International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 9 Issue 6 ǁ 2021 ǁ PP. 64-69

Prediction of Campus Placement Using Data Mining Algorithms-Random forest, j48 and REPTree

N.Premalatha[1] (PhD scholar), Dr.S.Sujatha[2] (Associate professor & Head)
[1]*(Dept of Computer Science, Dr.G.R.Damodaran College of Science (Autonomous),Coimbatore)*
[2]*(Dept of Computer Science, Dr.G.R.Damodaran College of Science (Autonomous),Coimbatore)*

ABSTRACT: *Data Mining is "the procedure of extracting beneficial information from a big scale information set". It is a effective device to be considered best within the field of education. Educational facts mining includes the new techniques and its strategies for coming across the expertise through reading the database sets to support the selection making procedure in academic organization. It translates an powerful technique for mining the students' performance primarily based at the database sets to expect and analyze whether a student (he/she) will be recruited or no longer within the campus placement. The placement of a student now not only relies upon on his instructional abilities however additionally includes the attributes which include co-curricular sports, communication abilities and so forth. In this paper the usage of these datasets and attributes, predictions are made the usage of the Data Mining Algorithm "Random forest","j48" and REPTree. The results obtained from each approaches are then compared with respect to their "performance", "accuracy" and "evaluation of class parameters" levels by graphical analysis and hence the decisions are made toward the high-quality prediction within the campus placement.*

KEYWORDS: *Data Mining, Educational Data Mining, random forest, j48, REPTree.*

Date of Submission: 25-05-2021 Date of acceptance: 07-06-2021

I. INTRODUCTION:

Placements are taken into consideration to be very critical for every and every university. The basic fulfillment of the university is measured with the aid of the campus placement of the scholars. Every student takes admission to the colleges by means of seeing the proportion of placements in the college. Hence, in this regard the method is about the prediction and analyses for the location necessity in the faculties that enables to build the colleges as well as students to improve their placements. The model is built via the usage of the information mining strategies. The algorithms used for building the model are "random wooded area" and "J48","REPTree''. The performance/accuracy of every version is visualized and tested and based totally on the performance evaluation, every model results are mentioned.

A. Introduction to Data Mining:

Data Mining is the technique of extracting useful facts from huge scale dataset. In different phrases, Data Mining is the method of mining know-how from structured and unstructured records. It is also called understanding discovery technique from huge unstructured records. Data Mining is the vital step inside the system of information discovery (KDD). The following are the diverse steps worried inside the know-how discovery procedure:

- Cleaning the Data set: Here, the process is to cast off the noise and inconsistent information.
- Integration of Data: Here, a couple of statistics sources are incorporated.
- Selection of Data: From the database, the statistics relevant to the undertaking are retrieved in this step
- Transformation of Data: It is a system to carry out aggregation or summary obligations, i.e., information may be transformed into the paperwork that are appropriate for KDD.
- Mining the Data: In this stage, to extract beneficial records styles various intelligent techniques are applied.

B. Introduction to Educational Data Mining:

The uses of Data mining strategies in the training environment are called as Educational Data Mining (EDM). EDM is described via the International Educational Data Mining Society as "an unindustrialized discipline, concerned with growing methods for exploring the only varieties of information that come from instructional settings and using those strategies to higher apprehend pupils/college students and the settings which they collect in".

II. METHODOLOGY

The architectural design consists of the various strategies used in the data mining process. The detail of the system model is presented in the Figure 1.

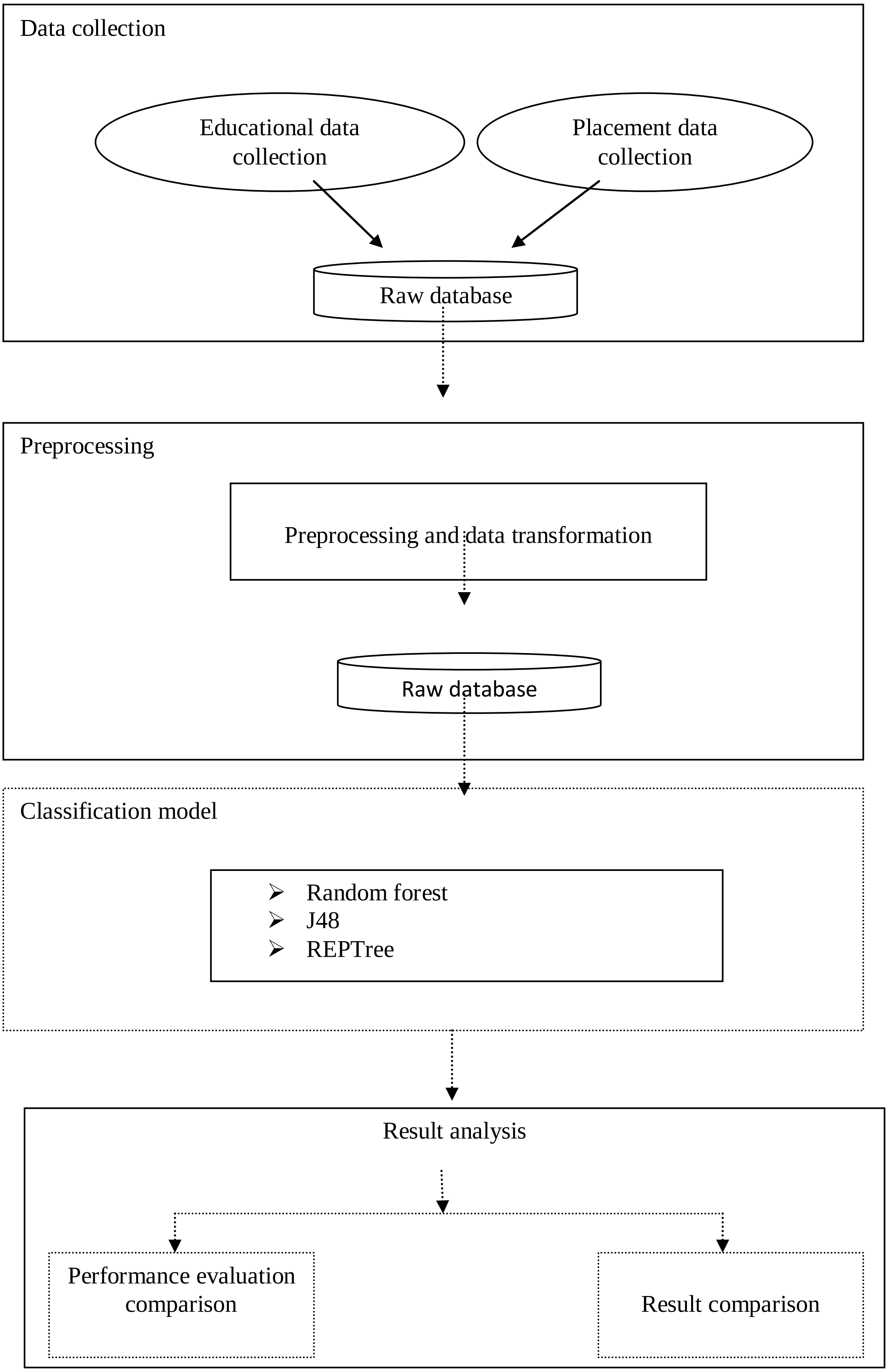

Figure.1 System Architecture

In the first section of this model, the student's information set may be amassed from the educational establishments. The statistics set is then cleaned and pre-processed manually by means of verifying all the attributes entries and making changes using Microsoft workplace excel. The various records mining strategies are implemented to discover the know-how. To expect pupil's placement records, the subsequent algorithms could be used to construct the prediction version

- Random forest
- J48
- REPTree

In the final segment of the version, the classification result is analysed and as compared as a step inside the system of KDD (knowledge discovery). Thus, the accuracy of every model is presented as a very last final results step. In the first phase of this model, the scholar's records set will be amassed from the academic institutions. The records set is then wiped clean and pre-processed manually via verifying all the attributes entries and making adjustments the use of Microsoft workplace excel. The numerous information mining strategies are implemented to find out the know-how. To predict pupil's placement statistics, the subsequent algorithms could be used to construct the prediction version

- Random forest
- J48
- REPTree

In the ultimate phase of the model, the classification result is analyzed and compared as a step within the manner of KDD (know-how discovery). Thus, the accuracy of each version is presented as a final final results step.

III. IMPLEMENTATION

A. Data Collection:

Data set entails the records of the scholars accumulated from the instructional institutions. The facts consists of academics and personality development skills of the scholars. Here the attributes taken into consideration are Gender, percentage_SSC, board_SSC, board_CBSC, Board_ICSE, Percent_PTC, Specialization_PTC, marks_ communication, mark projectwork,placement,profits of the student. So right here statistics units of 120 college students are taken into consideration. In which 310 data units are used for education set which might be used for constructing the model and ultimate information sets are used as trying out records for validating the model.

B.Pre-Processing Module: After series of statistics set, it's far essential to pre-technique the information set. Pre-processing is an crucial section in facts mining and the dataset need to be pre-processed earlier than making use of the records mining algorithm. The pre-processing obligations include cleansing, transformation and integration. The information in the dataset is cleaned and pre-processed manually with the aid of checking the attributes entries. The modifications are made using the Microsoft excel format. This excel layout sheet is saved in Comma Separated Value (CSV) layout. The attributes taken into consideration are Gender, percentage_SSC, board_SSC ,board_CBSC, Board_ICSE, Percent_PTC, Specialization_PTC,marks_communication ,mark projectwork,placement,salary.

C. Classification Module: Classification of statistics is a phase manner. In phase one that is referred to as education segment a classifier is constructed using education set of tuples. The second phase is the type section, where the trying out set of tuples is used for validating the model and the performance of the model is analyzed. The algorithms to carry out such evaluation and validation are random woodland,j48 and REPTree. 1.Random forest is a supervised gaining knowledge of algorithm that is used for each type in addition to regression. But however, it's miles in particular used for class problems. As we recognize that a forest is made of bushes and more bushes approach more sturdy wooded area. Similarly, random forest set of rules creates selection bushes on statistics samples after which gets the prediction from each of them and sooner or later selects the fine answer by vote casting. It is an ensemble method which is better than a single decision tree because it reduces the over-becoming by way of averaging the end result.2.C4.Five set of rules is a category set of rules generating decision tree primarily based on facts concept.C4.5 is from Ross Quinlan (recognized in Weka as J48 J for Java).3.REPTree: algorithm is a quick decision tree learner it is also based totally on C4. Five set of rules and may produce class (discrete final results) or regression bushes (non-stop outcome). It builds a regression/choice tree the use of records gain/variance and prunes it the use of decreased-errors pruning (with returned-fitting).In this, the schooling statistics set is saved, so that for a new unclassified document a classification may be detected through evaluating it to the maximum associated/comparable information in the training set. The models output the prediction consequences of the pupil's placements and the accuracy of every version is calculated and performance of every model is as compared in terms of accuracy where accuracy is the percentage of trying out set examples efficiently labled through the classifier. The two fashions output the prediction consequences of the scholar's placements and the accuracy of each model is calculated and performance of every model is compared in phrases of accuracy wherein accuracy is the percentage of checking out set examples successfully classified by using the classifier.

IV. RESULTS AND ANALYSIS:

The two approaches data set used for is further splitted into two sets consisting of two third as training set and one third as testing set. The approaches use the 310 data sets as the testing sets and the attributes include Gender, percentage_SSC, board_SSC, board_CBSC, Board_ICSE, Percent_PTC, Specialization_PTC,marks_communication, mark projectwork,placement,salary. and the fields are actual result and prediction result. The efficiency of the three approaches is compared in terms of the accuracy and evaluation of class parameters.

The accuracy of the prediction model/classifier is defined as the total number of correctly predicted/classified instances. Accuracy is given by using following formula:

Accuracy= TP+TN/ TP+FN+FP+TN*100

Where TP, TN, FN, FP represents the number of true positives, true negative, false negative and false positive cases.

CORRECTLY AND INCORRECTLY CLASSIFIED INSTANCES ANALYSIS FOR RANDOM FOREST: When the test data is submitted to the random forest model in percentage split correctly classified instances is 78 and incorrect classified instances is 55.and cross validation correctly classified instances is 213 and incorrect classified instances is 178.

Test	Parameters	Correctly classified instances	Incorrect classified instances
Percentage split	Random forest	78	55
Cross validation	Random forest	213	178

Table.1 classification Results for random forest

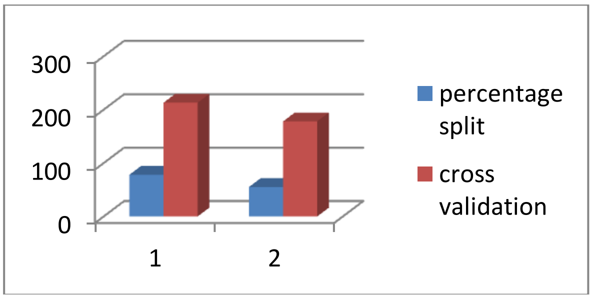

Figure.2 Performance Analysis for random forest

CORRECTLY AND INCORRECTLY CLASSIFIED INSTANCES ANALYSIS FOR J48:
When the test data is submitted to the j48 model in percentage split correctly classified instances is 76 and incorrect classified instances is 57.and cross validation correctly classified instances is 217 and incorrect classified instances is 174

Test	Parameters	Correctly classified instances	Incorrect classified instances
Percentage split	J48	76	57
Cross validation	J48	217	174

Table.2 classification Results for J48

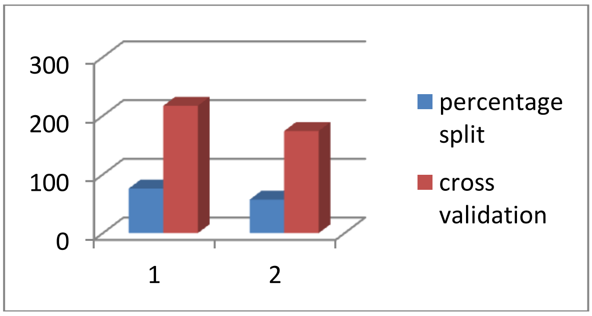

Figure.3 Performance Analysis for J48

CORRECTLY AND INCORRECTLY CLASSIFIED INSTANCES ANALYSIS FOR REPTREE: When the test data is submitted to the REPTree model in percentage split correctly classified instances is 76 and incorrect classified instances is 57.and cross validation correctly classified instances is 218 and incorrect classified instances is 173.

Test	**Parameters**	**Correctly classified instances**	**Incorrect classified instances**
Percentage split	**REPTree**	**76**	**57**
Cross validation	**REPTree**	**218**	**173**

Table.3 classification Results for REPTree

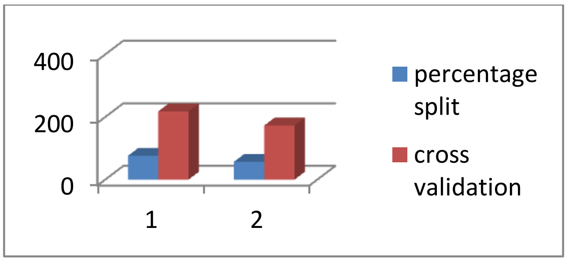

Figure.4 Performance Analysis for REPTree

From above figures and tables the percentage split in random forest algorithm gives the best performance because correctly classified instances is 78 with the accuracy value 58.6466% compared with other algorithm.

Parameters	**Accuracy of percentage split**	**Accuracy of cross validation**
Random forest	58.6466	54.4757
J48	57.1429	55.4987
REPTree	57.1429	55.7545

Table.4 Accuracy

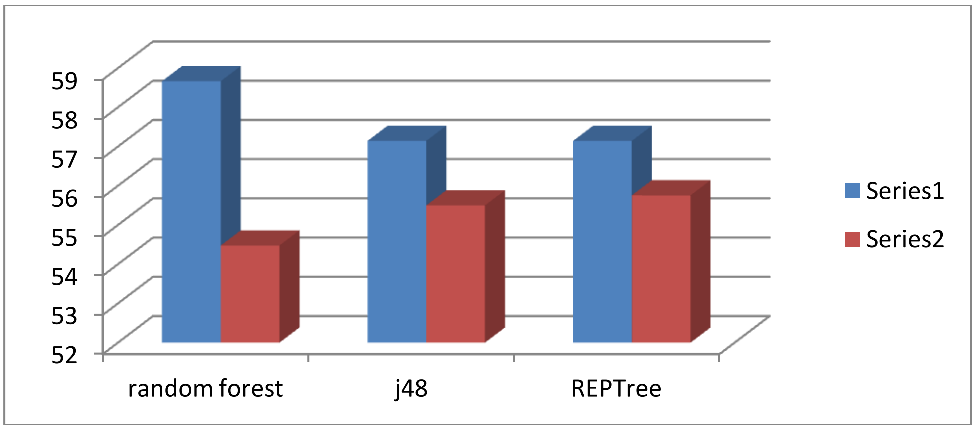

Figure.5 Comparison of accuracy

From the above figure random forest algorithm gives the best result for placement prediction in percentage split test because the accuracy level is 58.6466% compared with the other algorithms accuracy value.

E.**COMPARISON OF CLASS PARAMETERS:**

Class	TP RATE	FP RATE	PRECISION	RECALL	F-MEASURE
Computer networking	0.5	0.312	0.538	0.5	0.519
IT	0.704	0.484	0.625	0.704	0.662
Computer engineering	0	0.008	0	0	0

Table.5 Class parameters

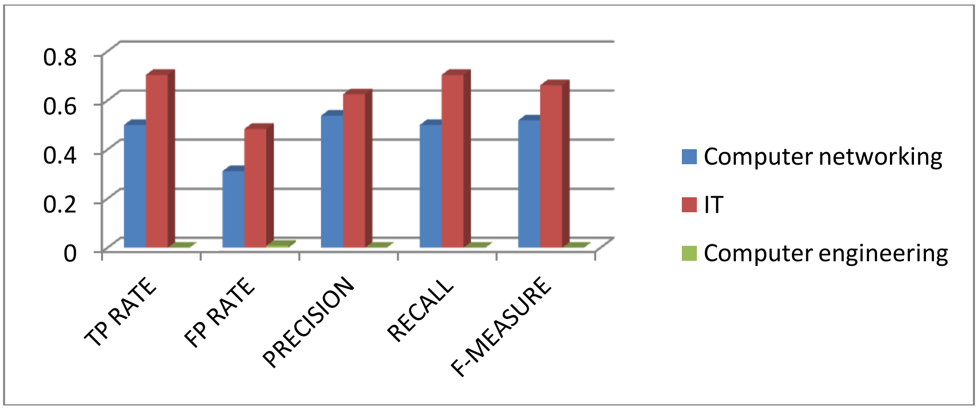

Figure.6 Class parameter comparison

Here percentage split with random forest algorithms class parameters have taken and analysed in this IT students get more placed in placement compared with other department.

V. CONCLUSION:

The campus placement pastime is very much vital as institution point of view in addition to scholar point of view. In this regard to improve the scholar's overall performance, a work has been analysed and anticipated the use of the algorithms random forest, J48and the REPTree algorithms to validate the techniques. The algorithms are applied at the information set and attributes used to construct the model. The accuracy acquired after analysis for random forest algorithm in percentage split accuracy is **58.6466%.** Hence, from the above stated evaluation and prediction it'd be better if the random forest is used to predict the placement result.

REFERENCES

[1]. "Student Placement Analyzer: A Recommendation System Using Machine Learning" 2017 International Conference on advanced computing and communication systems (ICACCS-2017), Jan 06-07,2017, Coimbatore, INDIA.

[2]. "Prediction Model for Students Future Development by Deep Learning and TensorFlow Artificial Intelligence Engine" 2018 4th IEEE International Conference on Information Management.

[3]. "Data Mining Approach for Predicting Student and Institution's Placement Percentage", Professor. Ashok M Assistant Professor Apoorva A ,2016 International Conference on Computational Systems and Information Systems for Sustainable Solutions

[4]. "Student Placement Analyzer: A Recommendation System Using Machine Learning", Senthil Kumar Thangavel, Divya Bharathi P, Abijith Sankar, International Conference on Advanced Computing and Communication Systems (ICACCS -2017), Jan. 06 - 07, 2017, Coimbatore, INDIA

[5]. "A Placement Prediction System Using K-Nearest Neighbors Classifier", Animesh Giri, M Vignesh V Bhagavath, Bysani Pruthvi, Naini Dubey, Second International Conference on Cognitive Computing and Information Processing (CCIP), 2016.

[6]. "Class Result Prediction using Machine Learning", Pushpa S K, Associate Professor, Manjunath T N, Professor and Head, Mrunal T V, Amartya Singh, C Suhas, International Conference On Smart Technology for Smart Nation, 2017 [