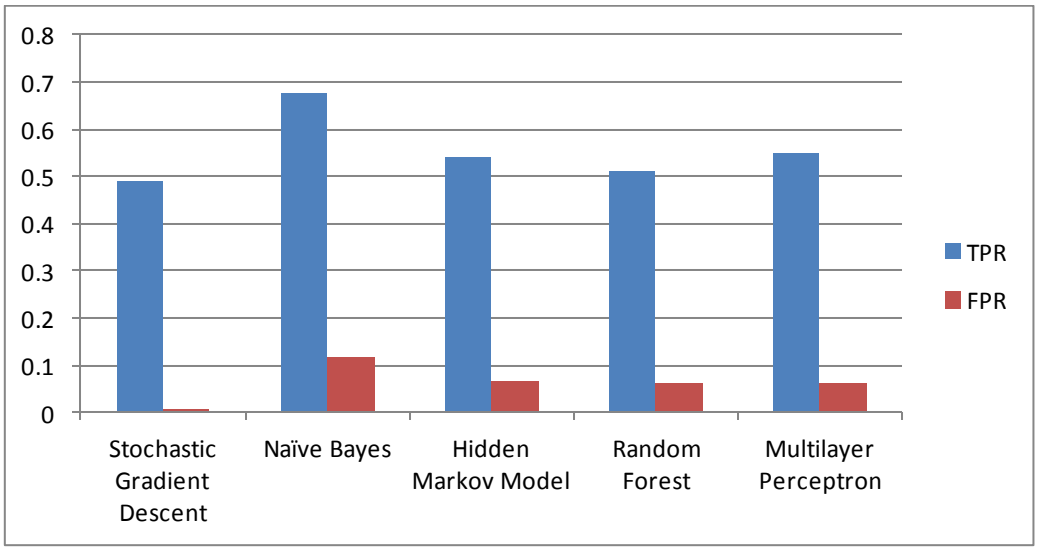


Fig. 5. 8 TPR vs. FPR

Classification evaluation parameters such as Sensitivity, Specificity and Accuracy calculated value for different algorithm has been shown in table 5.10.this table shows that the highest value of accuracy i.e. 84.6% achieved at Multilayer Perceptron algorithm and least value of accuracy i.e. 83.2% achieved at Naïve Bayes algorithm.

Table 5. 10 Evaluation Parameters Sensitivity, Specificity and Accuracy

S.No.	Algorithm Name	Sensitivity	Specificity	Accuracy
1	Stochastic Gradient Descent	0.4903	0.9943	0.834
2	Naïve Bayes	0.6774	0.8814	0.832
3	Hidden Markov Model	0.5418	0.9339	0.840
4	Random Forest	0.5127	0.9375	0.836
5	Multilayer Perceptron	0.5494	0.9392	0.846

On the basis of table 5.10 the graphical plot of classification parameter sensitivity, specificity and accuracy are given in fig 5.9

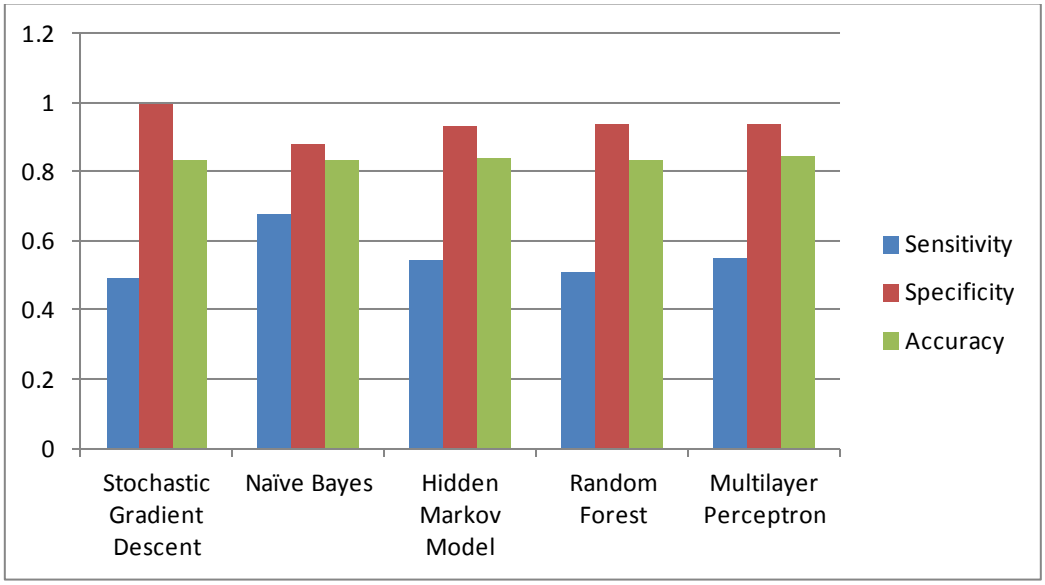


Fig. 5. 9 Classification Sensitivity, Specificity and Accuracy Plot

Classification evaluation parameters such as Precision and ROC calculated value for different algorithm has been shown in table 5.11.This table shows that the highest value of precision i.e. 0.7325 achieved at Multilayer Perceptron algorithm and least value of accuracy i.e. 0.6431 achieved at Naïve Bayes algorithm.

Table 5. 11 Evaluation parameter Precision and ROC

S.No.	Algorithm Name	Precision	ROC
1	Stochastic Gradient Descent	0.7270	0.86
2	Naïve Bayes	0.6431	0.57
3	Hidden Markov Model	0.7211	0.81
4	Random Forest	0.7214	0.82
5	Multilayer Perceptron	0.7325	0.90

On the basis of table 5.11 the graphical plot of classification parameter ROC and precision are given in fig 5.10

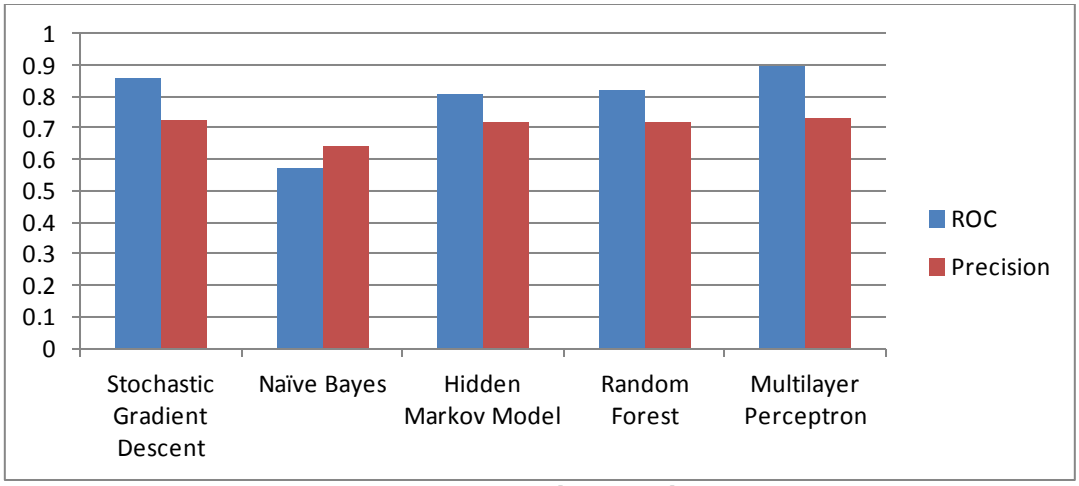


Fig. 5. 10 ROC and Precision Plot

5.2.5 Measuring performance of the best algorithms using the concept of Ensembling

Ensemble technique includes consolidating various multiple predictions determined by various learning algorithms so as to make a stronger overall prediction and get best outcomes. In this experiment approach the dataset that has been used on standalone machine also predicted on mahout environment.

Snapshot 5.1 represents the result of 9768 instances after applying Ensemble model. Total 9768 instances have been classified from which 8389 comes under the category of correctly classified instances and 1379 instances come under the category of incorrectly classified instances.

```
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances          :       8389          85.8824%
Incorrectly Classified Instances        :       1379          14.1176%
Total Classified Instances              :       9768

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       <--Classified as
1518    826       |   2344      a       =  A
553     6871      |   7424      b       =  B

=======================================================
Statistics
-------------------------------------------------------
Kappa                                           0.59
Accuracy                                        85.8824%
```

Snapshot 5. 1:Ensemble model Result of 9768 instances

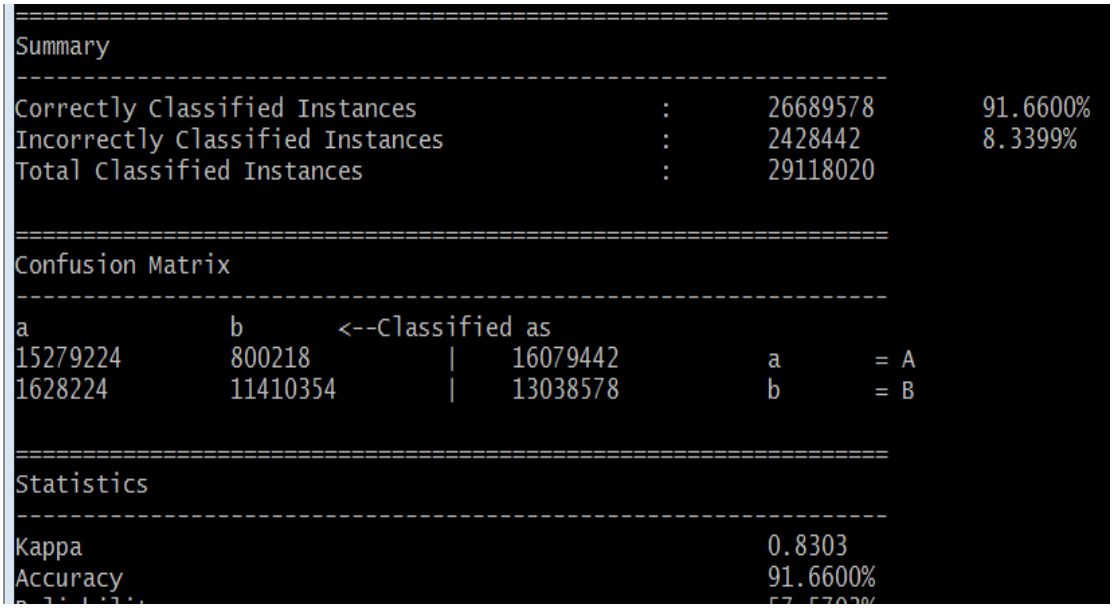
Following evaluation parameters has been evaluated. These parameters are represented in tabular form in table 5.12:

Table 5. 12 Parameters value of classification for 9768 instances

S.No.	Parameter	Results
1	TP	6871
2	TN	553
3	FP	826
4	FN	1518
5	TPR	0.892
6	FPR	0.3523
7	Precision	0.892
8	Recall	0.9255
9	F-Square	1.5092
10	Sensitivity	0.892
11	Specificity	0.892
12	ROC	2.530
13	Accuracy	85.8824%
14	Kappa	0.59

The experimental result of ensemble model for the dataset that has been used on apache mahout cloud environment for 29118020 instances has been evaluated.

Snapshot 5.2 represents the result of 29118020 instances after applying Ensemble model. Total 29118020 instances have been classified from which 2428442 comes under the category of incorrectly classified instances and 26689578 instances come under the category of correctly classified instances. Accuracy value of this huge dataset has been evaluating 91.66% and Kappa value for this dataset is 0.8303.



Snapshot 5. 2Ensembling Result of 29118020 instances

The remaining parameters of classification have been shown in Fig 5.13this table contains all parameters value of classification for 29118020 instances.

Table 5. 13Parameters value of classification for 29118020 instances

S.No.	Parameter	Results
1	TP	11410354
2	TN	15279224
3	FP	800218
4	FN	1628224
5	TPR	0.8751
6	FPR	0.0497
7	Precision	0.9344
8	Recall	0.8751
9	F-Square	1.4836
10	Sensitivity	0.8751
11	Specificity	0.9502
12	ROC	17.5845
13	Accuracy	91.6600%
14	Kappa	0.8303

5.3 Comparison between the results

In this section, results of the different platforms have been compared. When the standalone machine has been used to run R-Tool on 9768 instances then it gives the accuracy 77.03% on Support Vector Machine model but when ensemble the model of top three accuracies. It gives the result 82.12% of accuracy.

After that for a large dataset of 29118020 instances as well as chunk dataset of this huge dataset of 9768 instances run on Apache Mahout cloud environment then it gives the accuracy 84.6% on Multilayer Perceptron model. But ensemble model gives the surprise result as compared to predicting the result on the standalone machine using R-Tool. For 9768 instances ensemble model gives 85.88% accuracy which is greater than R-Tool ensemble model. Similarly for 29118020 instances ensemble model gives 91.66% of accuracy. All comparison Accuracy result for different instances in different platform shown in table 5.14:

Table 5. 14 Comparison between Accuracy for different instances in different platform

S.No.	Machine Platform	Instances	Accuracy Result
1	R-Tool on standalone machine	9768	82.12%
2	Apache Mahout on Cloud Environment	9768	84.6%
3	Apache Mahout on Cloud Environment	29118020	91.66%

The graphical histogram representation of above table has shown in fig. 5.11 from this figure accuracy curve shows that for large dataset the designed ensemble model gives the highest accuracy for the same dataset.

In Table 5.15 the dataset of 9768 and 29118020 instances parameters results has been compared, their confusion matrix values results such as True Positive, False Positive , True Negative, False Negative, True Positive Rate, False Positive Rate values has been compared.

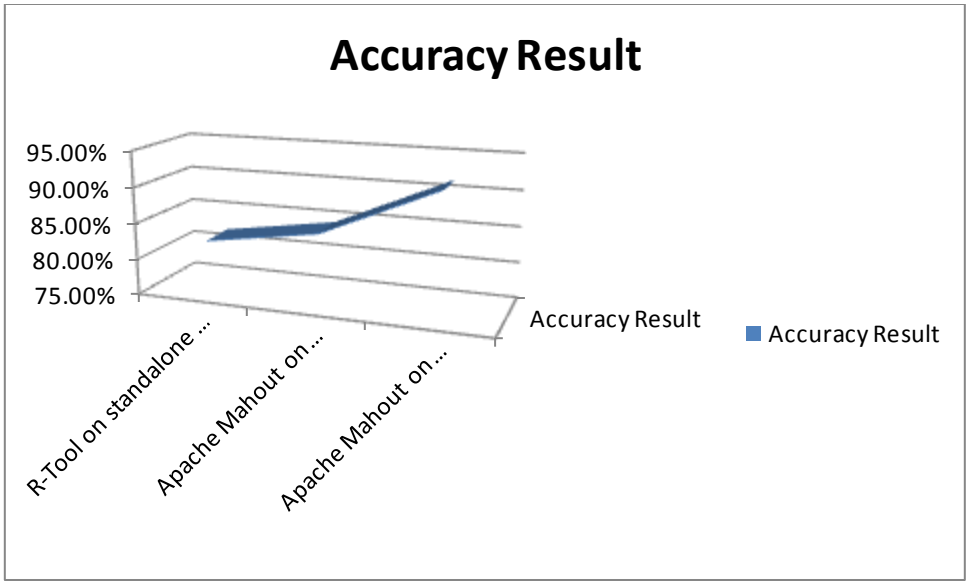


Fig. 5. 11Accuracy comparison on different platform

The classification essential parameters Precision, Recall F-Square, Sensitivity, Specificity, ROC, Kappa values has been compared in the table 5.15:

Table 5. 15 Comparison result between parameters for two different dataset

S.No.	Parameter	Result_1	Result_2
1	TP	6871	11410354
2	TN	553	15279224
3	FP	826	800218
4	FN	1518	1628224
5	TPR	0.892	0.8751
6	FPR	0.3523	0.0497
7	Precision	0.892	0.9344
8	Recall	0.9255	0.8751
9	F-Square	1.5092	1.4836
10	Sensitivity	0.892	0.8751
11	Specificity	0.892	0.9502
12	ROC	2.530	7.5845
13	Accuracy	85.8824%	91.6600%
14	Kappa	0.59	0.8303

On the basis of the classification evaluation parameter comparison between Result 1 and Result 2 their graphical representation shown in fig 5.12:

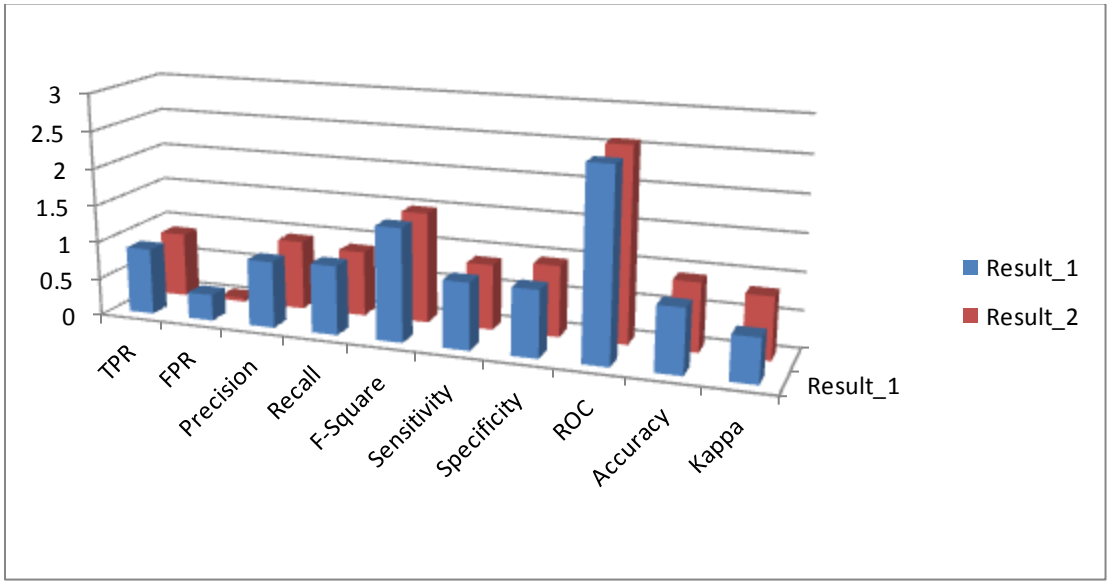


Fig. 5. 12Graphical repersentation of evaluation parameters

From this overall comparison, it has been concluded that when the new model design approach i.e. Ensembling has been applied on cloud environment for large dataset as well as small dataset it always gives well performance as compared the prediction result of standalone machine using R-Tool.

CHAPTER-6
CONCLUSION AND FUTURE SCOPE

This chapter states a brief description of the work done and list out the scope for further research under the future scope section.

6.1 Conclusion

The thesis work on Facebook check-ins dataset, this is a large dataset and it is very difficult to handle the given dataset on a standalone machine. To resolve this problem, the work is divided into two modules.

In the first module, the given dataset is taken in chunks from its original dataset. R programming tool has been used on this small dataset. The proposed model increases the prediction accuracy of location id in the Facebook check-ins dataset as compared to the existing work. The three models i.e. k-Nearest Neighbors, Conditional Inference Tree, Support Vector Machines has been used to create multilevel ensemble model. A novel multilevel ensemble model is developed for prediction and it produces high accuracy, correlation, R-square, root mean square error. The proposed model is compared with existing Facebook check-ins model and validated on a benchmark dataset. To check the robustness of proposed model, repeated k-fold cross-validation is used.

In the second module, to handle large dataset, Mahout Cloud environment has been used. In this case, a unique multilevel ensemble model is produced for the expectation of most prevalent places in Facebook and its length is imperative for training the model and it is proficient to utilize the variable length of the location. In this ensemble approach, three diverse machine learning model has been Ensembled to predict the variable length of the most well-known location. The proposed model of Facebook check-ins has accomplished highest of accuracy. In the current framework, the ensemble model is compared and it has been found that the proposed model produces better results. Thus, an endeavor has been settled on to propose an intelligent decision support model that can help Facebook specializes in foreseeing most mainstream area in view of the verifiable information of Facebook check-ins.

6.2 Future Scope

In the current work, only classification algorithms and their Ensembling model have been implemented however in future different machine learning models can be tried in the environment of the cloud. More important attributes can be added to the current dataset to expand the effectiveness of the proposed model. Real and huge Facebook most likely location information can be gathered from different live applications, and also all the previously mentioned strategies can be connected to it in the cloud environment for improving a model with more noteworthy accuracy. The proposed model can further be executed for finding the behavior of the Facebook user. Further research is necessary because of the ongoing diffusion and use of mobile devices by consumers in an increasingly globalized economy.

REFERENCES

[1.] Top 15 Most Popular Social Networking Sites and Apps [Online]. Available: https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/

[2.] T. Huang, H. Mi, C. Y. Lin, L. Zhao, L. L. Zhong, F. B.Liu & Z. X. Bian, "MOST: most-similar ligand based approach to target prediction." *BMC bioinformatics* 18, no. 1 (2017): 165.

[3.] Huang, L.Gallegos, & K. Lerman, "Travel analytics: Understanding how destination choice and business clusters are connected based on social media data." *Transportation Research Part C: Emerging Technologies* 77 (2017): 245-256.

[4.] N. Gupta, N. Ahuja, S. Malhotra, A. Bala, & G. Kaur, "Intelligent heart disease prediction in cloud environment through ensembling." *Expert Systems* (2017).

[5.] G. Manogaran, C. Thota, D. Lopez, V. Vijayakumar, K.M. Abbas, & R. Sundarsekar, "Big Data Knowledge System in Healthcare." In *Internet of Things and Big Data Technologies for Next Generation Healthcare*, pp. 133-157. Springer International Publishing, 2017.

[6.] M. Holub, "Twitter Data Mining", Aalborg University Copenhagen, 2016.

[7.] C. Yang & P. Srinivasan, "Life satisfaction and the pursuit of happiness on Twitter." PloS one 11, no. 3 (2016): e0150881.

[8.] J. Vernack "Predicting check-ins/location using Facebook data." PhD diss., GHENT UNIVERSITY, 2016.

[9.] S. C. R Gangireddy, "Supervised Learning for Multi-Domain Text Classification." (2016).

[10.] J. Lin, R. Oentaryo, E. P. Lim, C. Vu, A. Vu & A. Kwee, "Where is the Goldmine?: Finding Promising Business Locations through Facebook Data Analytics." In *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, pp. 93-102. ACM, 2016.

[11.] S. Kumar, D. Kulkarni, B. Padmanabha & N. Y. Murali, "Facebook: Check-Ins Prediction."(2015)

[12.] J. D. Zhang, C. Y. Chow & Y. Li, "iGeoRec: A personalized and efficient geographical location recommendation framework." *IEEE Transactions on Services Computing* 8, no. 5 (2015): 701-714.

[13.] P. Luarn, J. C. Yang & Y. P. Chiu, "Why people check in to social network sites." International Journal of Electronic Commerce 19, no. 4 (2015): 21-46.

[14.] S. L. Lo, R. Chiong & D. Cornforth, "Using support vector machine ensembles for target audience classification on Twitter." PloS one 10, no. 4 (2015): e0122855.

[15.] J. H. Vo, "Check-In Frequency with Friends on Location-Based Social Networks: A Look at Homophily and Relational Closeness." (2015).

[16.] Gupta, "Learning Apache Mahout Classification". Packt Publishing Ltd, 2015.

[17.] M. P. Kalghatgi, M. Ramannavar & N. S. Sidnal, "A neural network approach to personality prediction based on the big-five model." International Journal of Innovative Research in Advanced Engineering (IJIRAE) 2, no. 8 (2015): 56-63.

[18.] K. Huseynov, "Hadoop-based P2P botnet traffic classification using random forests." (2015).

[19.] B. Hao, L. Li, R. Gao, A. Li, & T. Zhu, "Sensing subjective well-being from social media." In International Conference on Active Media Technology, pp. 324-335. Springer International Publishing, 2014.

[20.] L. Wang, "Classification of clinical tweets using Apache Mahout". University of Missouri-Kansas City, 2014.

[21.] S. M. Nahiyan, "Use of Facebook to Boost Marketing." (2014).

[22.] J. Mahmud, J. Nichols & C. Drews, "Home location identification of twitter users." ACM Transactions on Intelligent Systems and Technology (TIST) 5, no. 3 (2014): 47.

[23.] M. Sethi, & S. G. Batra, "Sentiment Polarity Classification of Trendy Topics." PhD diss., 2014.

[24.] M. Dumoulin, "Personalized Large Scale Classification of Public Tenders on Hadoop." PhD diss., Université Laval, 2014.

[25.] T. Eriksson, "Automatic web page categorization using text classification methods", 2013

[26.] M. De Choudhury, M. Gamon, S. Counts & E. Horvitz, "Predicting Depression via Social Media." In ICWSM, p. 2. 2013.

[27.] J. B. Gomes, C. Phua, & S. Krishnaswamy, "Where will you go? Mobile data mining for next place prediction." In International Conference on Data Warehousing and Knowledge Discovery, pp. 146-158. Springer Berlin Heidelberg, 2013.

[28.] L. Safko, “The social media bible: tactics, tools, and strategies for business success”. John Wiley & Sons, 2010.

[29.] S. Asur, & B. A. Huberman, "Predicting the future with social media." In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, vol. 1, pp. 492-499. IEEE, 2010.

[30.] K. C. Lee, & H. Cho, "A general bayesian network-assisted ensemble system for context prediction: An emphasis on location prediction." In International Conference on Future Generation Information Technology, pp. 294-303. Springer Berlin Heidelberg, 2010.

[31.] S. S.Wang, S. I. Moon, K. H. Kwon, C. A. Evans, & M. A. Stefanone, "Face off: Implications of visual cues on initiating friendship on Facebook." Computers in Human Behavior 26, no. 2 (2010): 226-234.

[32.] D. Shrestha, "Text mining with Lucene and Hadoop: Document clustering with feature extraction." Wakhok University (2009).

LIST OF PUBLICATIONS

Conferences

[1.] S. Kashyap and M. Kaur, "Machine Prediction of Facebook Most-Likely Location Using Ensemble Approach", In: Proceeding of the International Conference on Current Research in Engineering and Technology (ICCRET), 2017.

[Accepted]

[2.] S. Kashyap and M. Kaur, "Ensembled Approach for Predicting Facebook Check-ins Using R-Programming", In: Proceeding of the ISRD-Recent Advances in Engineering and Technology Summit (RAETS) Conference, 2017.

[Accepted]

Journal

[1.] S. Kashyap and M. Kaur, "Intelligent Cloud Environment for Prediction of Facebook Most Popular Location through Multi-Model Ensembling Approach"

[Communicated]

VIDEO LINK

https://youtu.be/EtQ_uyAgBaE