Applying Data Mining Techniques in MOOC Recommender System for Generating Course Recommendations

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

Master of Technology

In

Computer Science and Applications

Submitted by

Harshit Jain

Roll No. 601534006

Under the supervision of:

Ms. Anika

Lecturer, CSE Department

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, "*Applying Data Mining Techniques in MOOC Recommender System for Generating Course Recommendations*", in partial fulfillment of the requirements for the award of degree of Master of Technology in *Computer Science and Application* 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 *Ms. Anika* 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.

(Harshit Jain)

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

13-July-17

(Ms. Anika)

Lecturer, CSE

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 **Ms.Anika**. 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 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.

(Harshit Jain)

Date: June, 2017
Place: Thapar University, Patiala

ABSTRACT

With the increase in Massive Open Online Courses across different platforms and domains, the course related information is being overloaded. It becomes a very tedious task for the learners to search for the required courses matching their individual goals, knowledge and interest. MOOCs recommender system plays a vital role by easing this task and providing courses of interest within an efficient time frame. This research work proposes an effective MOOC recommender system with the help of various data mining techniques. It encashes upon the fact that the learners involved in the MOOCs can be easily bifurcated into two categories of active and passive learners based upon their activity logs. It first channelizes the learners into these categories and then provides separate recommendations by applying different data mining approaches. Through this technique a significant level of accuracy has been achieved over other basic methods of course recommendations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

MOOC	Massive Open Online Courses
RF	Random Forest
DT	Decision Tree
LR	Logistic Regression
CN2	CN2 Rule Induction
CBR	Content-based Recommendation
CF	Collaborative Filtering
KNN	K-nearest Neighbor
PCA	Principal Component Analysis

CHAPTER 1

INTRODUCTION

In this era of big data lots of information is available on the internet, but from a user's perspective, it requires lots of efforts in terms of time, for finding the most relevant resources amongst these. A recommender system proposes a solution, by inculcating information filtering technique that provides relevant data based on one's previous search, other similar users' interests, contextual information, etc. Earlier, they were mostly being used in E-commerce sites for increasing their sales and providing personalized recommendation to their users, but now they have extended to different domains like recommendations for movies, items, news, hotels, courses, etc. The vast amount of e-learning material is available on the internet in the form of Online Educational Resources, MOOC etc., which provide thousands of courses, study materials, and certificates to learners [1].

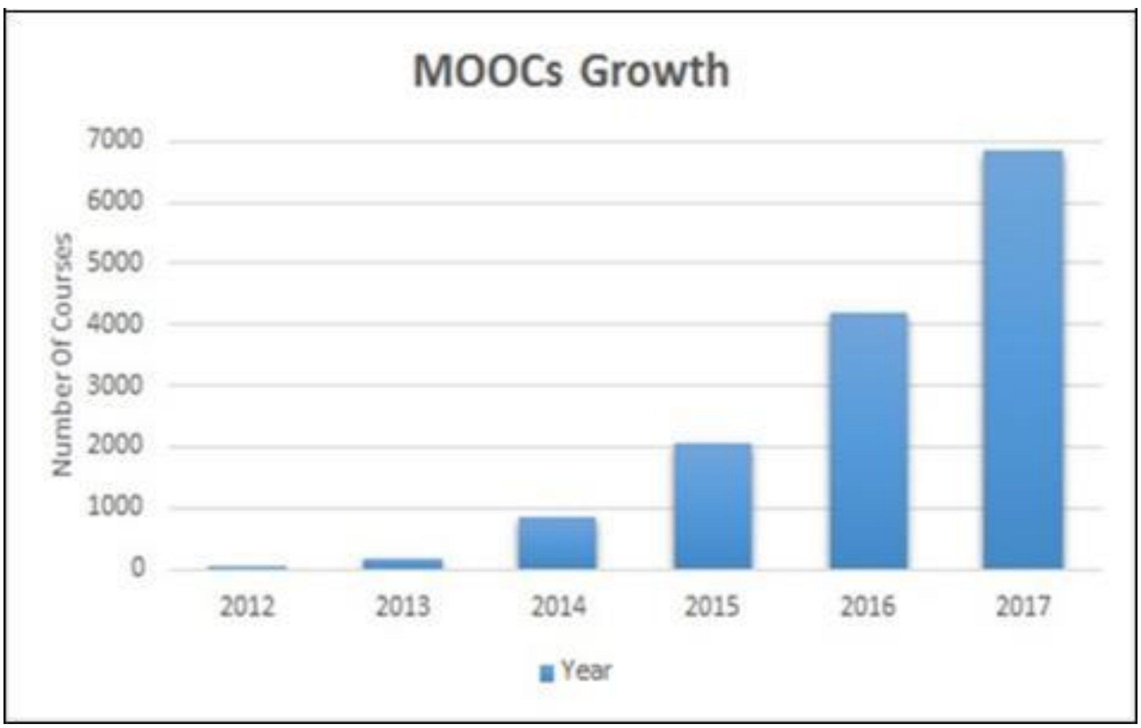

Figure 1.1: MOOCs growth in successive years

The focus of this thesis will be mainly Massive Open Online Courses. MOOC has several features like openness, flexibility, networked and lifelong learning which attracts learners across all over the world. MOOCs provide suitable resources to these learners according to their learning characteristics. Now-a-days lots of universities are offering various online courses via different platforms like coursera, Edx, Udemy, etc. It is gaining much more positive feedback based on its new way of teaching and different pedagogy. However, along with the expansion of these MOOCs as shown in Figure 1.1, it becomes difficult for the learner to find suitable courses for their career growth out of a large number of courses. So, an efficient system is needed that can provide a personalized recommendation to the learners, according to their learning pace, interests and context for choosing most relevant courses, within a short time span.

1.1 Recommendation Systems for E-Learning

An e-learning task recommender is a recommendation framework that would recommend a learning errand to a learner in light of the assignments effectively done by the learner and their success and in view of tasks. The likeness of the learners could be built up utilizing learner profiles or could be found on basic past get to patterns made by other comparative learners. The likeness of the learners could be built up utilizing learner profiles or could be found on past trends. On a fundamental level, there are two noteworthy parts in the plan of such a specialist: a learning module that gains from past access patterns and surmises an individual and an advising module that applies the educated model at offered times to suggest activities. There are numerous approaches to executing this procedure, for example, Data clustering, association rule mining, Collaborative Filtering and so forth [2].

1.1.1 Recommender System Approaches

There are different sorts of recommender framework: Content-based Recommendation, Collaborative Filtering Recommendation, Knowledge-Based, Hybrid Based and Demographic, as shown in Figure 1.2.

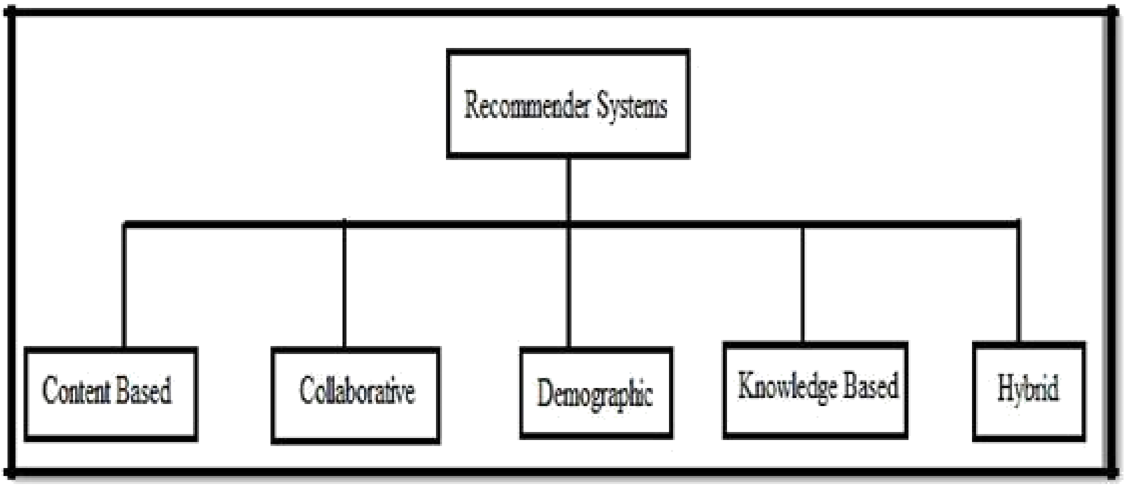

Figure 1.2: Diverse sorts of Recommender System [33]

- CBR recommends things that are like things the user enjoyed previously. Data about items is stored and utilized for the recommendations. Things chosen for recommendation are the things that similarity index most fulfill the user need is recommended on the top.
- CF are the recommendation method that requires the suggestion searchers to express their inclinations by rating things. It basically centered on thing closeness or client similarity.CF computed the likeness either between clients or between things. It is possible that it produces suggestion in light of client's comparability score i.e. on the off chance that client has a specific decision, at that point it seeks the client with a similar decision and prescribe their things to the objective client or using comparability score i.e. on the off chance that client loves a specific thing, at that point he will like the items that are like by other similar client. In this calculation, the parts of recommendation searcher (a client) and inclination supplier are blended; the more clients rate things, the more exact the recommendation moves toward becoming.
- Knowledge-Based Framework is situated in light of the obstructions brought by client inclinations and necessities.
- Frequent groupings can also build recommender frameworks. For instance, if a client often rates things we can utilize the successive example to prescribe different things to him.

- Random expectation calculation is a calculation that haphazardly picks things from the arrangement of accessible things and prescribes them to the client. Since the thing's choice is done arbitrarily, the exactness of the calculation depends on luckiness; the more noteworthy the quantity of things is the shot of good choice brings down.
- Hybrid based proposal calculation is the mix of at least two suggestion calculation to take profit of both the calculation.

1.2 Predictive Analytics

Predictive analytics is the utilization of information, factual calculations and machine learning procedures to distinguish the probability of future results in view of chronicled information. Predictive models utilize known outcomes to create or prepare a model that can be utilized to predict values for various or new information. The modeling brings about forecasts that represent to a likelihood of the objective variable in view of assessed criticalness from an arrangement of input variables.

The predictor is the central element which is used to foresee the future behavior. It utilizes measurable calculations and machine learning strategies to recognize the likelihood of the future outcome. The predictive modeling uses three or more predictive variables for implementing the model.

1.3 Data Mining technique used in MOOC for Course Recommendation

There are many data mining techniques which are used for generating a recommendation for the Learner.

1.3.1 Decision Tree

Decision Tree is a non-parametric supervised learning algorithm (having a predefined target variable) used for Classification and regression. It works for both continuous and categorical variables and its objective is to make a model that predicts the estimation of an objective variable by generating a decision rule from data features. It starts at the root node like tree structure contains an input training dataset (shown in Figure 1.3). For each possible value child node is constructed which contains decision about courses and their respective probability. A heuristic approach is used to choose split attributes which will produce the best decision tree [3].

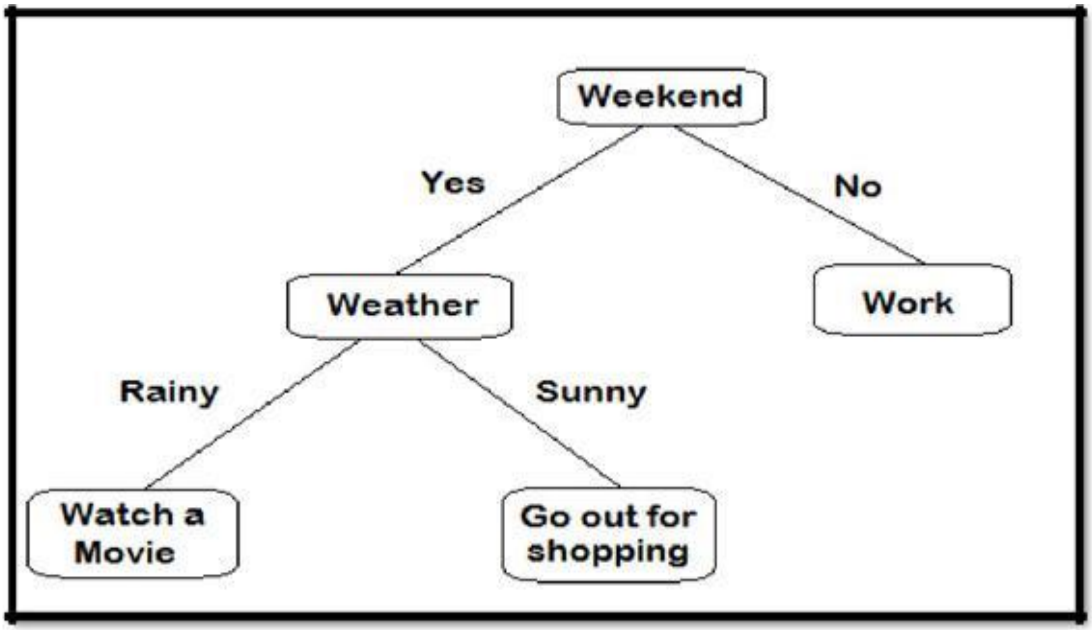

Figure 1.3: Decision Tree example [33]

1.3.2 K-Nearest Neighbor

K-Nearest Neighbor is a non-parametric managed learning method utilized for Classification work. The guideline behind closest neighbor strategies is to discover a predefined number of preparing tests nearest in separation to the new point and foresee the name from these. The quantity of tests can be a user characterized steady or change in view of the nearby density of points. Standard Euclidean distance is the most widely recognized decision for metric measure [4].

1.3.3 CN2 Rule Induction

CN2 Rule is a Classification technique used for classification purpose. It is designed for getting efficient induction rules of forms ‘if <condition> then predict <class>'. CN2 has two main functions: the search function and control function. Search procedure performs beam search in order to get rule and control function continuously executes the search. Cn2 contains three evaluation measures: Laplace accuracy, WRAcc (Weighted Relative Accuracy), and entropy [5].

1.3.4 Random Forest

RF is ensemble technique utilized for Classifications work. It develops numerous decision trees at preparing time and yielding the class either for order (method of the classes) or for regression of the single Classification trees, which is shown in Figure 1.4. For Classification assignment, the gathering of Classification tree votes in favor of the suitable or well-known class. In the regression task, reactions arrive at the midpoint of to get needy variable. Tree ensemble may prompt a critical change in order exactness. Every Classification tree is developed by taking a bootstrap test from the preparation information. Subsequent to building each tree, vicinity is figured for every last case and if two cases contain a similar end hub, their propinquity is expanded by one. Toward the end, their propinquity is standardized by isolating by the quantity of single trees [6].

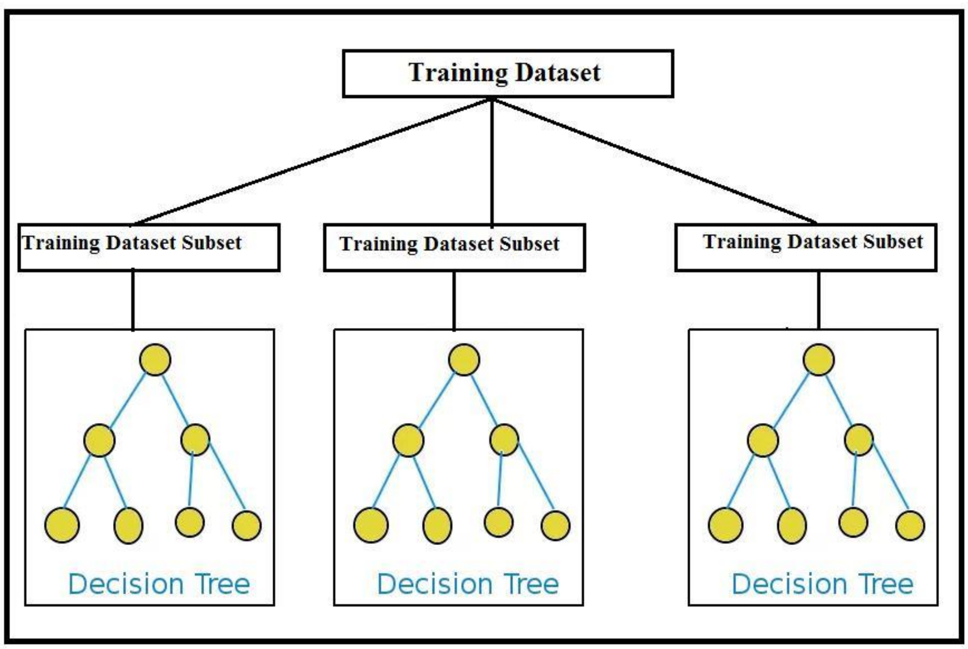

Figure 1.4: Random Forest example

1.3.5 Association Rule Mining

Association rule mining or learning is a control-based strategy to discover visit examples, connection, and relationship from datasets which are found in different databases, for example, a transaction database, social database, and whatever another archive. It has two parts: an antecedent (if) and a consequent (then). Association rules utilized the criteria support and certainty for getting most continuous if-then standards. It is the ramifications of the frame $X \rightarrow Y$, where X and Y are itemsets. For instance: if a Learner purchases a bread, he is 80% prone to likewise buy milk [7].

1.3.6 Logistic Regression

Logistic regression or logit regression is the statistical method used for analyzing a dataset in which dependent variable is binary or dichotomous i.e. it can take only two values 1(success, true, yes) or 0 (failure, false, no) and one or more independent variables that determine the result. It has two sorts: Binary LR i.e. when a subordinate variable has just two classifications and Multinomial LR i.e. when a subordinate variable has more than two classifications. It is also a predictive analysis that allows prediction of a discrete variable by discrete and continuous predictors [8].

1.3.7 K- Mean Clustering

It is a type of unsupervised learning algorithm used when unlabelled data i.e. data without defined groups or categories. The aim of this algorithm to find groups in the dataset that are related to each other, with the number of the cluster is represented by the value of k. Silhouette score is a measure of user similarity with its own cluster as compared to another cluster. The silhouette score value ranges from -1 to 1, here high value defines how well the user is matched to its clusters. The K-means clustering parameter includes distance measure and the value of k [9].

1.4 Thesis Structure

Rest of this theory is organized in an following way:

- Chapter 2 tells about the past work done in the field of MOOCs recommender frameworks for producing course suggestion. It likewise outlines all the work done in the field of MOOCs recommender frameworks and course recommendation.
- Chapter 3 tells about the issue proclamation which demonstrates the whole display in the MOOCs recommender frameworks.
- Chapter 4 depicts the experimental setup used for MOOCs recommendation. It

deals with the Active and Passive Learner datasets, data mining techniques and tools which are used for that task.

- Chapter 5 describes the implementation part. It shows the stepwise procedure to implement the proposed system which is mentioned in Chapter 4.
- Chapter 6 focuses on the proposed system accuracy, validation work and a comparison result between both the learners.
- Chapter 7 concludes the work and furthermore examines the future extent of this work.

CHAPTER 2

LITERATURE SURVEY

Heaps of research has been done in MOOCs suggestion framework which concentrated on understudy exercises. They manage the diverse proposal and information digging procedure for anticipating proper MOOC to the Learner.

DeBarr [10] proposed a strategy for recognizing spam utilizing bunching, RF and dynamic learning (shown in Figure 2.1). They utilized TF and DF for message portrayal. Clustering was utilized for naming the message, the RF was utilized for Classification and dynamic learning for amending the classifier performance.

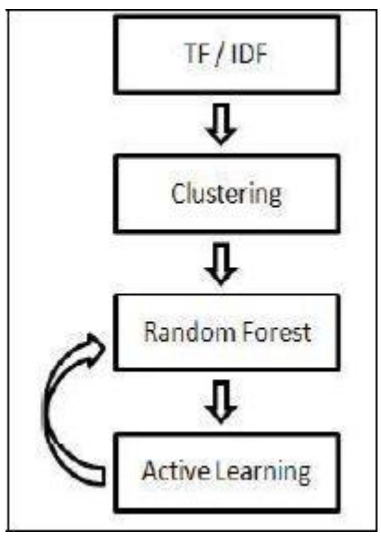

Figure 2.1: Workflow [10]

Gaddam and Shekhar [11] proposed a strategy for detecting anomaly utilizing a mix of k-mean clustering and ID3 Classification tree. Right off the bat, K-mean clustering divided the preparation information into k groups utilizing Euclidean distance as a similitude measure. From that point forward, on each cluster, they manufactured an ID3 decision tree to take in the subgroups inside the group to segment the choice space into more precise characterization regions. They also compare their classification performance with the individual ID3 decision tree and K-mean clustering.

Ouertani [12] represented a system that is built on the top of MOOCs provider. This system has two main functionality: first, it will improve the searching process for the Learner to find suitable courses among different MOOCs providers. Second, it will recommend courses to the Learner based on previous Learner experience.

Garg and Tiwari [13] suggest an efficient MOOC recommender system based on the K-mean clustering and collaborative filtering technique. First, the implicit rating is generated from learning the behavior of users. After that system produces the neighborhood clusters from user's database. This system is then trained to determine a feature vector for predicting the recommendations for its user. The workflow of this system is shown in Figure 2.2.

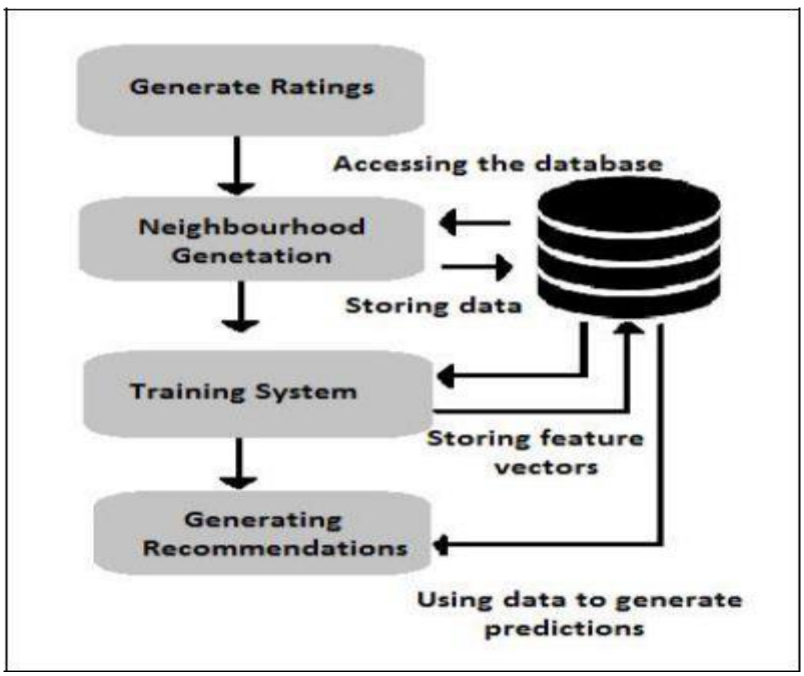

Figure 2.2: Workflow [13]

Bousbahi [14] recommended a framework (appeared in Figure 2.3) which has three layers: Interface layer, System Function Layer, and System data layer. The interface layer main function is to receive a request from the learner. System function layer provides functionality to create and process the case. System data layer has three sections: case base, the learner case, and comparable learner case. This system will take learner request as a query and recommend appropriate MOOCs to the learner. Essentially, they utilized Case-

Based Reasoning (CBR) and recovery systems to prescribe MOOCs to the learner based on their learning behavior and needs and knowledge.

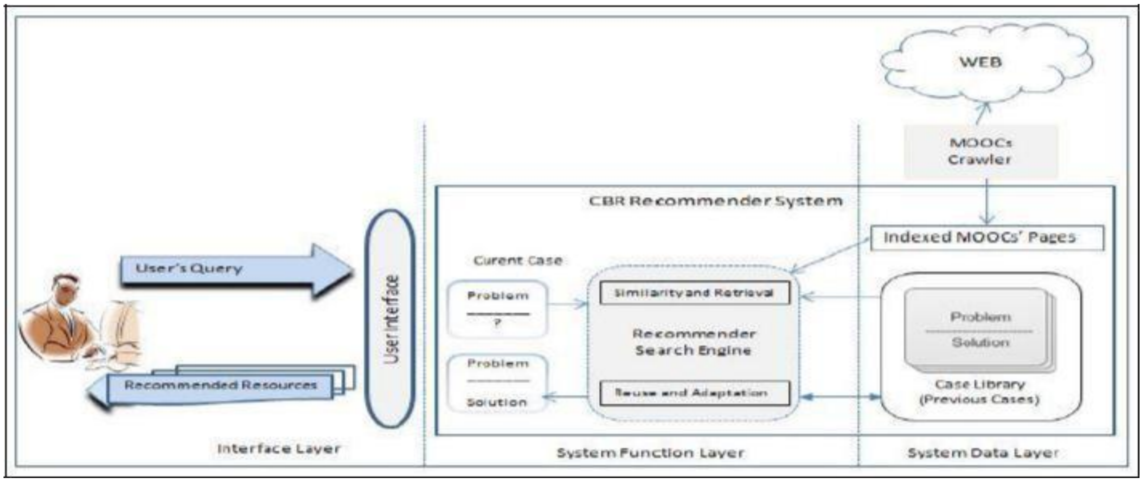

Figure 2.3: CBR Architecture [14]

Symeonidis [15] developed a MoocRec (shown in figure 2.4) website that consists of the recommendation engine, web crawler, and the database. It suggests courses to the learner that are useful for their required job. They used Matrix Factorization and collaborative filtering technique to predict the courses trend and prediction.

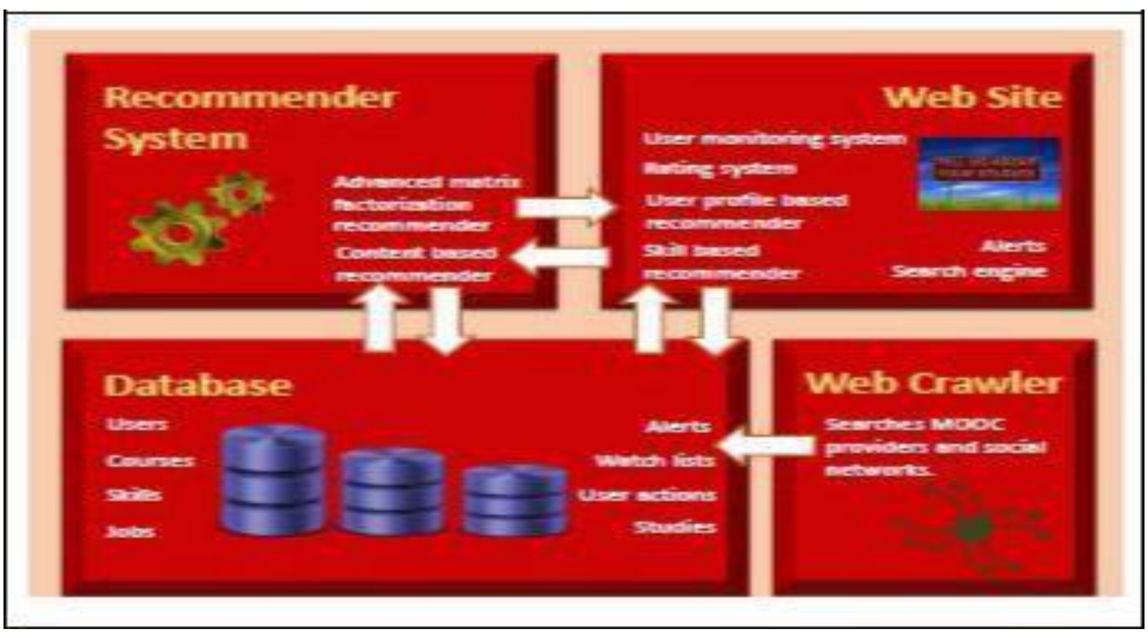

Figure 2.4: MoocRec Architecture [15]

Ahre and Lobo demonstrated [16], how information mining approach, for example, clustering and rule mining– Apriori are helpful in prescribing courses identified with the decision of other comparative Learners. Subsequently, it can prescribe the course to another Learner who as of late joined a course for e.g. C programming, the new course to select e.g. Java Programming. They compared their outcome with the open source information mining apparatus Weka. The outcome acquired utilizing joined approach is superior to anything individual methodologies. The proposed structure for course recommender framework is appeared in Figure 2.5.

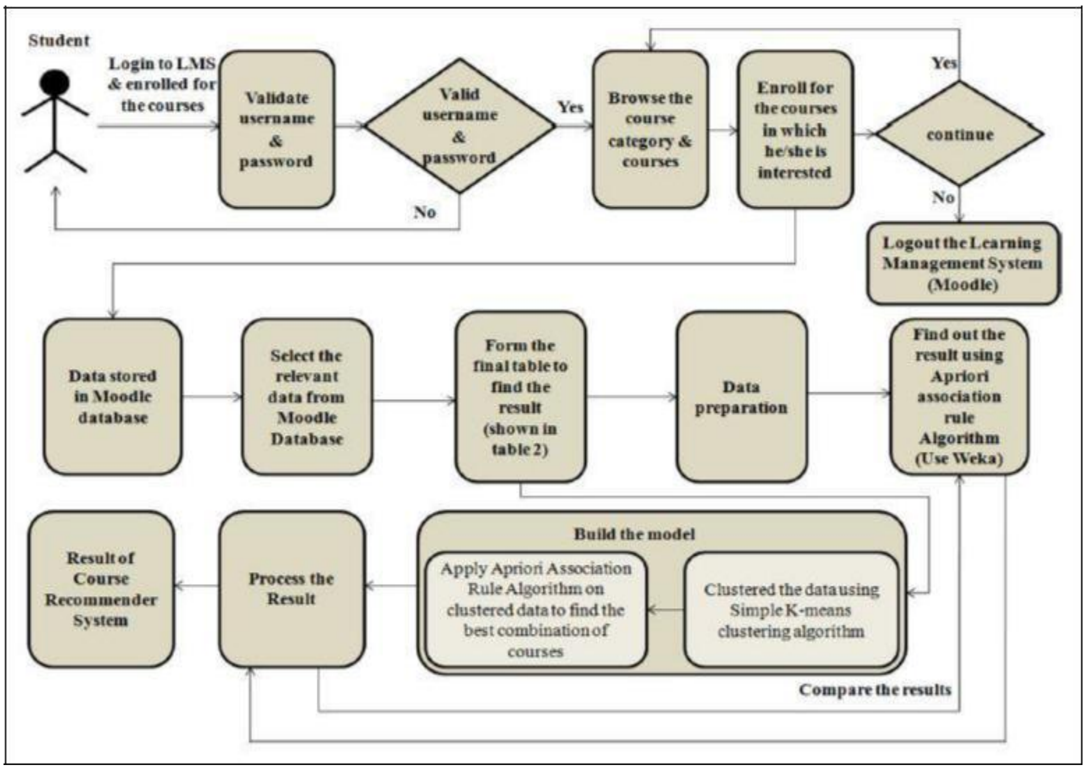

Figure 2.5: Framework for Course Recommender System [16]

FU and LIU [17] suggested a framework for undergraduate level recommender system (shown in Figure 2.6) in which particular features of Learners like Knowledge background, learning behavior, personal interest, and cognitive level are used to generate a recommendation for learner using content-based and collaborative filtering recommendation techniques.

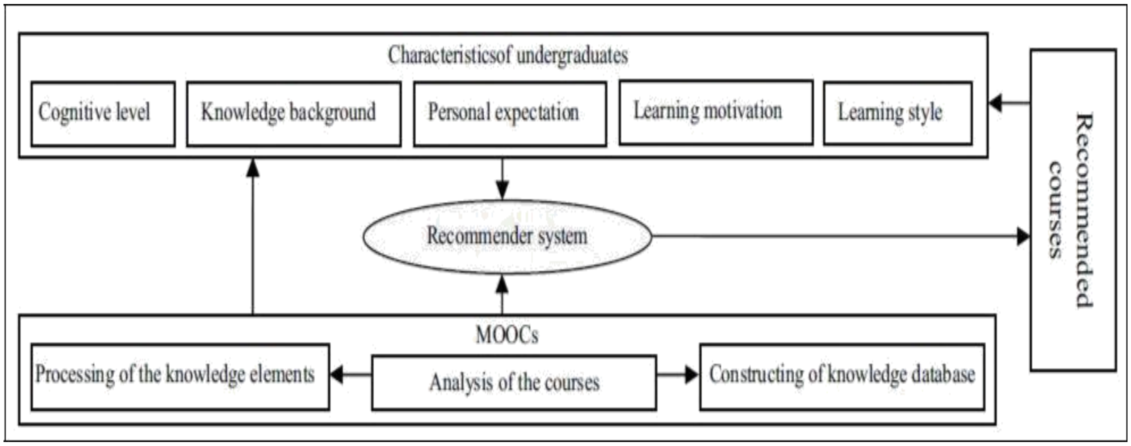

Figure 2.6: Undergraduate-oriented framework of MOOCs Recommender System [17]

MCRS (course suggestion framework for MOOC) [18] has made extraordinary change in the course recommendation framework. It depends on appropriated calculation structure. The Algorithm utilized as a part of MCRS is appropriated association rules mining method, which is a change of Apriori method. It is additionally helpful to discover concealed course relationship. Hadoop is utilized for information pre-handling to enhance the productivity of the calculation. Then Spark is used to mine association rule from data. Parallel, Sqoop is used to transfer course recommendation information into MySQL to improve the courses retrieval efficiency.

Zaiane [19] proposed the web mining systems prescribe on-line learning exercises in a course web page in view of learners' history and get to example to enhance course route and help the online learning process. They designed a recommender agent using data mining approach named association rule mining to recommend learning the material and other activities.

Upendran and Chatterjee proposed a course recommendation system that takes students 12th grade as a sign of their previous academic performance and learning ability of the student. They constructed the model from the former student data who have finished the course. The basic ideology behind this approach is that when a former student with some skills has completed the course then another new student with the same set of skills will also complete the course successfully [20].

Castro and Vellido provides an effective and efficient information about current research and data mining applications in e-learning such as to know about student failure, classification based on students performances, e-learning system recommendation, clustering of students and much more [21].

Carmona and Castillo proposed an adaptive machine learning technique to know about student preferences in a relative interval. Firstly they build an initial model using student background knowledge and their learning behavior. This model can be updated after getting more information about student learning styles with the system [22].

Elbadrawy and Karypis showed that how student and the academic feature can regulate the enrollment data. They used this features to specify the student and course group at various level. They defined how this feature can be important to design prediction model and top-n course model using collaborative filtering, matrix factorization [23].

Huang proposed a semantic e-learning framework to improve the productivity and viability of learner in online course framework and in addition to giving customized bolster. In this system, Learner can also modify their learning contents during learning and their result shows that this system can improve learner efficiency [24]

The Degree Compass [25] system is implemented to recommend courses for students in post-secondary degree programs. The main motive to develop this system to point out the course selection process that will best suit the student and also help the students to perform better in their academic to achieve a good rank. This system predicts the courses to overcome the problem, which students faced while selecting the courses.

The Course Agent [26] is another course recommendation system which contains details of available courses. This system collects the data from the students regarding their career interest and goals. It is a community-based system and the information is manipulated based on feedbacks from the user. The two types of feedback are used: implicit and explicit feedbacks. The implicit feedback is extracted from the user actions and explicit is collected from the user directly.

RARE [27] or Recommender System based on association rules which recommend the courses and the sequence of courses to be taken by the student. This system uses data mining along with user ratings to predict the courses. It uses historical data to mine the course rules. That rules are used for the recommendation and it also provides user to rate the courses to future improve the course rules.

Engin and Aksoyer discussed the implementation of two educational systems. The first system is course advising system which is developed to recommend courses to undergraduates and the second system suggests, a student about their scholarship based on their eligibility [28].

Garcia and Romero described a methodology to maintain and develop the web-based course system. They utilized information mining strategy names association rule mining to discover the course connection as IF-THEN principles. Afterward, they utilized CF method to score and offer the principles acquired by instructive specialists [29].

CHAPTER 3

PROBLEM STATEMENT

As discussed in the Literature Survey, a large number of systems are found to exist for making the course recommendations in MOOCs recommender system scenarios, but still a number of gaps have been identified that needs to be tackled over.

In most of the existing systems, the focus of course recommendations lies in the utilizing the learners' valuable insight and captured preferences information as well as educational background only. However, since the day by day exponential expansion of the MOOC ecosystem it becomes hard for an automated system to capture and recommend efficient courses to the learners based on these previously defined limited categories.

After close examination of the MOOCs educational datasets, it has also been found that despite of the fact of diversified backgrounds and learning paces, the learners have different learning behaviors too. For example, only some learners are found to be active throughout the course in which they are registered as compared to most of the others who have not even visited the course website after their registration in the course.

Another issue in the existing system's framework is to legitimately extract the data from the learner action and utilize that data appropriately for better course recommendations.

So, there is a need for an efficient MOOC recommendation system that recommends courses to the learners, according to their past learning history, educational background, learner's contextual information, learning behavior and which also goes in pace with the current educational and technological trends.

CHAPTER 4

METHODOLOGY

The methodology adopted for generating relevant Course Recommendations is shown in Figure 4.1.

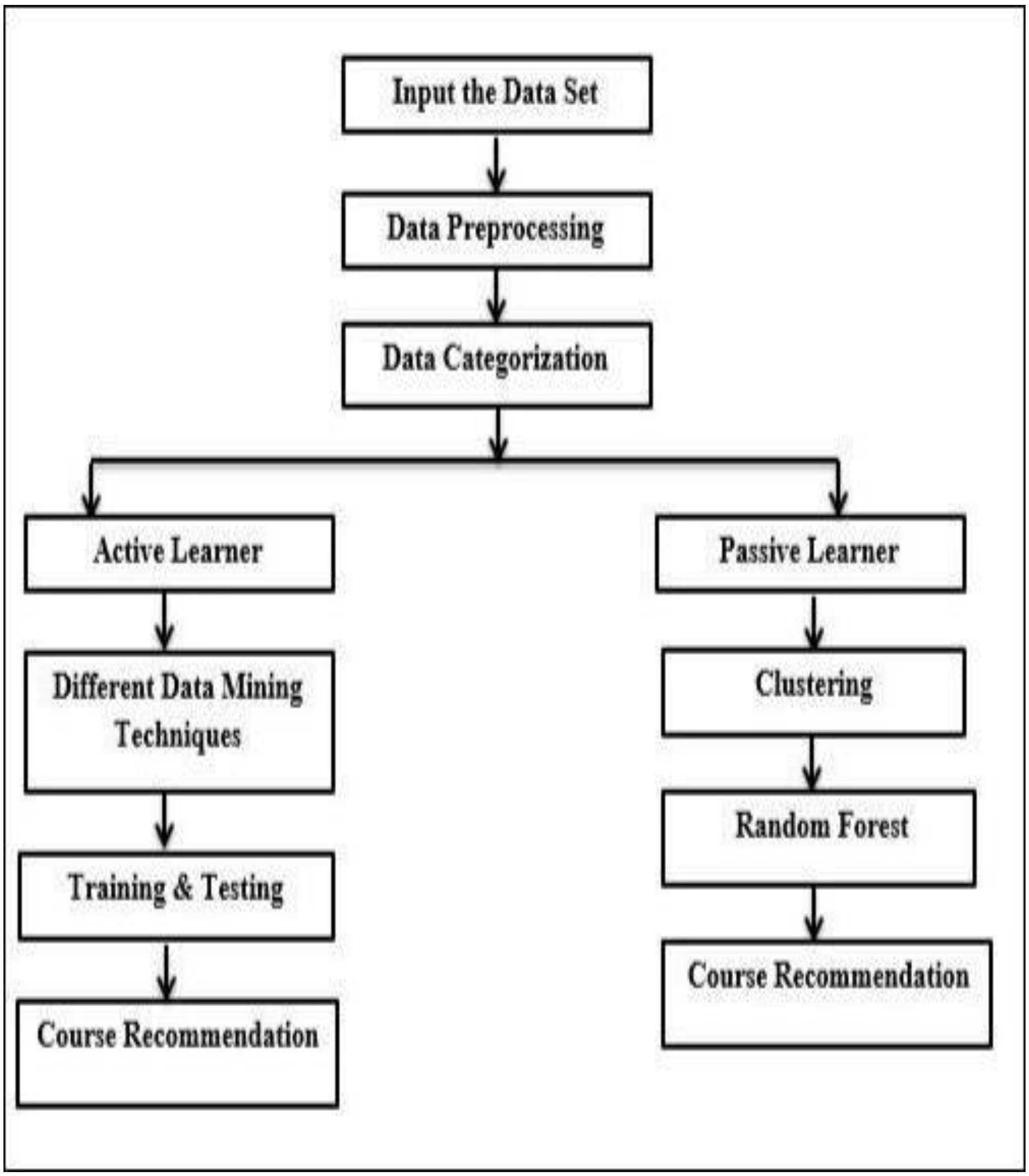

Figure 4.1: Workflow of generating Course Recommendations

4.1 Data Preprocessing

It is a strategy to change over crude information into the reasonable frame. In this dataset, data cleansing was applied to remove all the outliers and missing attribute rows. In data discretization, student age was calculated with respect to course completion year which is 2014. After that their age was gathered into discrete interim of age gatherings: 16-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 51-60, and 61-65.

4.2 Feature Selection

For this dataset, Recursive Feature Elimination strategy was utilized for disposing of elements that were not discovered important for instance enrolled field was wiped out. As every one of the applicants considered were enlisted in the particular courses and consequently this field contained just 1 as its esteem. At last, after components were discovered significant and were considered:

- ***Course_id:*** It represents the institution name, course name and its time period.
- ***Userid_DI:*** It represents the unique user id of the student.
- ***Registered:*** its value is 1 if the student is registered for the course.
- ***Viewed:*** its value is 1 if the student has accessed the course tab (videos, exam, problem set).
- ***Explored:*** its value is 1 if the student has explored at least half of the chapters of course which they registered.
- ***Certified:*** its value is one if the student has got the certificate otherwise 0.
- ***Final_cc_cname_DI:*** It contains the location of the student.
- ***LoE:*** it contains the student highest designation.
- ***YoB:*** it contains the year of birth of the student. For example 1988.
- ***Gender:*** possible value male (m), female (f), other (o).
- ***Grade:*** grade of a student in the course.
- ***nevents:*** Add up to number of connections with the course.
- ***ndays_act:*** Number of particular day understudy follow up on course.
- ***nplay_video:*** Number of video understudy played.

- ***nchapters:*** Number of parts, students act on.
- ***nforum_posts:*** Number of posts student has posted on a discussion forum.

4.3 Data Categorization

In MOOCs scenario, two basic types of learner were identified: Active Learner and Passive Learner based on their activity. Hence, the dataset was categorized into two categories. Table 4.1 shows the basic difference between these two different types of learner based on various parameters.

Table 4.1: Comparison between Active Learner and Passive Learner

Parameters	**Active Learner**	**Passive Learner**
Number Of course studied	Two or more	Only one
Viewed and explored the course	Yes	No
Certificate and grade	Got certificate and grade > 50%	No certificate
Number of features	21	12

The Active Learner dataset contained 21 features, first 5 features are User_id, Final_cc_cname_DI, YoB, LoE and Gender and last 16 features are courses which the learner was selected and its snapshot is shown in Figure 4.2. Course value is 1 if the learner has selected that course otherwise 0.

User_Id	final_cc_(	YoB	LoE_DI	gender	HarvardX/	HarvardX/	HarvardX/	HarvardX/	HarvardX/	MITx/14.7	MITx/2.01	MITx/3.09	MITx/3.09	MITx/6.00	MITx/6.00	MITx/6.00	MITx/6.00	MITx/7.00	MITx/8.02	MITx/8.MR
30093887	India	21-25	Secondar	m	1	0	1	0	1	1	0	0	0	0	0	0	0	0	1	0
30117026	United St	26-30	Bachelor'	m	0	1	1	1	0	0	0	0	0	0	0	1	0	1	0	0
30145567	United St	56-60	Bachelor'	m	1	0	1	0	0	0	0	0	0	0	0	0	1	0	1	1
30293570	United Ki	31-35	Bachelor'	m	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0

Figure 4.2: Active Learner Dataset

The Passive Learner dataset contained 12 features which are Course_id, Userid_DI, Final_cc_cname_DI, LoE, YoB, Gender, Grade, nevents, ndays_act, nplay_video, nchapters and nforum_posts and its snapshot is shown in Figure 4.3.

Course_id	Userid_DI	Final_cc_cname_	LoE	YoB	Gender	Grade	nevents	ndays_act	nplay_vid	nchapters	nforum_posts
HarvardX/PH278x/2013_Spring	30066539	United Kingdom	Secondary	1993	f	0.43	979	7	139	4	0
HarvardX/PH278x/2013_Spring	30210335	Other Europe	Bachelor's	1978	f	0.43	693	8	3	5	0
HarvardX/ER22x/2013_Spring	30328842	India	Bachelor's	1988	f	0.43	512	3	0	10	0
MITx/2.01x/2013_Spring	30163794	Spain	Master's	1984	m	0.43	641	6	35	5	0

Figure 4.3: Passive Learner Dataset

4.4 Data Mining Techniques used for Active Learner

The following techniques have been applied for Active Learner which is described below to get a better recommendation.

4.4.1 KNN

KNN is lethargic learning as well as a non-parametric algorithm which is used for Classification and Regression. In KNN distance measure is used to determine closeness between instances. KNN acts as a collaborative filtering technique, which stores all possible events and classifies new events based on the similarity with other similar events. Euclidean distance is used to find the separation amongst students and classified them into the similar neighbor groups. A number of neighbors allowed is set to 5 and uniform weight is used as the parameter value.

4.4.2 Decision Tree

A decision tree is a conventional, tree shape structure which is used to determine every possible result and statistical probability. Decision tree constructed for dataset using a recursive procedure which is shown is Figure 4.4. In this work, Course_Ids is selected as the target variable. Minimum number of instances for leaves is fixed to 2 and maximum tree depth is set to 100.

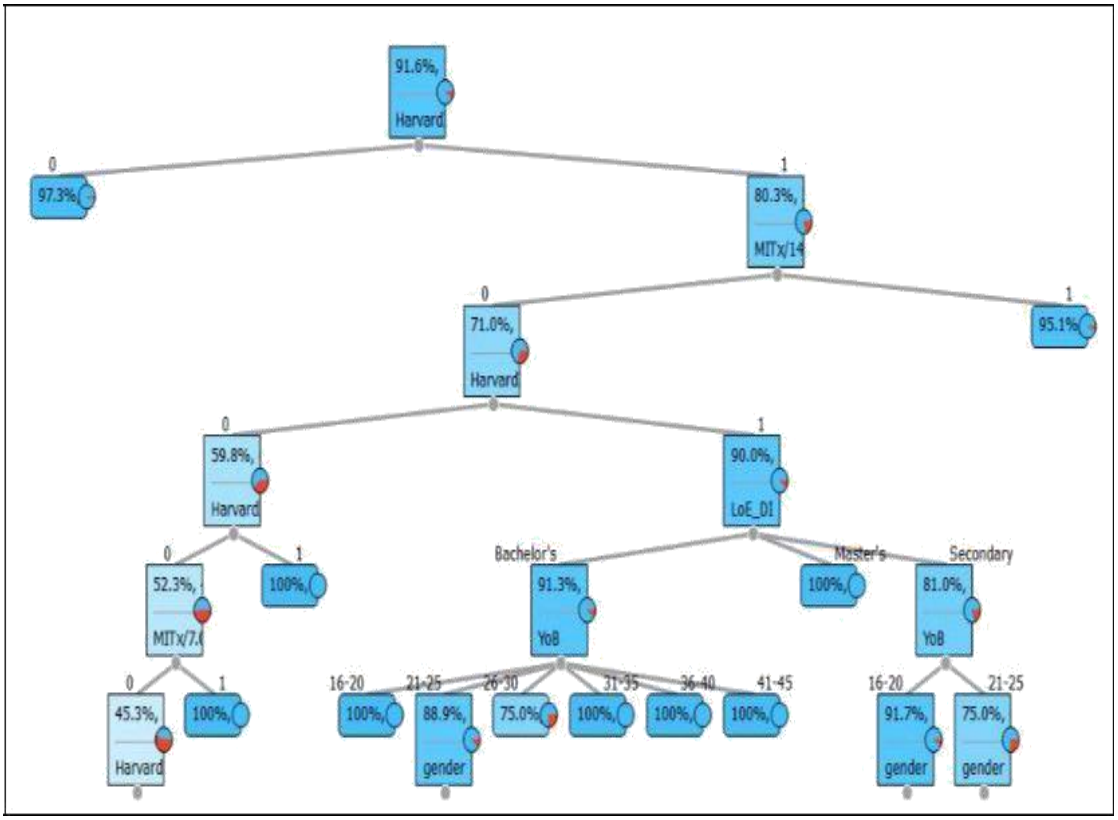

Figure 4.4: Pictorial representation of a Decision Tree

4.4.3 CN2 Rule Induction

CN2 rule induction is a supervised learning algorithm for getting propositional principles of the shape “if condition then predicts class” which is shown in Table 4.2. In this system following parameter of CN2 has been adjusted for obtaining better results: beam width as 5, entropy as evaluation measure and ordered rule ordering as a parameter.

Table 4.2: Rule mined by CN2 Rule Induction

S No	Rule Set(Top 25)
1	IF HarvardX_CS50x_2012!=0 THEN MITx_6.002x_2012_Fall=1
2	IF MITx_2.01x_2013_Spring!=0 THEN MITx_6.002x_2012_Fall=1
3	IF YoB==16-20 AND MITx_3.091x_2012_Fall!=0 THEN MITx_6.002x_2012_Fall=1

4	IF YoB==31-35 AND HarvardX_CS50x_2012!=0 THEN MITx_6.002x_2012_Fall=1
5	IF HarvardX_CS50x_2012!=0 THEN MITx_3.091x_2013_Spring=1
6	IF MITx_6.00x_2013_Spring!=0 THEN MITx_3.091x_2013_Spring=1
7	IF MITx_14.73x_2013_Spring!=0 THEN MITx_3.091x_2012_Fall=1
8	IF MITx_6.002x_2012_Fall!=0 AND gender==f THEN MITx_3.091x_2012_Fall=1
9	IF HarvardX_PH207x_2012_Fall!=0 THEN MITx_3.091x_2012_Fall=1
10	IF HarvardX_PH278x_2013_Spring!=0 THEN MITx_3.091x_2012_Fall=1
11	IF HarvardX_CB22x_2013_Spring!=0 THEN MITx_14.73x_2013_Spring=1
12	IF MITx_6.00x_2012_Fall!=0 THEN MITx_14.73x_2013_Spring=1
13	IF MITx_6.002x_2013_Spring!=0 THEN MITx_14.73x_2013_Spring=1
14	IF MITx_3.091x_2013_Spring!=0 THEN MITx_8.02x_2013_Spring=1
15	IF YoB==16-20 AND MITx_6.00x_2013_Spring!=0 THEN MITx_8.02x_2013_Spring=1
16	IF HarvardX_CS50x_2012!=0 THEN MITx_8.02x_2013_Spring=1
17	IF MITx_6.002x_2012_Fall!=0 AND gender==f THEN MITx_2.01x_2013_Spring=1
18	IF MITx_7.00x_2013_Spring!=0 THEN MITx_2.01x_2013_Spring=1
19	IF MITx_14.73x_2013_Spring!=0 THEN MITx_8.MReV_2013_Summer=1
20	IF MITx_8.MReV_2013_Summer!=0 THEN HarvardX_ER22x_2013_Spring=1
21	IF HarvardX_CS50x_2012!=0 THEN HarvardX_ER22x_2013_Spring=1
22	IF MITx_6.00x_2013_Spring!=0 THEN HarvardX_ER22x_2013_Spring=1
23	IF HarvardX_ER22x_2013_Spring!=0 THEN HarvardX_CS50x_2012=1 IF HarvardX_CS50x_2012!=0 AND final_cc_cname_DI==Russian Federation THEN MITx_6.002x_2013_Spring=1
24	IF HarvardX_PH207x_2012_Fall!=0 AND HarvardX_CS50x_2012!=0 THEN HarvardX_ER22x_2013_Spring=1

25	IF HarvardX_CS50x_2012!=0 AND final_cc_cname_DI==Russian Federation THEN MITx_6.002x_2013_Spring=1

4.4.4 Logistic Regression

Logistic regression or logit regression is a form of regression analysis conducted when the reliant variable is in the binary form shown in Figure 4.5. In this work, target variables are Course_Ids and represents whether a student has registered for that course or not. The parameters used for its implementation are: Ridge (L2) as regularization type and cost strength are set to1.

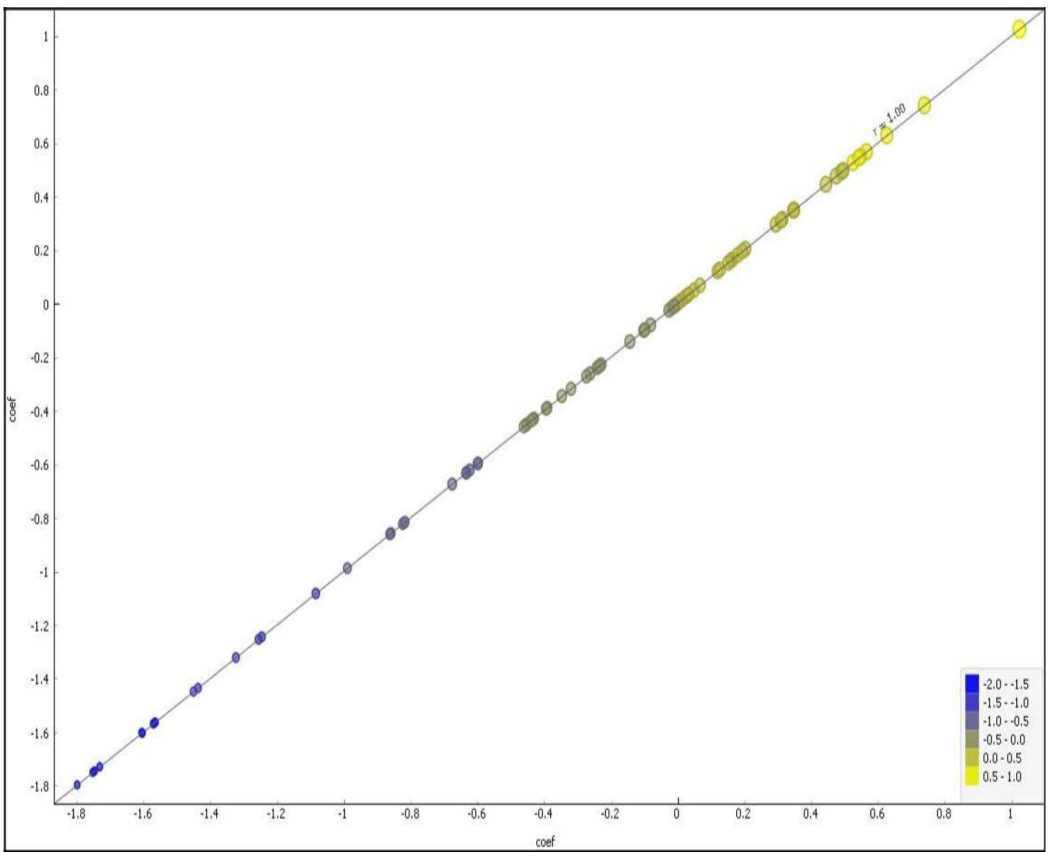

Figure 4.5: Graph of Logistic Regression

4.5 Data Mining Technique used for Passive Learner

The techniques used for Passive Learner datasets are as follows.

4.5.1 K-mean Clustering

K-means clustering is an unsupervised learning method which is utilized when initials information marks are not given. The pictorial portrayal of grouping for the given dataset of the Passive Learner is appeared in Figure 4.6. There are confirmations where clustering is consolidated with the Random Forest for getting better outcomes in other application areas [30]. In this framework, Clustering is utilized to isolate Passive Learner dataset into comparative learner groups. For this framework, Silhouette score was utilized to quantify the learner comparability and k estimation of 4 with high Silhouette score was picked as cluster size.

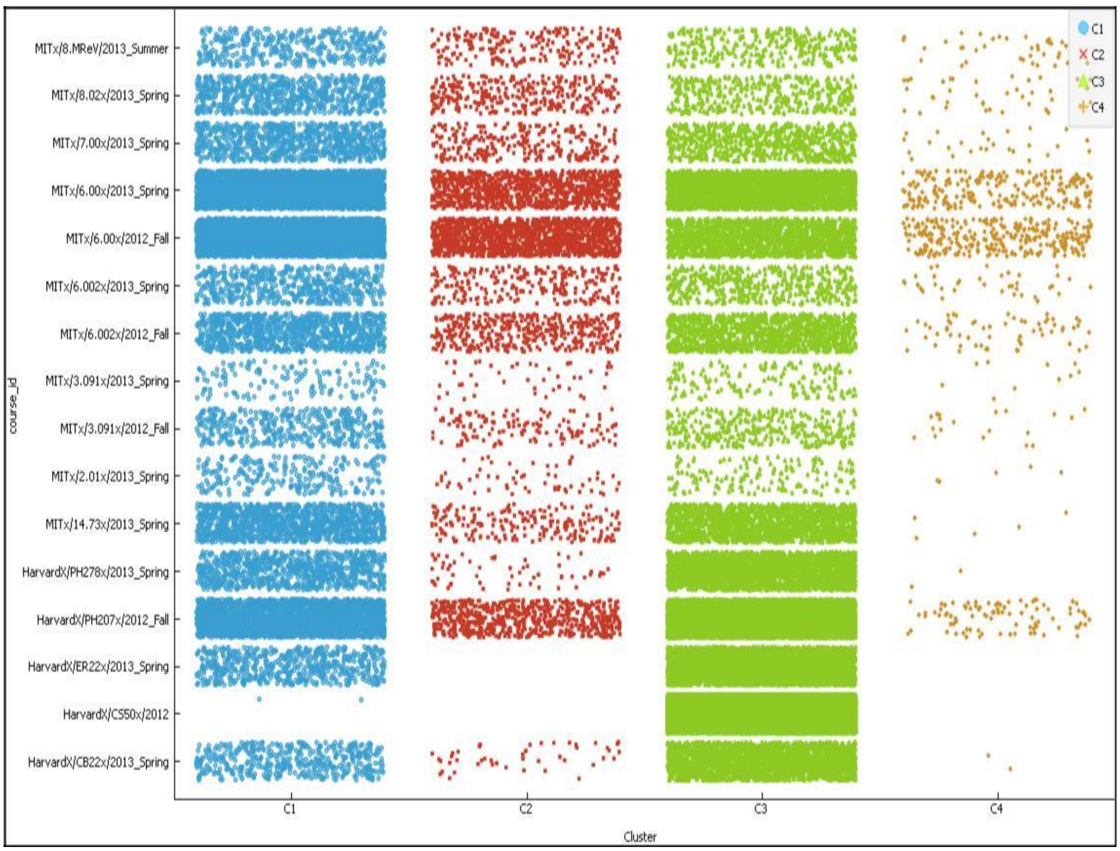

Figure 4.6: Each cluster containing a different set of courses.

4.5.2 Random Forest

RF is a gathering of Classification Trees, which are utilized to decide final outcome and can be utilized for classification tasks. Every characterization tree is constructed from a bootstrap sample of the training dataset. While constructing single trees, a random subset of features is taken out and best features are selected for the split. The classification is based on the absolute majority of the individual Classification Tree in the forest. In this work, the following parameters have been adjusted and used for the Random Forest: a number of trees in the forest are 10 and number of elements to be drawn at each node is 5. Pictorial representation of Random Forest is shown in Figure 4.7.

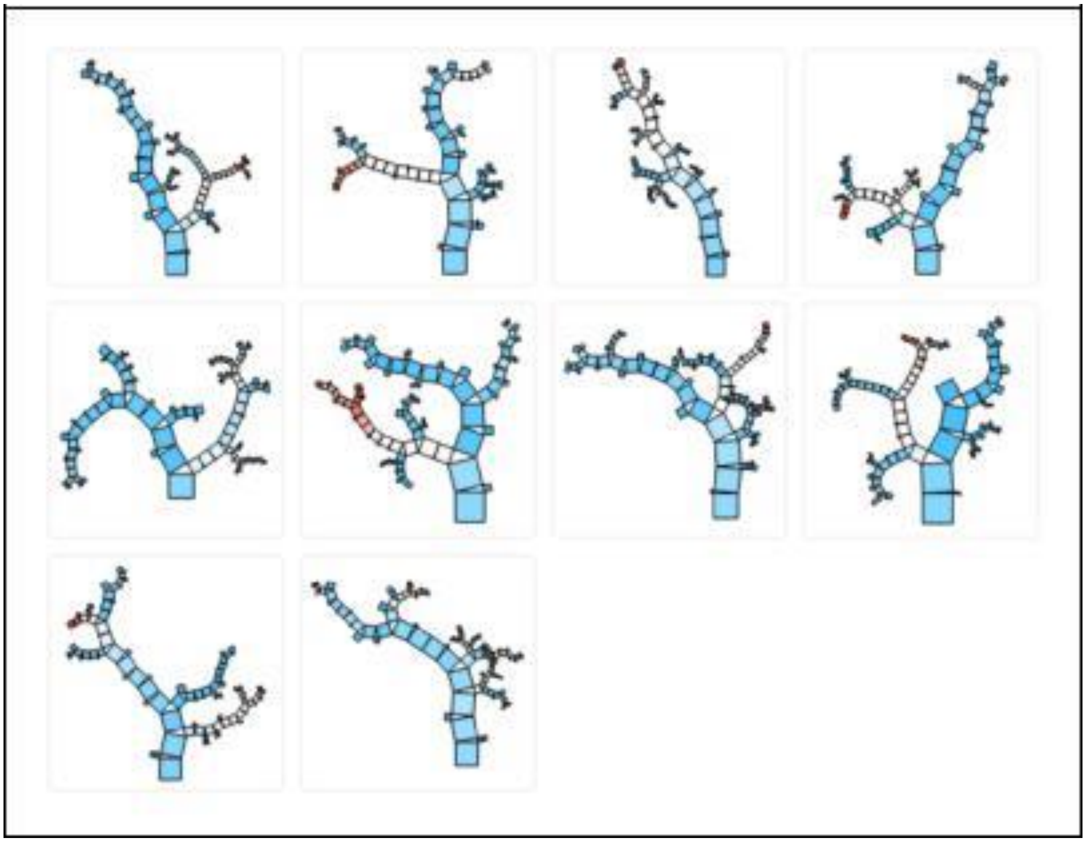

Figure 4.7: Pictorial representation of Random Forest using pythagorean tree

CHAPTER 5

IMPLEMENTATION

5.1. Dataset Description

The proposed system has used the dataset of Harvard and MIT of first year open online courses which they released in 2014 [31]. Dataset contains more than 600000 student's records which are registered on courses released in the scholarly year 2013(Fall 2012 - Summer 2013). Each record contains individual activity of Learner on any 16 courses which they studied on edX platform.

5.2. Tools and Technology

Python programming dialect was utilized for execution work. This is a scripting language that gives an inbuilt library for machine learning, information mining, recommendations, mining and substantially more. It supports all data mining operations with the help of Panda and orange library which is used for data mining task. Because of all these components, Python was utilized as an execution language. Panda python library was also used for data preprocessing and data categorization task. Orange python library was used for implementing the data mining algorithms which are useful for generating the recommendation for the Learner in MOOCs environment.

5.3. Data Mining using Orange

Orange3 [32] is used for implementation purpose and it provides data visualization in an interactive manner and it is also available in the form of python library. Orange3 is visual programming software that provides services such as data visualization, data mining, and machine learning. It provides simplicity to the researcher to use its graphical interface for their research work. As it is a tool that contains widgets in its interface that takes the user input, process that data and simply provide output to the user without any hurdle as well as provides graphical results. Figure 5.1 demonstrates the trial setup used for Active

Learner dataset during data mining step.

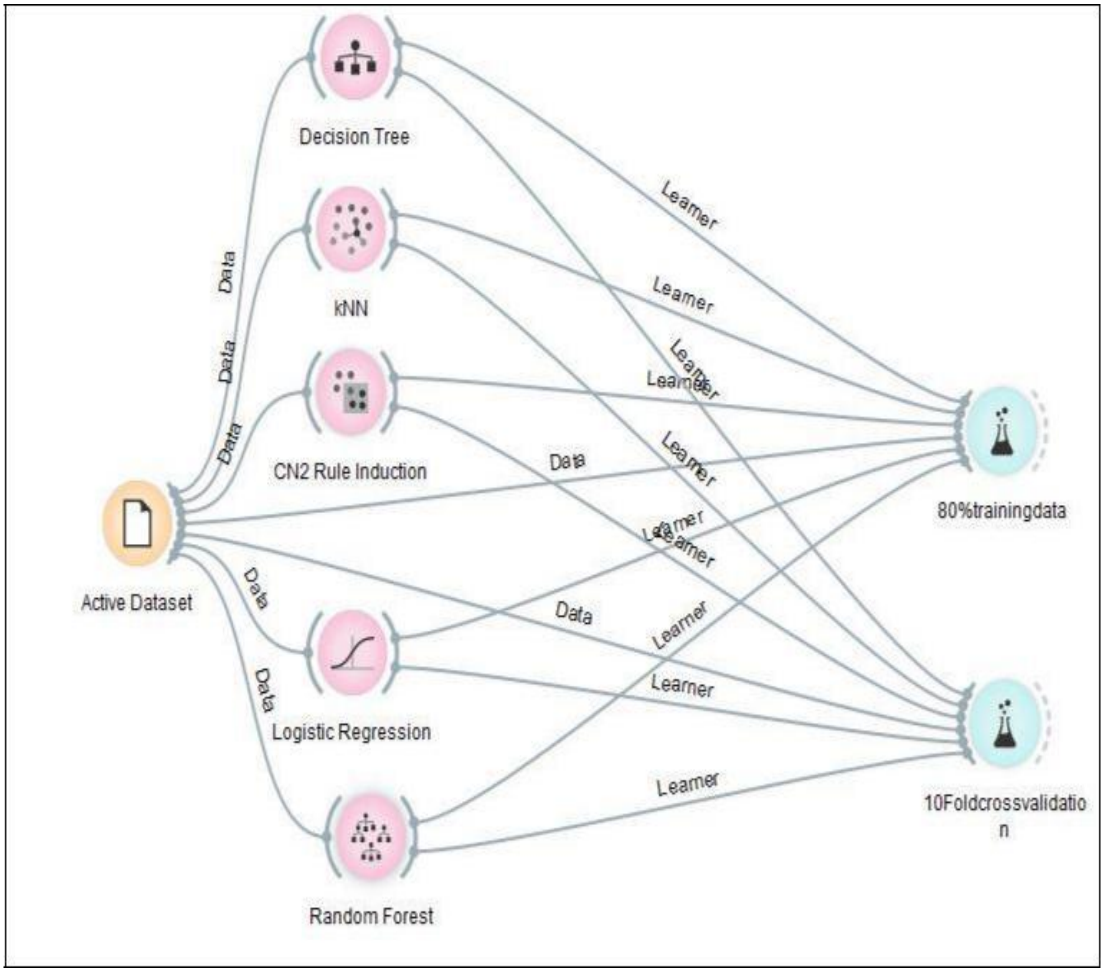

Figure 5.1: Experimental setup for Active Learner

It comprises of the accompanying steps (Figure 5.1):

- Active Learner data file is uploaded into File Widget and target variable is fixed.
- Perspective of the information record can be seen by Data Table Widget.
- Various Classification algorithm is being applied on Active Learner dataset by using their respective widgets.
- At that point classifiers are validated and tested using Test Widget of Orange3.

Figure 5.2 and 5.3 shows the experimental setup used for Passive Learner dataset to generate a recommendation for the Passive Learner.

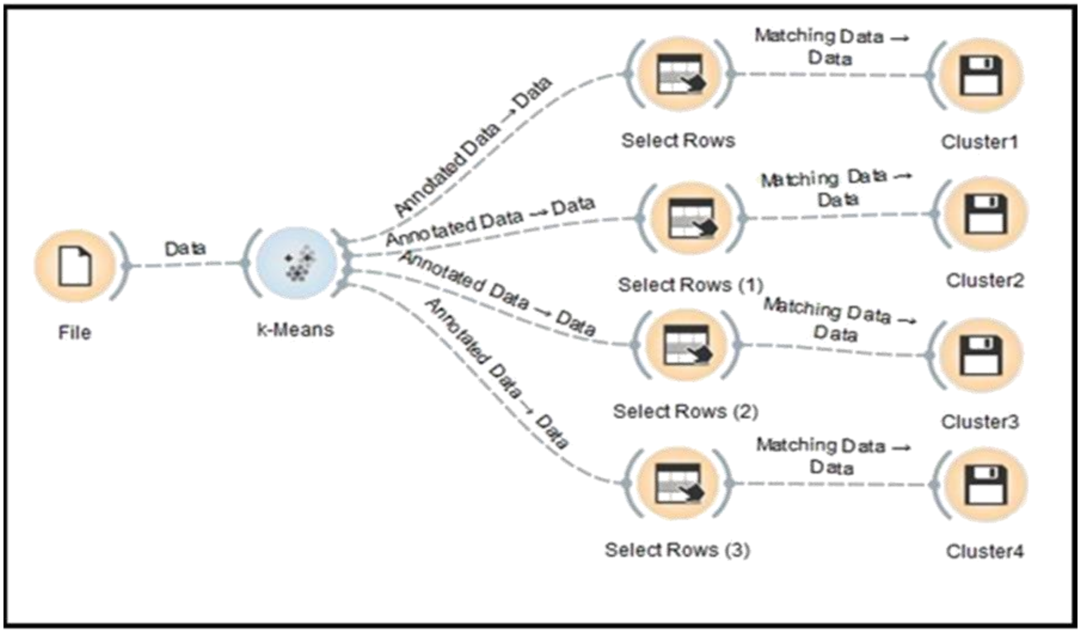

Figure 5.2: Experimental setup for clustering of Passive Learner.

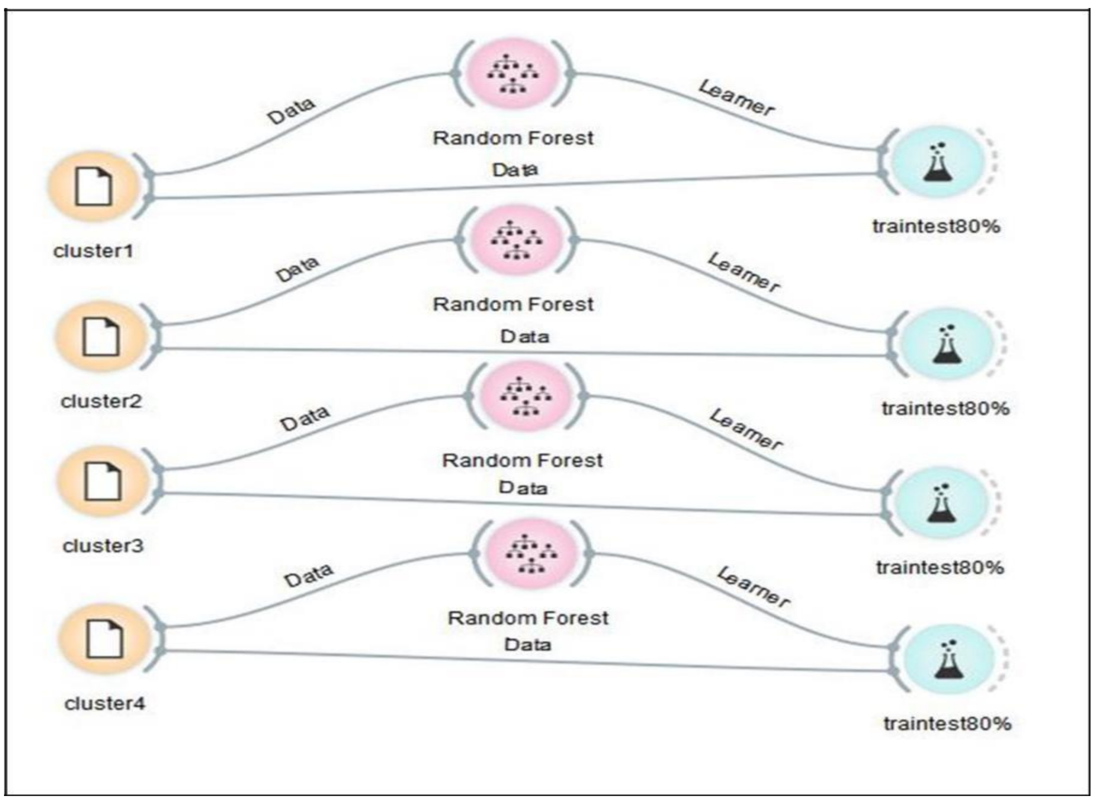

Figure 5.3: Experimental setup for Passive Learner

They consist of the following steps (Figure 5.2 and Figure 5.3):

- Passive Learner dataset is uploaded into the File Widget.
- K-mean clustering is being applied on this dataset using Clustering Widget.
- All k clusters were stored on k different CSV files using Save Data Widgets.
- All CSV file is being uploaded on k individual File Widgets.
- On each file, Random Forest classifier is applied using Random Forest Widget.
- Then these classifier is tested using Test Learner Widget with 80% training data (shown in Figure 5.4).

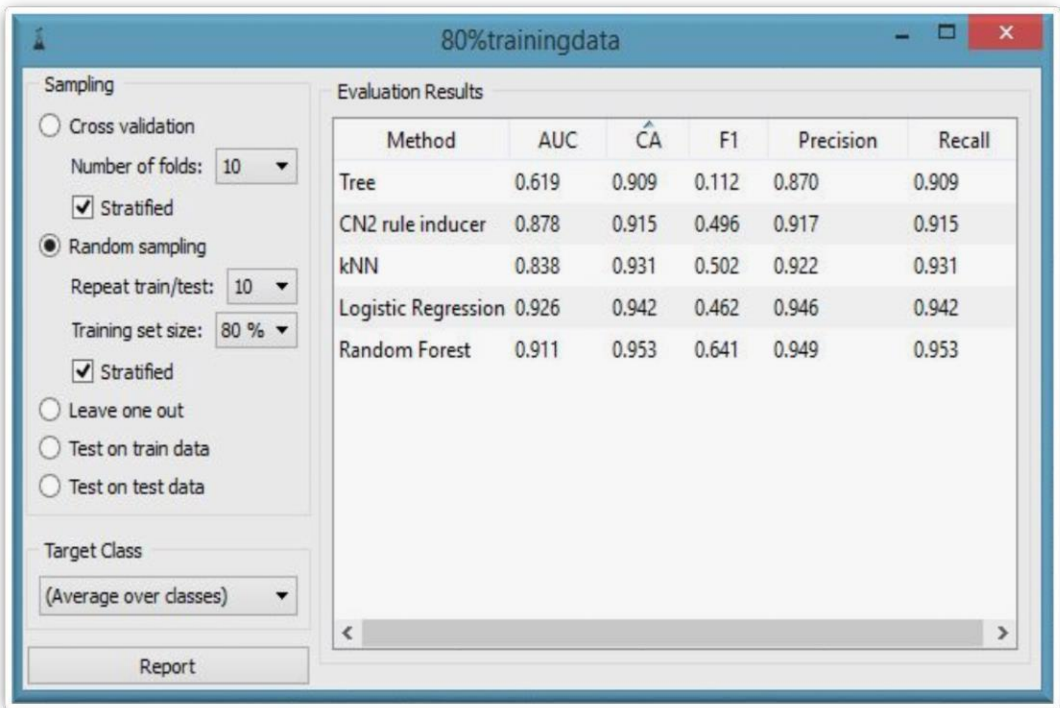

Figure 5.4: Experimental setup for testing of dataset

CHAPTER 6

RESULTS AND DISCUSSION

The considered dataset contained 641,138 Learner records and 16 courses. The dataset was categorized into two major groups. The proposed system was trained and tested with 80% training data and 20% testing data on Active and Passive Learner groups separately. The proposed system was then judged on following performance measure: AUC, CA, F1 Score, and Precision.

- ***AUC (Area Under the Curve):*** It is used in Classification to analyze which of the models predicts the classes better. It determines the performance of binary classification.
- ***F1 Score:*** In binary classification, F1 score measures the test accuracy. It is calculated by considering both recall and precision. Here, P is precision and R is recall.

$$F1\ score = 2 * (P.\ R) / (P + R) \quad (1)$$

- ***CA (Classification Accuracy):*** It is the ability of the classifier to predict the classes accurately.

$$Accuracy = (TP+TN) / (TP+TN+FP+FN) \quad (2)$$

Here, TP is true positive, TN is a true negative, FP is false positive, and FN is a false negative.

- ***Precision:*** It is a positive prescient esteem which is the ratio of TP and (TP+FP) where TP is a total number of True Positive and FP is a total number of False Positive.

$$Precision = TP / (TP+FP) \quad (3)$$

6.1 Results for Passive Learner

On the Passive Learner dataset, Clustering and Random Forest were applied consecutively and the results obtained after its training and testing are shown in Table 6.1.

Table 6.1: Performance measures of Passive Learner Dataset

Method	AUC	CA	F1	Precision
Random Forest	0.843	0.510	0.474	0.459

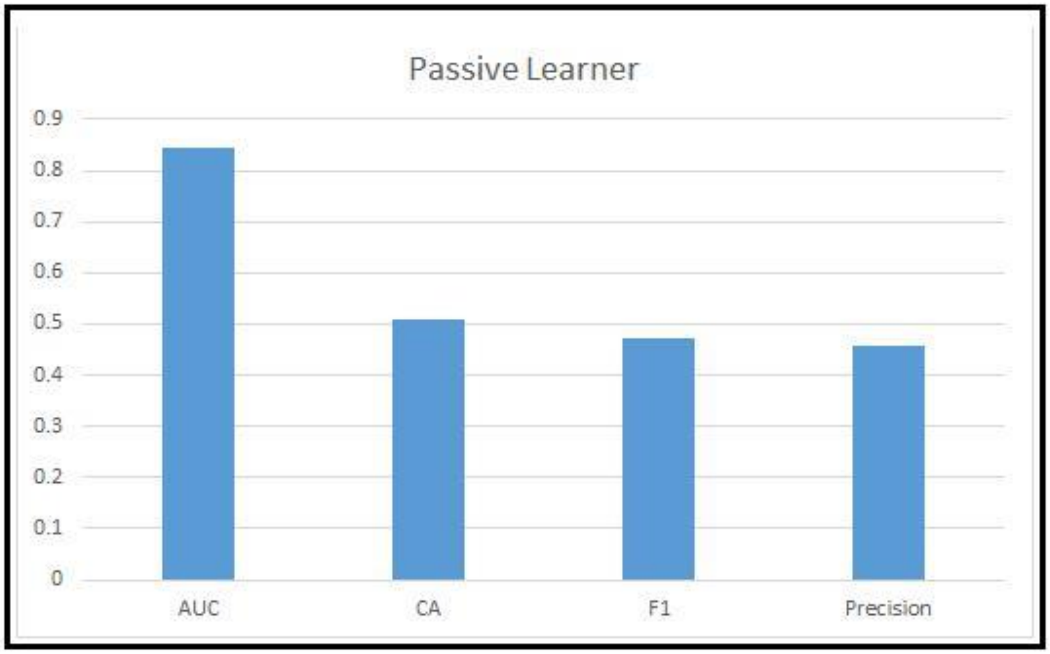

Figure 6.1: Comparative analysis of different performance measures using Random Forest for Passive Learner

6.2 Results for Active Learner

On the Active Learner dataset, 5 different data mining techniques such as Decision Tree, Random Forest, CN2 Rule Induction, Logistic Regression and KNN were applied for building a Learner model and these models were then tested for efficiency against different performance measures. Figure 6.2, 6.3, 6.4, 6.5, and 6.6 represent the graphs for the 5 different data mining techniques, used in case of Active Learners along 4 performance measures across 16 courses (C1, C2… C16). Logistic Regression proved to be the best amongst all other Classification techniques used.

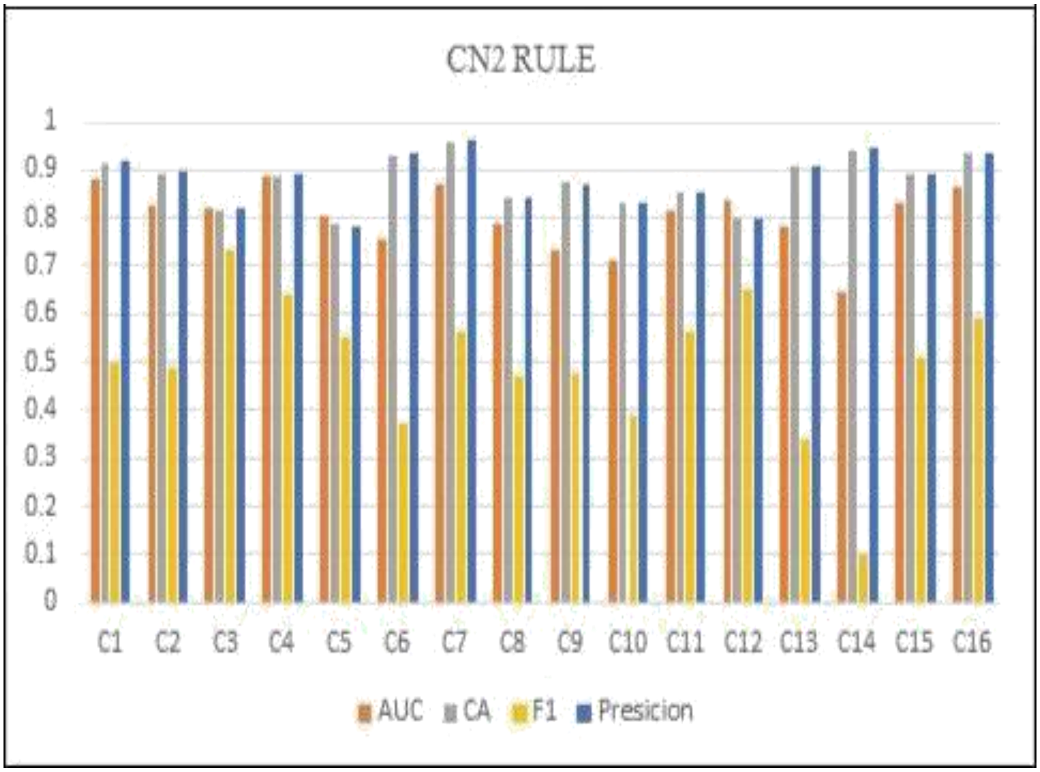

Figure 6.2: Comparative analysis of different performance measures using CN2 rule

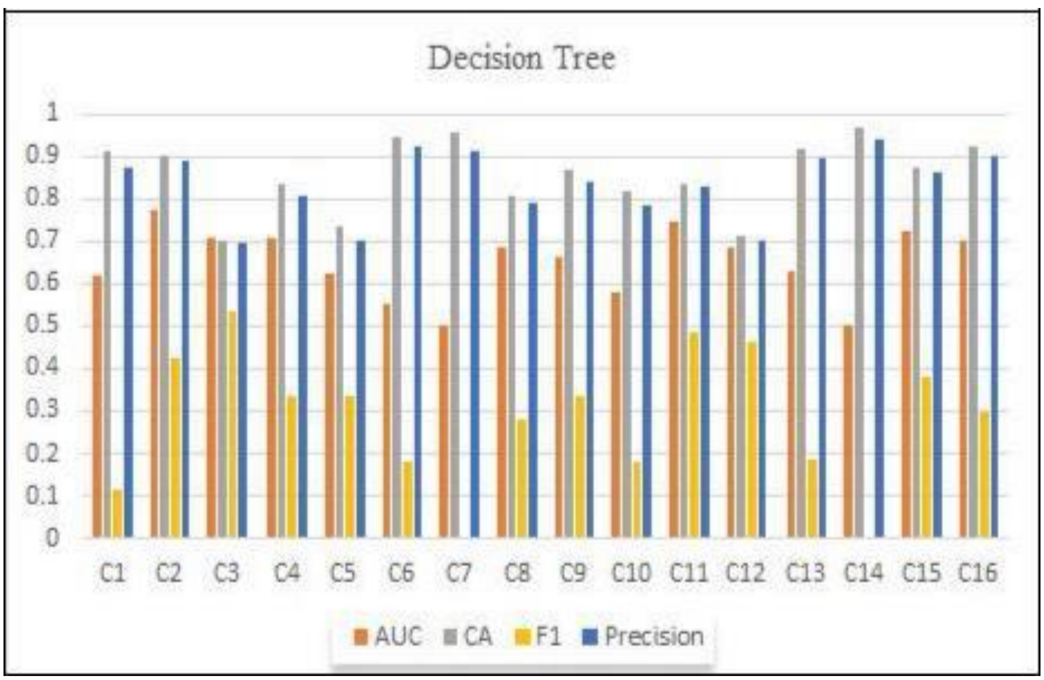

Figure 6.3: Comparative analysis of different performance measures using DT

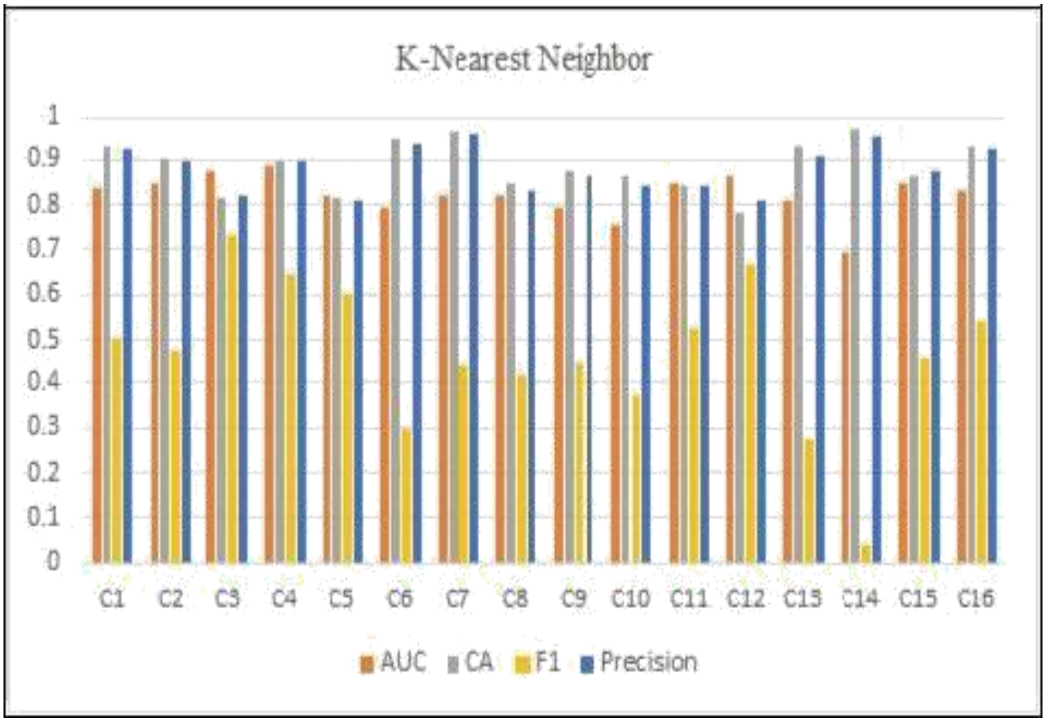

Figure 6.4: Comparative analysis of different performance measures using KNN

Figure 6.5: Comparative analysis of different performance measures using LR

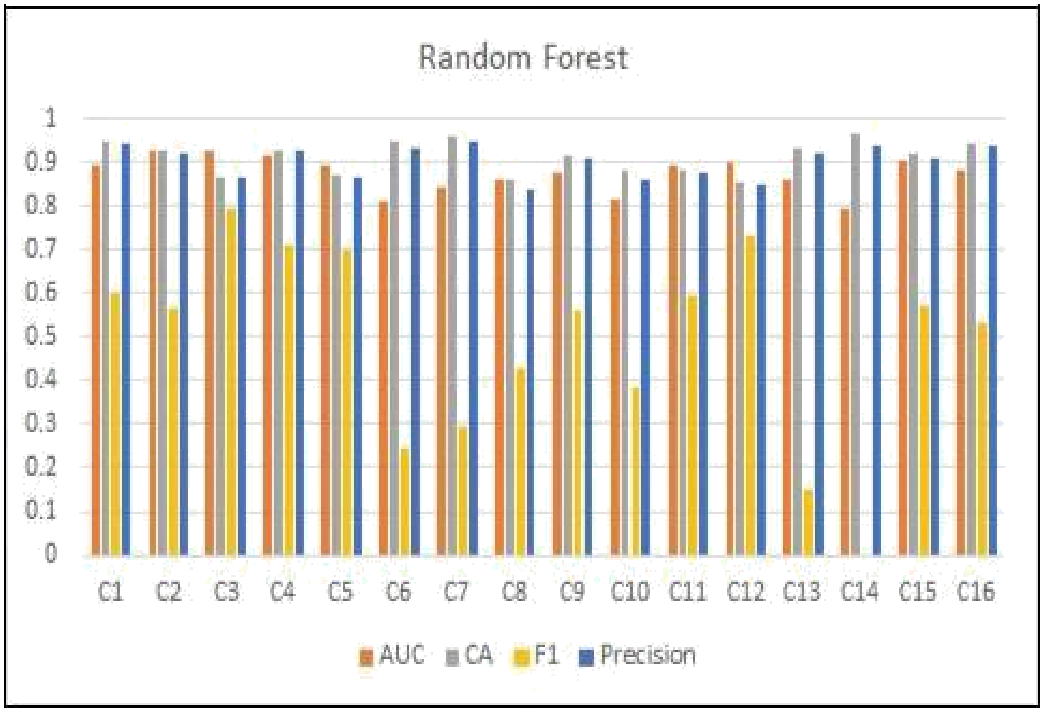

Figure 6.6: Comparative analysis of different performance measures using RF

In Active Learner dataset, all 16 courses with their respective classifiers (Decision tree, KNN, CN2, Random Forest, Logistic Regression) accuracy are shown in Table 6.2 and classifier normalized accuracy with other performance measures is shown in Figure 6.7.

Table 6.2: Accuracy of each Course Recommendation using different classifiers

Courses	Decision Tree	KNN	CN2 Rule	Random Forest	Logistic Regression
HarvardX_CB22x_2013Spring	0.909	0.931	0.915	0.949	0.942
HarvardX_CS50x_2012	0.898	0.901	0.891	0.928	0.937
HarvardX_ER22x_2013Spring	0.703	0.813	0.813	0.866	0.918
HarvardX_PH207x_2012Fall	0.832	0.896	0.886	0.927	0.941
HarvardX_PH278x_2013Spring	0.734	0.814	0.785	0.872	0.916

MITx_8MReV_2013Summer	0.944	0.945	0.931	0.949	0.946
MITx_201x_2013Spring	0.956	0.964	0.959	0.96	0.965
MITx_600x_2012Fall	0.809	0.849	0.842	0.858	0.872
MITx_600x_2013Spring	0.865	0.873	0.876	0.915	0.931
MITx_700x_2013Spring	0.818	0.865	0.833	0.88	0.909
MITx_802x_2013Spring	0.835	0.841	0.851	0.884	0.896
MITx_1473x_2013Spring	0.711	0.78	0.796	0.853	0.885
MITx_3091x_2012Fall	0.916	0.928	0.907	0.934	0.943
MITx_3091x_2013Spring	0.968	0.968	0.941	0.966	0.968
MITx_6002x_2012Fall	0.875	0.863	0.892	0.918	0.899
MITx_6002x_2013Spring	0.921	0.93	0.935	0.943	0.946

Figure 6.7: Normalized performance measure for Active Leaner using different classifiers

6.3 Discussion

Logistic Regression outnumbered the other data mining techniques in the case of Active Learner because target variables, in the form of courses are in the binary form and Logistic Regression performs better in the case of binary data. In Passive Learner's case, single data mining techniques, as well as a combination of Clustering with other data mining techniques were applied, but the classification accuracy of k-means combined with Random Forest only crossed the set threshold limit of 50% in the respective classification.

The performance measure results have shown that Active Learner performance was good as compared to the Passive Learner, which is appeared in Figure 6.8. The explanation for this lies in the way that the Active Learner's dataset was able to capture sufficient information about the learning behavior of the learners as compared to the Passive Learners. So, the recommendations generated in case of Active Learners were much better across all the performance measures, as compared to the less certain information captured for Passive Learners.

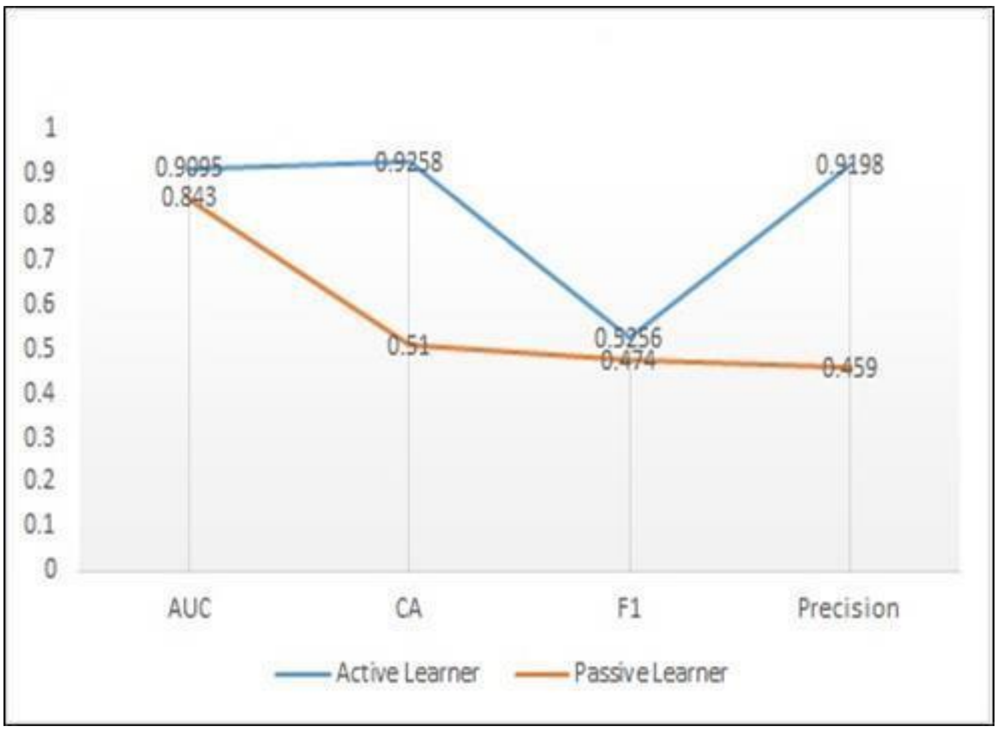

Figure 6.8: Comparison between the performance measures of Active and Passive Learners

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

In this work, the efforts were made to capture the learning behavior of the learners in MOOCs scenarios. In light of the activity of these online learners, the framework classified the learners into two groups: Active Learner and Passive Learner. Different datamining techniques are applied on the Active Learner dataset to get the relevant outcomes. But, with the help of Logistic Regression technique in Active Learner case, along with ample information at the backend, very high prediction accuracy was achieved. Whereas in the case of Passive Learners, where available information was limited, instead of a single Classification technique combination of Clustering and Random Forest was used, which proved to be efficient and considerable.

7.2 Future Work

Some future recommendation in this field can be:

- Passive Learner cluster can be further enhanced by experimenting and implementation of various hybrid data mining techniques. Hybridization of Logistic Regression with K-mean clustering can be a promising future aspect of this work.
- More contextual features can be included like geographic location, which can further help in improving the course recommendation efficiency.
- Other Feature extraction methods such as PCA, has shown efficient results in other recommender system scenarios. So, integration of PCA in the proposed model can lead to better result outcomes.
- The proposed system can also be enhanced by capturing and including current market educational and technological parameters for improving the learner's profile.

[1] K. Hone and G. El Said, "Exploring the factors affecting MOOC retention: A survey study", *Computers & Education*, vol. 98, pp. 157-168, 2016.

[2] J. Bobadilla, F. Serradilla, and A. Hernando, “Collaborative filtering adapted to recommender systems of e-learning,” *Knowledge-Based Systems*, vol. 22.4, pp. 261–265, 2009.

[3] Y. H. Cho, J. K. Kim, and S. H. Kim, “A personalized recommender system based on web usage mining and decision tree induction,” *Expert Systems with Applications*, vol. 23.3, pp. 329–342, 2002.

[4] S. Sun and Q. Chen, “Hierarchical Distance Metric Learning For Large Margin Nearest Neighbor Classification,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25.7, pp. 1073–1087, 2011.

[5] P. Clark and R. Boswell, “Rule induction with CN2: Some recent improvements,” *Lecture Notes in Computer Science Machine Learning — EWSL-91*, pp. 151–163.

[6] L. Breiman, "Random forests," *Machine learning*, vol. 45.1, pp. 5-32, 2001.

[7] W. Lin, S. A. Alvarez, and C. Ruiz. "Efficient adaptive-support association rule mining for recommender systems." *Data mining and knowledge discovery*, vol. 6.1, pp.83-105, 2002.

[8] D. W. Hosmer, S. Lemeshow, and R.X. Sturdivant,” Applied logistic regression.” vol. 398. John Wiley & Sons, 2013.

[9] K. R. Žalik "An efficient k′-means clustering algorithm." *Pattern Recognition Letters*, vol.29.9, pp.1385-1391, 2008.

[10] D. DeBarr and H. Wechsler. "Spam detection using clustering, random forests, and Active learning." *Sixth Conference on Email and Anti-Spam. Mountain View, California*, 2009.

[11] S. R. Gaddam, V. V. Phoha and K. S. Balagani. "K-Means+ ID3: A novel method for supervised anomaly detection by cascading K-Means clustering and ID3 decision tree learning methods." *IEEE Transactions on Knowledge and Data Engineering*, vol. 19.3, 2007.

[12] H. C. Ouertani and M. M. Alawadh. "MOOCs Recommender System: A Recommender System for the Massive Open Online Courses." *Innovations in Smart Learning*. Springer Singapore, pp.137-140, 2007.

[13] V. Garg and R. Tiwari. "Hybrid Massive Open Online Course (MOOC) Recommendation System using Machine Learning." *International Conference on Soft Computing techniques in Engineering and Technology*, 2016.

[14] F. Bousbahi and H. Chorfi. "MOOC-Rec: a case based recommender system for MOOCs." *Procedia-Social and Behavioral Sciences*, vol. 195, pp. 1813-1822, 2015

[15] P. Symeonidis and D. Malakoudis. "MoocRec.com: Massive Open Online Courses Recommender System." *RecSys Posters*, 2016.

[16] S. B. Aher and L. M. R. J. Lobo. "Combination of machine learning algorithms for recommendation of courses in E-Learning system based on historical data." *Knowledge-Based Systems*, vol. 51, pp. 1-14, 2013.

[17] D. Fu et al. "The Undergraduate-Oriented Framework of MOOCs Recommender System." *International Symposium on Educational Technology*, 2015.

[18] H. Zhang, et al. "MCRS: A course recommendation system for MOOCs." *Multimedia Tools and Applications*, pp. 1-19, 2017.

[19] O. R. Zaíane. "Building a recommender agent for e-learning systems." *International Conference on Computers in Education*, 2002.

[20] D. Upendran, et al. "Application of Predictive Analytics in Intelligent Course Recommendation." *Procedia Computer Science*, vol. 93, pp. 917-923, 2016.

[21] F. Castro, et al. "Applying data mining techniques to e-learning problems." *Evolution of teaching and learning paradigms in intelligent environment*. Springer Berlin Heidelberg, pp. 183-221, 2007.

[22] C. Carmona, G. Castillo and E. Millán. "Discovering student preferences in e-learning." *Proceedings of the international workshop on applying data mining in e-learning*, 2007.

[23] A. Elbadrawy and G. Karypis. "Domain-aware grade prediction and top-n course recommendation." *RecSys*, 2016.

[24] C. H. Y. Ji, and R. Duan. "A semantic web-based personalized learning service supported by on-line course resources." *Networked Computing (INC)*, 2010.

[25] T. Denley. "How predictive analytics and choice architecture can improve student success." *Research & Practice in Assessment*, vol. 9, 2014.

[26] R. Farzan and P. Brusilovsky. "Social navigation support in a course recommendation system." *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, 2006.

[27] N. Bendakir, and E. Aïmeur. "Using association rules for course recommendation." *Proceedings of the AAAI Workshop on Educational Data Mining*, vol. 3, 2006.

[28] G. Engin, et al. "Rule-based expert systems for supporting university students." *Procedia Computer Science*, vol. 31, pp. 22-31, 2014.

[29] E. García, et al. "An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering." *User Modeling and User-Adapted Interaction*, vol. 19.1, pp. 99-132, 2009.

[30] T. Shi and S. Horvath. "Unsupervised learning with random forest predictors." *Journal of Computational and Graphical Statistics*, vol.15.1, pp. 118-138, 2006.

[31] A. D. Ho, J. Reich, S. O. Nesterko, D. T. Seaton, T. Mullaney, J. Waldo, and I. Chuang, “HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013,” *SSRN Electronic Journal*.

[32] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina et al. "Orange: data mining toolbox in Python." *The Journal of Machine Learning,* vol. 1, pp. 2349-2353.

[33] A.Gupta , “Applying Data Mining Techniques in Job Recommender System for Considering Candidate Job Preferences,” M.E. thesis, Dept. Com. Sci., Thapar Univ., Patiala, India, 2014.

PUBLICATION

H. Jain and A. "Applying Data Mining Techniques in MOOC Recommender System for Generating Course Recommendations." *International Conference on Advanced Computing and Intelligent Engineering*, 2017. (**Communicated**).

VIDEO PRESENTATION LINK

- https://youtu.be/kRKWSSKpORg

PLAGIARISM REPORT

ORIGINALITY REPORT

%5	%2	%4	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Angiuli, Olivia, Joe Blitzstein, and Jim Waldo. "How to de-identify your data", Communications of the ACM, 2015. Publication	%2
2	O.R. Zaiane. "Building a recommender agent for e-learning systems", International Conference on Computers in Education 2002 Proceedings CIE-02, 2002 Publication	%1
3	rspa.royalsocietypublishing.org Internet Source	<%1
4	"Construction on Learning Analytics Object for Sharing and Interoperation of Educational Big Data", Lecture Notes in Computer Science	<%1