A Novel Ensemble based Approach for Predicting Internet usages using Data Mining

Mahima Goyal[1], Vivek Gupta[2], Surabhi Kataria[3]
[1]*Department of Computer Science and Engineering,*
[23]*Assisant Professor, Department of Computer Science and Engineering,*
Rajasthan Technical University, Kota, Rajasthan, India

Abstract- In the present scenario, the internet has a significant impact on every facet of operations in various sectors. The adoption process of the Internet becomes a key challenge because many information systems such as Decision Support Systems, are rapidly moving to the Internet platform and facilitating remote access and group cooperation. In our paper, we studied important working environmental factors those affect the usage of the Internet. Also, facilitating constraints and social factors are taken into account in predicting the use of the Internet. Therefore, a more comprehensive framework is designed by data mining process using a hybrid approach. A java based tool called WEKA is used for performing experiments. The results show that the proposed framework effectively increase the prediction accuracy for Internet usage.

I. INTRODUCTION

The Internet plays an important role in every aspect of human life. Its usage is extended to numerous fields, including education [1], healthcare, business, and much more. Individuals may access the Internet through numerous devices like desktops, tablets, smartphones, among others. However, the number of individuals accessing the Internet through their smartphones is increasing rapidly. The usage of mobile internet is increasing day by day [2]**.**

This brings along a myriad of problems ranging from data storage, data management to data security. Despite the vast number of networks involved, the Internet does not have a central authority. There is no control on how individuals use the Internet or censorship on the information placed on the network.

Businesses and leading industries are viewing these huge data repositories as a tool to design future strategies, prediction models by analyzing patterns and gaining knowledge from this unstructured data by applying different data mining techniques. Internet service providers will be benefited by such studies of internet usage prediction. Apart from this plenty of other datasets like medical datasets can be used for early disease prediction and serve the humanity. Therefore, this learning will help valued understanding of historical data and generation of revolutionary models for meaningful classifications. It is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends in data mining, association rules are created by analyzing data for frequent if/then patterns, then using the support and confidence criteria to locate the most important relationships within the data.

Support is how frequently the items appear in the database, while confidence is the number of times if/then statements are accurate [3]**.**

In this paper, a comparative study will be done on the dataset of the people using the Internet. By doing the analysis of the dataset and applying the algorithms of machine learning we can predict the internet usage of people and anticipate the future internet demands. The data contains general demographic information on internet users and comes from a survey conducted by the Graphics and Visualization Unit at Georgia Tech. [4]**.** The colossal dataset consists of about 72 different attributes with thousands of entries. In this project, we will deeply analyze all the attributes and reduce the dataset desirably to predict the internet usage [5]**.**

The number of internet users worldwide has proliferated over the years. Internet users are defined as persons who accessed the Internet in the last 12 months from any device, including mobile phones. As of June 2017, 51 \% of the world's population has internet access. In 2015, the International Telecommunication Union estimated about 3.2 billion people, or almost half of the world's population, would be online by the end of the year. Of them, about 2 billion would be from developing countries, including 89 million from least developed countries [6]**.**

I. PROPOSED MODEL

The proposed framework aims the classification of Internet usage data to design the best classifier for data usage compared to existing work with a better accuracy. Firstly, we have classified the dataset and then analyzed the algorithm best suited for prediction of Internet Usage. For the experiment and prediction purpose various attributes of social behaviour, have been identified which affects the usage of internet [7]**.** The term Knowledge Discovery in Databases, or KDD for short, refers to the broad process of finding knowledge in data and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition,

databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases The KDD process includes a selection of relevant data; it's processing, transforming processed data into valid information and then extracting hidden information/pattern from it [8]. The KDD process can be categorized as under:

Selection

It includes selecting data relevant to the task of analysis from the database.

A. Pre-Processing

In this phase, we remove noise and inconsistency found in data and combine multiple data sources.

B. Transformation

In this phase transformation of data takes place into appropriate forms to perform mining operations.

C. Data Mining

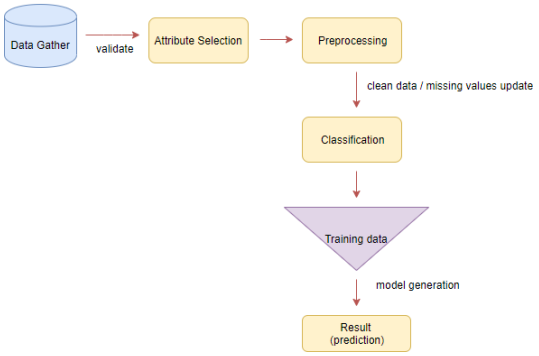
This phase includes applying a data mining algorithm appropriate for extracting patterns.

D. Interpretation/Evaluation

Interpretation/Evaluation includes finding the relevant patterns of information hidden in the data [9]

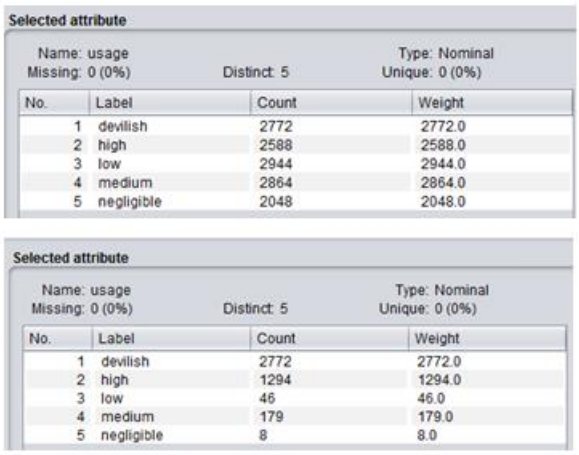
II. METHODOLOGY

Pre-processing & Attribute selection: Pre- processing is an essential part of data mining must. Firstly, the collected data is converted into the .arrf format} (uniform Format) and refined the data by identifying the missing values. Further, the collected data has various attributes but only a few attributes affect the final outcome. Therefore, the attribute selection technique is applied for reducing the attribute dimension and select the best-fitted attributes for making a prediction.



Classification: There are various classifiers available those predict nominal or numeric quantities such as decision-trees, support vector machines, instance-based classifiers, logistic regression and Bayes' nets. Each classifier provides an accuracy while applying on the dataset which can be evaluated in terms of Confusion Matrix. In our work, we merged multiple classifiers to increase the accuracy on the dataset.

Visualisation: Finally, the experimental results are graphically visualized in the 2D representation of data that easily interpret the solutions for the learning problem.



Selected attribute

Name: usage Missing: 0 (0%) Distinct: 5 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	devilish	2772	2772.0
2	high	2588	2588.0
3	low	2944	2944.0
4	medium	2864	2864.0
5	negligible	2048	2048.0

Selected attribute

Name: usage Missing: 0 (0%) Distinct: 5 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	devilish	2772	2772.0
2	high	1294	1294.0
3	low	46	46.0
4	medium	179	179.0
5	negligible	8	8.0

III. EXPERIMENTAL SETUP

For simulating the experiments in the Weka Toolkit, the collected data in the .arrf format is uploaded. Further, the attributes selection in done into two parts: Attribute Evaluator and Search Method. The attribute evaluator is the technique by which each attribute in the dataset is evaluated in the context of the output variable. Whereas the search method is the technique by which we try or navigate different combinations of attributes in the dataset in order to arrive on a short list of chosen features.

Here, the Correlation Attribute Eval technique is used with a Ranker Search Method.

In the Ranker Method, attributes are ranked by their individual evaluations. There are total 71 attributes are present in the dataset, out of the 23 attributes was selected to perform further operations. Classification techniques are supervised learning methods that are most widely used in data mining to classify the data in raw data, a value is assigned to each item in the set of data to group them in classes, these classifier models are mathematical techniques which are used to classify the data. Classification techniques such as decision tree, Bayesian classification, neural networks, support vector machine, association-based classification etc. are used for data mining. In our experiment following classifiers are evaluated on the dataset.

1) J48: It is a simple decision tree algorithm used to classify data. J48 is a supervised learning methodology for classification purpose. It is based on the divide and conquers approach. It divides the whole data into a subrange, which is based on present attribute values for the values that are already available in the sample training dataset.

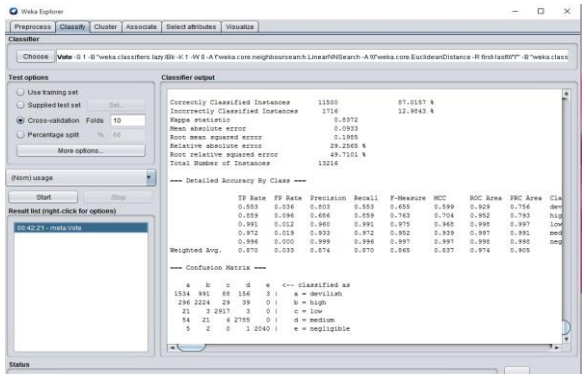
2) Random Tree: In mathematics and computer science, a random tree is a tree or a directed-graph form of a rooted tree

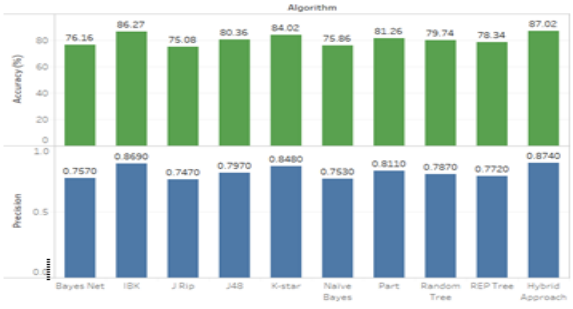
that is formed by a random process.

3) Naïve Bayes Classifier: This is quite helpful in the analysis of a large dataset which applied the probabilistic approach to classify the data according to their attributes.

4) IBK:} This algorithm, implements the k- nearest neighbour algorithm. In pattern recognition, the k-nearest neighbour's algorithm is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space [10]**.** The output depends on whether k-NN is used for classification or regression.

{5) KNN: It is used to classify the unknown instances which do not have similarities in behaviour so they are classified on the bases of relations of similarity function to known objects. K-nearest neighbour's classifier can select an appropriate value of K based on cross-validation.





IV. PERFORMANCE MEASURES

Accuracy: Accuracy is defined as up to what extent a predictor can predict the value of a predicated attribute for new data values. it is expressed as:

Accuracy = Number of correct prediction Total number of the cases

Accuracy = (TP + TN) (TP + TN + FP + FN)

Kappa Statistics: The K statistics are used to compare the accuracy of the system to the accuracy of the random system, calculated by the following formula:

$$K = P(A) - P(E) (1 - P(E)$$

Here P(A) denotes agreement percentage, P(E) represents agreement chances. If the value of K is equal to 1 then agreement is considered in a tolerable range of classifier and the true value. If the value of K is equal to 0 then it indicates that there exists a chance of agreement.

MAE (Mean Absolute Error): This is a coefficient used to find how much a prediction is close to the possible outcome.

For interpreting results, 10 fold cross- validation method is chosen and applied on nine major classifiers namely IBK, Random tree, Naïve Bayes, J48, J Rip, K-star, Bayes Net, Part and REP tree. the results of our hybrid classifier are compared with the aforementioned classifiers. Finally, we have applied a combination of various classifiers and obtained optimum accuracy. The merging of IBK, J Rip and J48 generate better accuracy measures. Three other hybrid approaches provide better results when combined with IBK than the former technique obtains alone. Furthermore, testing is performed on the selected model using a test dataset with some number of instances to predict the internet usage of each instance [11]**.** The following screenshots showed in Figure represents that the result of the test performed on internet dataset in terms of five different classes. Moreover, Precision, Recall and F-Measure are also used in order to determine the prediction classification accuracy listed in Table Conclusion The proposed framework facilitated the prediction of the internet usage. It makes organizations of multiple sectors to achieve decision making and better future results. In this paper, we have adopted a hybrid approach to predict the Internet usage. Approximate thirteen thousand data samples were collected. Furthermore, the proposed framework has been tested and has gained support for providing a better foundation for predicting future internet adoption.

V. REFERENCES

[1]. Anupam Khan and Soumya K. Ghosh, "Analysing the impact of poor teaching on student performance," 2017.

[2]. Valerio Gross Andrea Romei and Franco Turini, "Survey on using constraints in data mining," vol. 31, no. 2, 2017.

[3]. Lirong Qiu and Jia Yu, "CLDA: An Effective Topic Model for Mining User Interest Preference under Big Data Background," 2018.

[4]. María Alejandra Malberti, Graciela Elida Beguerí Raúl Oscar Klenzi, "Visualization in a Data Mining Environment from a Human Computer Interaction Perspective," vol. 22, no. 1,

2018.
[5]. "Applied optimization and data mining," vol. 249, no. 2, 2017.
[6]. Yanju Zhou, and Xiaohong Chen Xin Liu, "Mining Outlier Data in Mobile Internet-Based Large Real-Time Databases," 2017.
[7]. Muzhou Xionga, Deze Zenga, and Junfang Gongb Hong Yaoa, "Mining multiple spatial-temporal paths from social media data," vol. 87, 2018.
[8]. Tianlai Li Fangai Liu and Xinhua Wang, "A novel approach for mining probabilistic frequent itemsets over uncertain data streams," vol. 11, no. 3, 2018.
[9]. T. Sheik Yousuf and M. Indra Devi, "Frequent pattern sub-space clustering optimisation algorithm for data mining from a large database," vol. 13, no. 3, 2017.
[10]. ChenYingb Chen , Victoria, C.P.bSeoung and BumKimc Bancha Ariyajunyaa, "Data mining for state space orthogonalization in adaptive dynamic programming," vol. 76, 2017.
[11]. Jurjen Jansena and Paulvan Schaikc, "Testing a model of precautionary online behaviour: The case of online banking," vol. 87, 2018.