Predicting the User Behavior Analysis using Machine Learning Algorithms

Ashwini[1], K Viswavardhan Reddy[2]

[1]*Digital Communication Engineering Dept. Telecommunication, RV College of Engineering, Bengaluru, India*
[2]*Associate Professor, Dept. Telecommunication, RV College of Engineering, Bengaluru, India*

Abstract - *Due to COVID-19, there is a rapid growth in the usage of internet in regards of reading news, communicating with friends, playing games, working from home and surfing etc. Hence, there is a need for understanding user's web browsing behavior by the above activities carried via web browsing. With this we can improve their browsing experience. So in this paper, we present various machine learning(ML) algorithms to predict and analyze the current user behaviors. The main objective of this work is discriminating and classifying the close group to which user is most interested. The event related to user's surfing that data is collected using browsing history application. Classification, helps us to identify similar groups of data that has been browsed based on parameters such as most visited, time duration, type of website that person has browsed. Algorithms such as k-Nearest neighbor (KNN), Naives Bayes (NB), Support Vector Machine (SVM), and K means clustering has been compared. From the results it is observed that naive Bayesian has predicted with good accuracy of about 90±4.*

Key Words: **Machine learning; K-Nearest neighbor; Naives Bayes; Support vector machine; User Behavior analysis.**

1. INTRODUCTION

Web browsing is defined as searching for useful information from the web and displaying it to users. There are different ways to obtain desired information from the server logs. Web browsing is the process of extracting useful information or particular data from server. In this browsing process, browser find outs the what users are looking for and most interested in searching on the Internet. From the survey we found , most of the authors have carried out their work using Machine learning techniques. This is because, these algorithms gave a good accuracy result when compared to other techniques like Artificial neural networks, Deep learning algorithms.

In [1] authors have implemented Machine Learning (ML) algorithms, which have demonstrated their effectiveness in clustering. The work has been carried out by collecting a real time datasets from the hospital. The result analysis shown that the ML algorithms are most effective and suitable in clustering the data as well as in predicting user behaviors. Authors have proposed a generative model for their experiment. Here the developed model is Multi-site Probabilistic Factorization (MPF) in [2] . The data sets were collected from six different popular video websites. This model has captured two different such as cross-site as and site-specific preferences. From the work MPF model achieved accuracy of 97± 2%. Understanding, the user behaviors and access pattern has got a great importance to different content provider. Researchers in [3] has carried out work by employing the unique data collected from an Internet service provider (ISP) service, and they systematically analyzed the user behaviors and viewing patterns across the six most popular content providers. The classification of data in [4] is done with the help of ML algorithms such as Logistic regression and decision tree classifier. The decision tree has produced an accuracy about 86%. So, from the results they have concluded that Decision tree is the best compared to Logistic regression.

Authors in [5] has experimented on Web log along with individual users collecting data from website and discussed behavior of user using Long short term model (LSTM). Parallel FP-Growth (PFP), Large page sets based parallel FP-Growth (LPS-PFP) and most interesting pattern-based parallel FP-growth (MIP-PFP) were compared, and MIP-PFP outperformed well compared to other algorithms in [7]. Authors in [8], carried out an experiment using Support vector machines (SVM), and the results obtained as prediction performance is better than Back Propagation Networks (BPN) and got average prediction accuracy about 80%.

There are several issues being faced by the researchers from their work such as the amount of data required for the work must very large for ML algorithms so that result will be more accurate in preprocessing and prediction. Here, Preprocessing included challenges such as handling of large data. Sometimes that cannot fit in to the memory. Several potential research challenges has been faced in working with ML/DL[1].

In this paper, we address the above issues discussed by collecting real time datasets. For this, we made an experimental set up with 20 computers in a lab and requested all the students to browse for an hour without any monitoring. Later these real time datasets were collected from browsed history of various websites. Figure 1 represents some of social networks which users interested in browsing. These datasets are given to developed machine learning algorithms. The ML algorithms chosen are K-NN, Naives Bayes classifier and

Support vector machine (SVM). In order to analyze and predict the future usage , the prediction of user behavior is done using Machine learning(ML) approach. In which a set of various features are extracted from datasets, then the model for prediction is developed. The model is trained based on 80% of training and 20% of testing 80% approach. These algorithms are most suitable and gave best result in the survey. The User Behavior Analysis (UBA) and prediction is estimated by invoking developed model in python programming. And these results are compared with all the thee algorithms.

Fig 1: Different social media network

2. USER BEHAVIOR ANALYSIS AND PREDICTION

From the past many years, analysis of user has been focused on the intense efforts in marketing applications, buying intention of some online buyers etc. Obviously, the objective of this work is to adopt efficient and some other new specific marketing strategies. And these strategies are based on real time datasets. That is recorded dataset information from the systems. Which includes the past/previous activities of that users or clients. So this can be defined as a data-based behavioral analysis , it because analysis done on recorded data information. This analysis has found its importance in detecting fraud information and fighting against fraud etc. So now, this is not that surprise to see behavior analysis can enhance information communication technology, detect internal threats like targeted attack, accelerate some repetitive tasks, adapt software's to the users so organize more efficiently production tools etc.

The user model is a representation of single user or it can be a group of multiple users in system. This developed model includes a set of data/ parameters that are representative of the user's previous behavior. The development of user model starts with system designing which will be collecting all the data information needed for representing the users. The real time data obtained from browsing tool can be used to deeply understand the behavior of an user [4]. The model development is done based on certain features and parameters which tells about user behavior, where the user is most interested in surfing the data. This browsed information can be obtained through various applications from the web. From these developed model result we can know user interest in advance and then it can easy to provide personalized services. Suppose if any information is missing , that can be easily retrieved. And future activities and behaviors can be predicted

Behavioral Analysis

User Behavior Analysis(UBA) is the disciplinary way of analyzing behavior of that user. In an operational way it can be defined as, essentially collecting data, monitoring the obtained data, processing that data for analyzing. The required data sets for work is collected from the users which is history of browsed data are stored in separate files, databases, directories or data log files etc. The purpose of this collection of datasets is a process to provide desired parameters and from this data it is very easy to build usable and reliable models the user. In other words, it will precisely classify the user group and accurately characterize the users. For example, nowadays the Internet surfing has become most privileged space for this type of application. Indeed, nowadays technologies are so grown up in every aspects i.e, in order to collect data and then exploit the present, past and future behavior of individual users. The three pillars of UBA are : Analysis of data, integration of data and representation of data. The most difficult challenge faced is in analyzing and processing the huge amount of data. The analysis of UBA in must be fast in preprocessing huge data of the users. And selected developed ML algorithms should be appropriate to classify the users. Therefore, Machine Learning algorithms must run in real time, so that it will be easily accessing to complete data sets.

3. DATASETS COLLECTION

The dataset collection is done by an anonymized viewing of website by that particular users. From search engine we will know the what has been browsed previously. There are many browsing history tools which collects browsed data automatically as shown in table 1. Among those some are open source and some are licensed version tools. Different tools have different features to collect data. So we found tool named browsing history view tool which is suitable for the work.

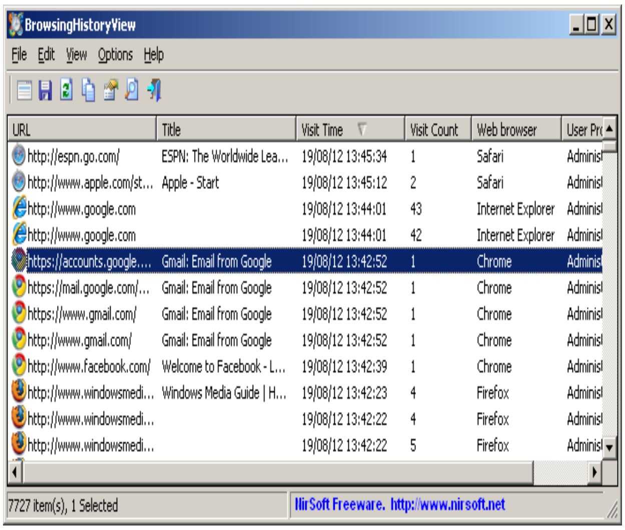
Table -1: Various Browsing tools

Sl no.	Tool	Availability	Advantages	Disadvantages
1	Time your web[10]	Free source	-Best for Google chrome -time bar graph	Does not support other websites
2	Activity watch[10]	Free source	-Available for Mozilla extensions	-Not good for Google chrome and other websites
3	Rescue Time[10]	Not free source	-Keeps complete	-It is not free source

			track of all online activities	
4	Browsing History View[9]	Free source	-Reads the history data from 4 different web browsing like Internet explorer, Firefox, Google chrome, Safar. -Requires less memory. -Displays history	-sometimes it bloat the 'Visit Count' number only for internet explorer.

Each log of the browsed data contains user information like user identity (ID), Uniform Resource Locator (URL) , Title, Visit Time, Visit Count, Web Browser, URL Length, Typed Count, History File, Duration, Record Id. Now by crawling and parsing the type of data has been browsed, we got information regarding URL, title, typed count and viewed website from the respective website browsing and content providers. Specifically, we have classified browsing into 5 different types. Such as social media browsing, educational purpose, shopping, news, entertainment and other purpose. The different content providers in web has their own way of naming conventions titles and other parameters. For example, in some of the logs we observed that at the beginning of data the content provider's name is embedded with the titles of the browsed websites. Then from these naming conventions, the manual modification is done to differentiate titles and website. By differentiating those parameter, mixing of parameters are avoided. Which later classify the data accurately and effectively. Figure 2 is captured view of how data is collected. The data is obtained is saved in excel or comma separated values (csv) file. This is done because it is easy to invoke these files later in programming.

The browsed data is collected at various time period i.e at morning, afternoon and night. This is done because browsing data varies from time to time. For example, some users are interested in news sites at the morning time, some user might browse study related sites and at the night time user might browse social media networks. The browsing data varies user from to user. So, for our work we have taken browsed data from morning to evening.



URL	Title	Visit Time	Visit Count	Web browser	User Pr
http://espn.go.com/	ESPN: The Worldwide Lea...	19/08/12 13:45:34	1	Safari	Adminis
http://www.apple.com/st...	Apple - Start	19/08/12 13:45:12	2	Safari	Adminis
http://www.google.com		19/08/12 13:44:01	43	Internet Explorer	Adminis
http://www.google.com		19/08/12 13:44:01	42	Internet Explorer	Adminis
https://accounts.google....	Gmail: Email from Google	19/08/12 13:42:52	1	Chrome	Adminis
https://mail.google.com/...	Gmail: Email from Google	19/08/12 13:42:52	1	Chrome	Adminis
https://www.gmail.com/	Gmail: Email from Google	19/08/12 13:42:52	1	Chrome	Adminis
http://www.gmail.com/	Gmail: Email from Google	19/08/12 13:42:52	1	Chrome	Adminis
http://www.facebook.com/	Welcome to Facebook - L...	19/08/12 13:42:39	1	Chrome	Adminis
http://www.windowsmedi...	Windows Media Guide | H...	19/08/12 13:42:23	4	Firefox	Adminis
http://www.windowsmedi...		19/08/12 13:42:22	4	Firefox	Adminis
http://www.windowsmedi...		19/08/12 13:42:22	5	Firefox	Adminis

7727 item(s), 1 Selected

NirSoft Freeware. http://www.nirsoft.net

Fig 2: Data Collection View

4. METHODOLOGY

Figure 3. shows complete process involved in classifying and analysing the behaviour of the user.

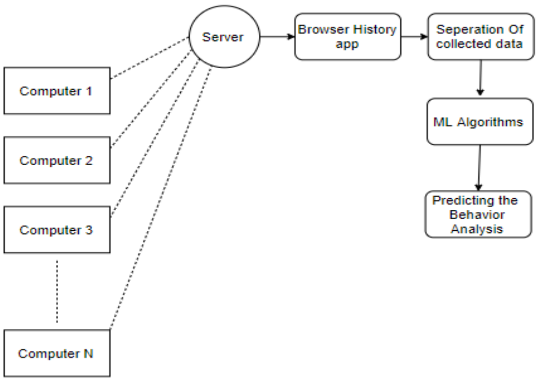


Fig 3. Process of classifying and analyzing UBA

General steps for Predicting user behavior based on web browsing history are

Step1: Arrange and setup nearly 5-10 in number of computers in the department. And these should be with the same configurations like hard disk storage(RAM, ROM), Power, Speed of the device etc.

Step 2: Then next step is to download and install web browsing history applications in each of the computers i.e Browsing History View.

Step 3: Now users are allowed freely for browsing data of their own interest, the browsing is done for certain time duration. Next step is to collect the browsed data from the application installed in the system

Step 4: The collected browsed data is given to machine learning algorithms. The algorithms do the classification based on similar data browsed. And those are grouped and classified.

Step 5: Thus, all features are classified and then we can predict the user behavior analysis and compare with ML algorithms.

A. Machine Learning Algorithms For Uba

Many of the ML approaches has shown the promising results required for the predicting user behavior on website. From the high-level model of information one can create direction for modeling for predicting user behavior. A high-level statistics cab be built on how exactly data propagates over time. In this visit time i.e duration spent and number of time visits made to that browser became useful method in understanding the UBA in Web. And also by taking various other features from obtained data we will get more detailed view of user behavior[6]. The feature parametric based approach for developing a model will provide a more precise model with good accuracy.

The problem formulation in predicting behavior of user in ML terminology cab be naturally performed in a straightforward way[6]. Here, the problem can be formulated as a classification task. The main goal is to predict the outcome results of user. For new sets of `test' samples one must build a predictive model M. From developed model we will get accuracy for that model. Some of classifier models are developed for our work which gave satisfactory results.

(i) K nearest Neighbor (k-NN)

The main idea of this K-NN techniques is to choose k neighboring vectors for all the input vectors. Let us consider x as an input vector. For selecting neighboring vectors of the input vectors is taken as distance metrics between the various data points of input vectors. Here for finding minimum distance , Euclidean distance calculating method is applied. Now next step is to find the similarity measure and then compare all the vectors. From the results of similarity measure we obtain the nearest K neighbors. In case of calculating the Euclidean space distance, the similarity measure i.e, S(p, q) between two vectors p and q are to be considered. Now equation 1 shows Euclidean distance :

$$s(p,q) = \sqrt{}\, \omega i(p - qi)2 \quad ....(1)$$

where p ,q be the two vectors, pi and qi are the ith entries of the all the input vectors. And 'n' is the total number of entries of data points. $\omega = \omega i$ is weights of all the vector. Which are correspondence to the importance in predicting the all vectors. In n-dimensional Euclidean space vectors pi is a point in this vector space. Now vector Y can be considered as predicting variable for prediction. Where Y is examined by majority group of kNN to this n-dimensional point. For example, if we got k = 5. And among five clusters, three are the nearest neighbors are representing the symbols of a class. which are labeled as `crosses', or `1', whether two others are of the class which are labeled as `0', or `circles', then variable Y will be equal to one. So, the class `crosses' is having the majority of neighbors. Figure 4 shows classification using k-NN approach.

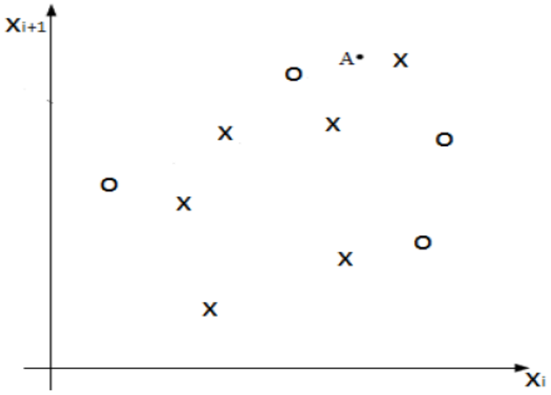


Fig 4: Classification using a k-NN approach

(ii) Naives Bayesian theorem

The NaivesBayes' theorem briefly explains about the probability. That is, probability of happening an incident or event taking the of previous information which is related to this event/ incident[7]. For instance, network traffic information related to the attack can be known with DoS attack information. Therefore, comparing with the network traffic and assessing this without the knowing the past network traffic information we can evaluate traffic of the network probability using Bayes' theorem. A common and efficient Machine Learning (ML) algorithm based on probability calculation is a Naive Bayes (NB) classifier. This NB classifier does the classification by estimating probability of the datasets. NB is a commonly known and used as a supervised classifier. This NB classifiers calculates posterior probability for the given data and then uses the Bayes' theorem to forecast's that the probability. Which does the feature sets of not labeled examples of NB classifier those examples fits a specific label of NB classifiers. Now considering an intrusion detection as example, NB is used as a classifier to classify this traffic as abnormal or normal. The advantages of NB classifiers are like ease of implementation, simplicity, applicability to binary and multi-class classification, robustness to irrelevant features and requirement of low training.

Let us take an example, which will classify X as a test instance, probabilistic model can be formulated with the approach by considering and calculating the posterior probability as $(p \mid w)$ for different (w) and later the largest posterior probability for the given data is predicted. From the rule of maximum a posteriori (MAP), we have Bayes Theorem equation 2 as

$$p(w \mid x) = \frac{p(\ W \mid X\ )\, p(w)}{p(\ x\ )} \quad ....(\ 2\ )$$

where $p(w)$ is calculations estimated by considering and counting all the proportion of class w data points in the training set. Then $p(x)$ can be ignored. Now we are comparing different 'w' on the same data points. So we need to consider $p(w \mid x)$. From this, suppose if we get an accurate estimate of $p(x \mid w)$ from the given training data it will be the best classifier in theory. Therefore, we get the resulting Bayes as a optimal classifier with the smallest error rate in calculation. However $p(x \mid w)$ estimation is not a straightforward thing. Therefore it is involved with the estimation of exponential numbers of joint-probabilities of that features. To make the good estimation some of assumptions are made. So here NB classifier assumes that given the class label will have all the 'n' features which are independent with one another within same class. Thus, we have equation 3 as,

$$p(x \mid w) = \prod_{i=1}^{n} p(xi \mid w) \text{ ....(3)}$$

Now we need to calculate probability for estimating every feature value in each of the class so that the conditional probability can be estimated. And after estimating each feature so that the calculation of joint-probabilities cab be avoided. Now in the training set, the naive Bayes classifier calculates estimates the probabilities of $p(w)$ for all classes. For which w € $\mathbf{w}$ and $p(xi \mid w)$ for all features of the class. That is i can be i = 1,2,3,......,n from the training set *xi.* In the test set, test instance which are labeled with *w* will be predicted and the condition it lays is, it is predicted only if *w* leads to largest value of all labels of class.

$$p(w \mid x) \text{ ά } p(w) \prod_{i=1}^{n} p(xi \mid w) \dots \dots (4)$$

As from ML, the NB techniques provides a very simple approach, with clear semantics, also for using, for representing, and for knowledge of learning probability. This NB classifier goal is to accurately predict the class of test instances.

(iii) Support vector machine (SVM)

Support vector machine is an non-probabilistic classifier. The assigning of labels for prediction is done with SVM model. Where it predicts into one or other category. The SVM builds a boundary between many of data points in n-dimensional space. Here n represents the total number of features. Considering a SVM classifier, the boundary between any of the two classes can be considered only after training this SVM classifier. Here two different classes are labeled as crosses and circles. Suppose for the new 'n' dimensional point if points lies above boundary line then it is classified and labeled as 'cross' and a 'circles' if not[8]. From the figure 5 we see that points A, B and C are classified as `crosses'. This is because they are above the boundary level.

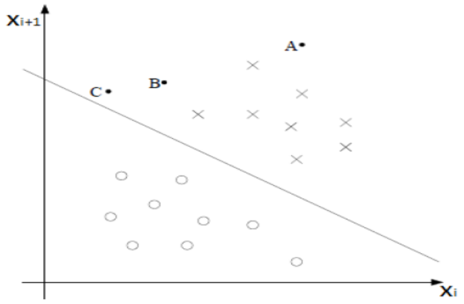


Fig 5: Classification using an SVM approach.

Analogously to other ML approaches, an implementation of SVM exists as most common ML technique in achieving the better accuracy of prediction.

(iv) k-clustering

k-clustering is an unsupervised method of ML approach in clustering the data. This clustering aims to discover k number of clusters from the input data. Where k refers to total number of clusters that need to be generated by algorithm for that input datasets. This method of clustering is basically calculated and implemented by iteratively allocating each data point to the various clusters. These data points are allocated to one of k clusters of total clusters according to the and minimum distance between the points and also based on various features. To get the ultimate result of clustering repeatedly trained and tested. The inputs of the algorithm are only the datasets and the k clusters. Firstly, the k centroids are estimated by Within cluster sum of squares (wcss) or other method. And then each of the sample in data is assigned to its one of its closest cluster. This assigning is done estimating centroids. Which is calculated using squared Euclidean distance between the all the points. Secondly, once all the data points of the samples are assigned to a specified cluster, and then again centroids are recalculated by taking mean of all sample values of cluster. The algorithm iteration continued to iterate until no sample of the data is left. The performance and accuracy of clustering is less precise than those of the other supervised leaning methods. In generating a labeled data it is difficult to generate. So, in this case unsupervised algorithms are good. The Unsupervised ML methods have many applications in security domain.

5. EXPERIMENTAL RESULTS

The ML algorithms have been implemented considering some of the features. Firstly, the real time datasets which is needed to our work is collected from user personal computers are installed with browser history tool. Now datasets are rearranged and modified. These datasets are divided as testing and training sets (20-80%) approach. The 80% training is train the developed ML algorithm and 20% testing is to test output results once classification is done

A. k-NN clustering

In this clustering method the data sample need to be clustered in different k-cluster. The value of k can be estimated by WCSS method. This estimation calculation is done using Euclidian distance this of sample. Figure 6 shows the graphical relationship between within cluster sum of squares (WCSS) and number of clusters. Then select the number of clusters where the change in WCSS begins to level off (elbow method).This is resultant graph obtained for our datasets.

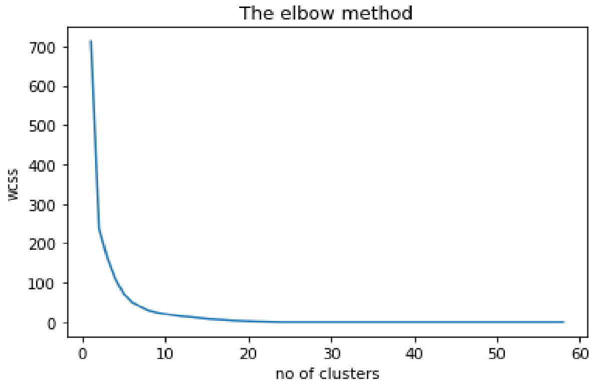


Fig 6: WCSS graph for finding K

After estimating value of k, from graph, k =5. So, number of clusters will be 5. Now clustered result is shown in figure 7

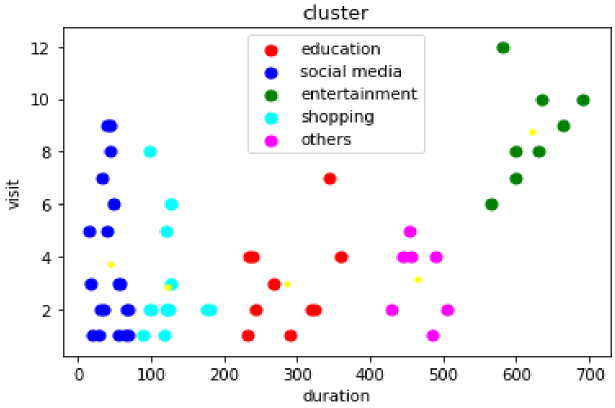


Fig 7: k-Clustered results

From the graph and result we can say that that user most interested in browsing data which is related to entertaining sites. This is because number of time visit and duration of time he spent is more compared to other clusters.

B. Naives Classifier

In this NB classifier algorithm datasets are trained and tested at different at different frequencies. The developed NB classifier algorithms results are tabulated as table 2.These are the accuracy performance results at various training set approach. The classifier got good accuracy at 75-25% training algorithm i.e about 93.33%.

C. KNN

Second algorithm chosen for classification is kNN algorithm, for this also same method is applied for training and testing. The accuracy result obtained are tabulated in table 2.

In this algorithm got best result at 80-20% training approach about 58% of accuracy.

D. SVM

The last algorithm chosen from the survey is SVM. So for this algorithm same kind of methodology is followed. The result got from this algorithm is tabulated in table 2.

In this algorithm we got best result at 70-30% training approach about 55% of accuracy.

Table 2: Comparison table

Algorithm	80-20%	75-25%	70-30%	60-40%
NB Classifier	93.33 %	91.66%	75.11%	72.22%
KNN	58%	53%	50%	31%
SVM	50%	53%	55%	33.3%

From comparison table 2 we can say best ML for our dataset is Naives Bayes classifier. This is because it is giving accuracy about 93.33%

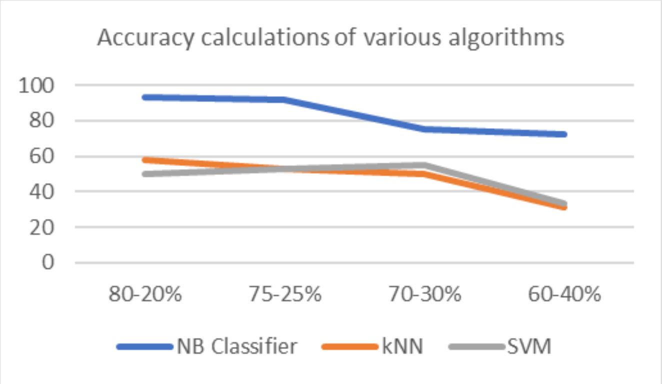


Fig 8 : Accuracy calculations of various algorithms

From the figure 8, it is observed that for all the percentage of data sets, NB is giving the best performance when compared with the other algorithms.

6. CONCLUSIONS

In this work, with user browsed data are obtained from browsing history viewing tool. Here we model browsed data in reference to social network, entertainment platform, others etc. Through real data analysis, we observe user where that person is most interested. So we developed some of ML algorithm to classify data. The Proposed models are used analyze and predict the UB from the dataset and then calculate the accuracy of each algorithm developed. The algorithm chosen here are KNN, NB classifier, SVM and for clustering the data k means is used.

Among these developed algorithms, we got good result for NB classifier. It gave an accuracy about 93%. From this we can conclude that Naives Bayes Classifier is the best among all other algorithms developed and we got satisfactory results for this work carried out.

In future we will be trying to predict and analyze by considering other parameter for the work. And we will try to compare with other algorithms too. In future other algorithms may give better accuracy performance than NB Classifier.

REFERENCES

[1]. M. Callara and P. Wira, "User Behavior Analysis with Machine Learning Techniques in Cloud Computing Architectures," in the preceeding of International Conference on Applied Smart Systems (ICASS), Medea, Algeria, 2018, pp. 1-6.

[2] H. Yan, C. Yang, D. Yu, Y. Li, D. Jin and D. Chiu, "Multi-site User Behavior Modeling and Its Application in Video Recommendation," in IEEE Transactions on Knowledge and Data Engineering.

[3].Ladekar, Ashwini, Pooja Pawar, Dhanashree Raikar and Jayashree Chaudhari. "Web Log based Analysis of User's Browsing Behavior." (2017).

[4] R., Virendra & V., Govind, "Prediction of User Behavior using Web log in Web Usage Mining" in the proceedings of International Journal of Computer Applications. 139. 4-7. 10.5120/ijca2016909228.

[5] V. Anitha and P. Isakki, "A survey on predicting user behavior based on web server log files in a web usage mining," in the proceeding International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, 2016, pp. 1-4.

[6] Sisodia, D. S., Khandal, V., & Singhal, R. (2018). Fast prediction of web user browsing behaviours using most interesting patterns. Journal of Information Science, 44(1), 74– 90.

[7] Liu, Qingchao & Lu, Jian & Chen, Shuyan & Zhao, Kangjia. (2014). Multiple Naïve Bayes Classifiers Ensemble for Traffic Incident Detection. Mathematical Problems in Engineering. 2014. 1-16. 10.1155/2014/383671.

[8] N. K. Gyamfi and J. Abdulai, "Bank Fraud Detection Using Support Vector Machine," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, 2018, pp. 37-41, doi: 10.1109/IEMCON.2018.8614994.

[9]. https://en.wikipedia.org/wiki/Web_browsing_history.

[10].https://hetmanrecovery.com/web-browser-history-viewer-software.htm.