An Intelligent Student Job Prediction using Educational Data Analytics

Sarvesh T S, T H Sreenivasa, Ambika V

Abstract: Mining of Educational data is an emerging field focused in data mining field to uncover required facts within educational data in order to assist educational institutions to increase their management design also student facilities. It provides essential knowledge about imparting the education, which is used to enhance the quality of teaching and learning. The implementation of the proposed system dataset provides details with respect to old students data. Mining of Educational data is implicated within data mining field to find the required facts inside educational data to assist institutions, to increase the management design along with learner facilities. The present study comes up with applying data science techniques over educational data. Association rule used within student's data to find some facts for assisting management design. Data Science algorithms implemented by considering grade of courses, also graduated student employment information for job prediction after completion of education. The outcome of this study gives better knowledge for student management design and prediction of job. The main objective of the proposed system is to find the correlations between the student educational parameters with the types of the job.

Keywords: Association Rule; job prediction; classification; Machine Learning

I. INTRODUCTION

The primary objective of learners admitted to professional branches in higher educational colleges is to obtain a job with greater income in a deemed organization. Professional teaching provides the education for either technical or managerial section as well. A technology course delivers technical tuition to learners in different fields such as Thermal Engineering, Information Technology, Computer Science, Electronics and Communications etc. This is targeted at producing learners' professional in advanced academic also experimental knowledge in different engineering courses. [9] The status of job prediction that learners are very likely to reach will assists learners to put more effort to get suitable betterment in venturing into a profession in different specialized fields [1]. It also assists the tutors and more inside an educational institution to gives suitable control to the

Sarvesh T S*, Student, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India.

Dr. T H Sreenivasa, Professor, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India.

Prof. Ambika V, Assistant Professor, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India.

advancement of learners within the period of course. The percentage value of placement is an important factor in raising the esteem of the institution. Therefore educational institutions need such a set-up has a remarkable position in some higher learning institutions.

II. LITERATURE SURVEY

The technology occurrence interrupt inside different areas has managed unexpected speed of generated data and various format of huge volume of data. These data generation method produces big data with volume, various and velocity. The appropriate approach for study of big data is required to extract the useful knowledge. The detection of required fact from huge group of data can be achieved using data mining technique. The main purpose of data mining is applying many algorithms to find pattern and discover useful knowledge which is masked inside dataset. Educational Data Mining exists in a recent research process inside data mining [2].

The major goal of Educational Data Mining is to detect required facts and mining required patterns in the academic data. The academic data contains enrolment of student's data, student profile, student job details and other live data of students available in the institutions [3].

In 2015, N. Buniyamin and Usamah bin Mat et al. [4] proved in their analysis that the Neuro-Fuzzy application categorizing inside a learners college data among the faculty members of Malaysian public university within electrical engineering. This analysis concluded that the result of the method can predict the likelihood of the learner to bring out outstanding grading although the learner has attained weak in some project or course.

In 2011, S. Pal and B.K. Baradwaj et.al. [5] Performed a study on categorizing function of learner datum to visualize the student's performances in the terminal academic examination. The achievement of student was categorized as four groups. This encourages the faculty to recognize the slow learners, dropouts and students who need specific attention in the form of mentorship.

In 2016, Amjad Abu Saa et.al. [6] Conducted a study using numerous data mining functions in order to build subjective expectation protocol to anticipate the learners grading within educational dataset. Author implemented four algorithms about decision tree method. The results of this study motivated the Universities and concluded that the performing data mining task on students by the University aids in improving learners achievements.

In 2017, E. Mnkandla and F. Matsebula et.al. [7] performed a analysis and suggested a planning to big data analytics in advanced education.

The planning consisted of many parts namely, Data assembly instrument that collects students data through numerous source of data (learners identity board, community gathering and learners data structure, Information depository and management structure that composes of big data administration, Data analytics mechanism which execute algorithms using particulars, Imagination of data that aid in making the decisions with working structure for delivering awareness, caution or directing learner or executives.

In 2017, Anindya Ojha et al. [8] analyzed the historical student placement data of Software Engineering department of SRM University from the year 2014-2016. In this analysis Linear regression has been implemented in working on the data and assesses the collected data using numerous parts that is CGPA, backlog history, apprised strengths, syllabus, and job training methods etc. The collected data has been correlated and resultant correlation was considered to produce useful predictions and suggestion to improve method of recruitment in place. Linear regression, decision trees and random forest algorithms were utilized to study the data to show correlations and main factors in student's appointment.

In 2017, Senthil Kumar Thangavel et al. [9] implemented machine learning techniques to analyze the placement of students by extracting the knowledge from running and old data that resides inside the databases of the institutions. Author presented a proposal method that predicts the students to posses unique among the five statuses of recruitment. Above mentioned five statuses of recruitment are Core Organization, Mass employment services, Dream Organization, Not suitable and Not focused in positioning.

In 2019, Munirah Alghamlas et al. [10] suggested a solution to assessing the call for employment, utilizing extracted data from Saudi employment market real time employment websites. Author utilizes machine learning algorithms in forecasting the eligibility of Information Technology students' ability for the employment inside Saudi employment market. With this background present project was designed with the aim to find out the effectiveness of Future job prediction by Data science technology. To evaluate the students job prediction, and correlate the student parameters with job types.

III. METHODOLOGY

The suggested model is as shown in figure 1. It shows the database from which the data was extracted. The dataset contains the students trained data. The system uses mining of data procedure to detect the facts within the dataset. Association rule algorithms are implemented to detect the relationship between the items. Apriori, ECLAT and SFIT algorithms were implemented in the system to generate the strong rules and predict the Jobs. The performance of algorithms calculated and displayed using graphical charts.
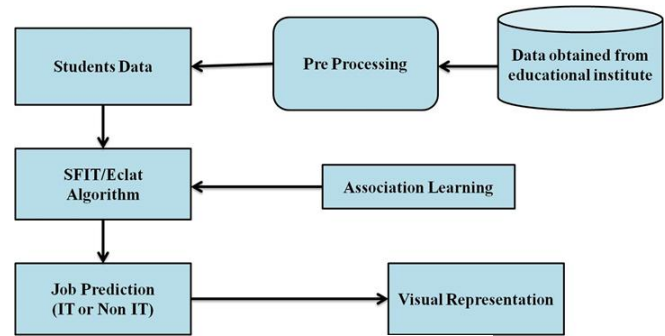

Figure1. Proposed Model

The main users are administrator, student and visitor. The administrator has the full control over the system. The administrator can add, modify and delete students in the system and provides them access control to the system by registering them to the system. All the students of the institutions are registered to the system during the admission to the institution. Since users are added by the administrator the users are authorized easily as the students of the system. The administrator can access job prediction patterns and graphical representation of algorithms performance for the given inputs. The administrator can view the student's queries which are pending and already answered. The administrator can also send the response to pending queries. The student is the second actor. The student who wants to access a job details logs into the website can access job prediction patterns. The students can view their queries which are pending and already answered. The students can also post a new query to the system. The third actor is the visitor to the website. The visitor can visit the details about the system. The visitor can have basic access to the system like contact information, team etc. Data mining can be defined as strategy of finding required information from a given dataset. To discover the required information, algorithms of data mining were implemented by utilizing associations rule mining method. Association rule mining implemented predominantly to identify the fully associated relationship in the middle of the parameters that have high level occurrence and strong interrelationship inside the dataset. Association rules help us to discover the parameters that repeatedly arise simultaneously in the system.

The proposed system implements the following algorithms:

1. Apriori: This is the fundamental algorithm for association rule mining and frequent item set mining which identifies the repeated itemset inside the dataset also enlarging the system to huge item sets allowing that those item sets occurs as much as possible in the dataset. Apriori algorithm functions like a horizontal mode go along with the Breadth-First Search mechanism in the graph. The pseudo code of Apriori algorithm is as follows:

1. Search the dataset and calculate the support(s) for each item
2. Create L1 (i.e. frequent one Itemset).
3. Utilize Lp-1, join Lp-1 to Create the set of candidate p item set.

4. Examine the candidate p item set and create the support of each candidate p item set.
5. Append to frequent item set, until C=Null set.
6. Generate all non empty subsets for each item in the frequent item set:.
7. For each non empty subset find the confidence. Append to Strong Association Rule if confidence ≥ specified confidence value

2. ECLAT Algorithm: The ECLAT stands for Equivalence Class Clustering and bottom-up Lattice Traversal. It is a better structured and extensible kind of the Apriori algorithm. Apriori algorithm functions like a horizontal mode go along with the Breadth-First Search mechanism in the graph, the ECLAT algorithm functions in a vertical manner like the Depth-First Search mechanism in the graph. This vertical process of the ECLAT algorithm arrive a faster algorithm compare to Apriori algorithm. Pseudo code of Apriori algorithm is as shown below.

1. Search the data-set and for each item find the support count(s). Add the transaction ids instead of specifying the actual support.
2. Create L1 (Frequent one item set) by comparing with minimum support count.
3. Create the candidate p- item set by utilizing Lp-1, join Lp-1 .
4. Traverse through the candidate k item set and create the support of all candidate p – item set. When we find support count of candidate items, we compare with previous step, no need to scan the data base again and compare with original dataset.
5. Uuntil C=Null Set, append to frequent item set,
6. Generate all non empty subsets for each item in the frequent item set.

7. Find the confidence for each non empty subset. Append to Strong Association Rule if confidence ≥ specified confidence value.

3.SFIT Algorithm: The SFIT algorithm stands for Set operation for Frequent Item set using Transactional database. It is a much better structured and extensible kind of Apriori algorithm. While Apriori algorithm functions like a horizontal mode go along with the Breadth-First Search mechanism in the graph, and ECLAT algorithm functions in a vertical manner like the Depth-First Search mechanism in the graph. SFIT algorithm functions by applying set operations for the frequent item sets. SFIT algorithm arrive faster algorithm compare to ECLAT algorithm.

Search the data-set and for each item find the support count(s) (both support and transaction ids).Add the support and transaction ids instead of specifying the actual support.

$$Support(Itemset) = \frac{Itemset}{Transactions}$$

$$Support(A\text{->}B) = \frac{No\ of\ Transactions\ containing\ both\ A\ and\ B}{Total\ No\ of\ Transactions}$$

1. Create L1 (Frequent one item set) by comparing with minimum support count.

2. Create L1 (Frequent one item set) by comparing with minimum support count.
3. Create the candidate p-item set by utilizing Lp-1, join Lp-1.
4. Traverse through the candidate p item set and create the support of all candidate p – item set. When we find support count of candidate items, we compare with previous step, no need to again scan the data base and compare with original dataset.
5. Until C=Null Set, append to frequent item set,.
6. Generate all non-empty subsets for each item in the frequent item set:
7. For each non empty subset find the confidence. Add to Strong Association Rule if confidence ≥ specified confidence value

Support value of an association pattern is the rate of Itemset measured in percentage, inside the inclusive Transactions for which the pattern is true.

Minimum Support value is the least value for considering repeated Itemset.

If we consider minimum support rate is 50%, a repeated Itemset is the single which occur in no less than of 50% of the entire transactions in the dataset.

The support count value of an association pattern is the Frequency of occurrence of an Itemset in the overall Transactions for which the pattern is true.

If minimum support count is 5 of 20 Transactions, a frequent itemset is the single that appears in at least 25% of the entire transactions.

E.g. Suppose Itemset Y is a repeated itemset if: support count (Y) >= minimum support count value

Confidence is defined as the compute of sureness or trustworthiness associated with each identified pattern.

$$Confidence(X \to Y) = \frac{Support(X \wedge Y)}{Support(X)}$$

$$Confidence(A\text{->}B) = \frac{No\ of\ Transactions\ containing\ both\ A\ and\ B}{No\ of\ Transactions\ containing\ A}$$

Certainty is characterized as the confidence related with each found example.

Association rules are generated by selecting frequent item sets and minimum confidence value.

IV. RESULTS AND DISCUSSION

In this system the dataset contains student details having more than thousand transactions. The parameters used in the dataset are SSLC, PUC, CPP, JAVA, CS, APTITUDE and RESULT. The parameters description and values are as shown in Table 1

Table 1 Student dataset parameters description

Parameters	Description	Values
SSLC	Course A Grading	{High(SSLC_DISTINCTION) Medium(SSLC_FIRST CLASS) Low (SSLC_SECOND CLASS)}
PUC	Course B Grading	{High(PUC_DISTINCTION) Medium(PUC_FIRST CLASS) Low (PUC_SECOND CLASS)}
CPP	Area of Interest 1	{High(CPP_A) Medium(CPP_B) Low (CPP_S)}
JAVA	Area of Interest 2	{High(JAVA_A) Medium(JAVA _B) Low (JAVA _S)}
CS	Area of Interest 3	{High(CS _A) Medium(CS _B) Low (CS _S)}
APTITUDE	Area of Interest 4	{High(APTI_A) Medium(APTI_B) Low (APTI_S)}
RESULT	Job of students	{IT,BUSINESS,BANK, TEACHING}

In this study, to discover the relationship between the parameters of the students, Association rule mining algorithms were used. The association rule provides two major steps. Primary step, creation of frequent item set, is a method to identify frequent item set by considering minimum value of support and confidence greater than designated threshold. Secondary step, creation of rules, exist a method to generate rules utilizing frequent item set. Using the minimum value of confidence the generated rules are assessed.

Table 2, Table 3 and Table 4 contains the Minimum support and Minimum confidence values that are tested in this project

Table 2 Minimum support and minimum confidence values for Apriori algorithm

Minimum Support	0.102	10.2% of frequent item set occurs in the student dataset.
Minimum Confidence	0.28	28% of rule has been observed to be true

Table 3 Minimum support count and minimum confidence values for ECLAT algorithm

Minimum Support count	108.7	10.2% of transactions occur in the student dataset. (i.e. Min Sup Count=0.102 * Total transactions in the dataset)
Minimum Confidence	0.28	28% of rule has been observed to be true

**Table 4 Minimum support count and minimum confidence values for SFIT algorithm**

Minimum Support count	108.7	10.2% of transactions occur in the student dataset. (i.e. Min Sup Count=0.102 * Total transactions in the dataset)
Minimum Confidence	0.28	28% of rule has been observed to be true

In the project, the rules generated using the minimum values of support and confidence by the system are shown in Figure 2

Rule X	->	Rule Y	Confidence
APTI_S	->	TEACHING	42.62%
APTI_S,PUC_FIRST CLASS	->	TEACHING	100.00%
APTI_S,SSLC_DISTINCTION	->	TEACHING	100.00%
CPP_B	->	BANK SECTOR	100.00%
CS_S	->	BANK SECTOR	89.02%
JAVA_A	->	IT	94.26%
JAVA_A,JAVA_S	->	IT	100.00%
JAVA_S	->	IT	52.75%
PUC_FIRST CLASS	->	TEACHING	91.40%
SSLC_DISTINCTION	->	TEACHING	88.89%

Figure 2 Rules generated by the algorithms

Performance calculated with respect to the running time in milliseconds of Apriori, ECLAT and SFIT algorithms by the system displayed using bar chart are shown in Figure 3
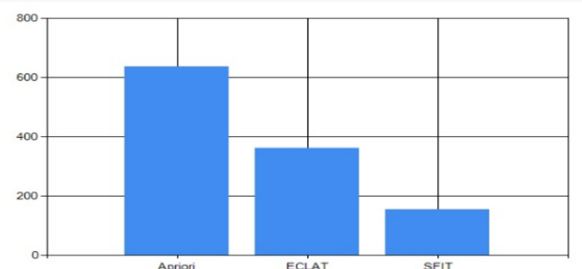
In the present project, three algorithms were used for Future job prediction. Among the three algorithms, SFIT showed more efficiency and fast rate of prediction in comparison to Apriori and ECLAT. In previous studies, authors have used only one algorithm to predict the future job opportunities or performances, the peculiarity of the present study is, we have used three algorithms and we have compared the performance of all three algorithms.

V. CONCLUSION

In the present study, we implemented the algorithms of data mining to find the required facts inside the academic dataset. Primarily, association rule method is utilized on dataset of students to response the query "how data mining can help student management working process". Outcome obtained from algorithms gives an important relationship in the middle of student parameters and job types. This result might be helpful for the educators, management personnel to plan their admission promotion.

The algorithms is implemented for students dataset having academic grading and job details to response the query "how data mining can predict the student's jobs". Then obtained outcome rule will predict the important areas to be chosen by the students that determines their future career.

FUTURE WORK

SMS/Email Module – In the present project, the admin was authorised to assign the Id and Password to the students and was intimated manually. So for future enhancement, we can add SMS/Email module, where the students receive an SMS or Email regarding the Id and Password.

Additional Parameters – In the present project, we have used few student parameters for job prediction, we suggest including various additional parameters like Post Graduation, Python etc, for better accuracy.

REFERENCES

1. Harinath S. Aksha Prasad, Suma H S, Suraksha A, Tojo Mathew, Student placement prediction using machine learning. International Research Journal of Engineering and Technology (IRJET). 2019.6(4):4577-4579.
2. Heikki, Mannila, Data mining: machine learning statistics, and database, IEEE, 1996.
3. Jiawei Han and Micheline Kamber (2011) Data Mining: Concepts and Techniques. 3 editions. Morgan Kaufmann.
4. Norlida Buniyamin, Usamah bin Mat and Pauziah Mohd Arshad. Educational Data Mining for Prediction and Classification of Engineering Students Achievement. IEEE-2015.
5. Brijesh Kumar Baradwaj and Saurabh Pal. Mining Educational Data to Analyze Students Performance, IJACSA- Vol. 2, No. 6, 2011.
6. Amjad Abu Saa. Educational Data Mining & Students' Performance Prediction. IJACSA Vol. 7, No. 5, 2016.
7. F. Matsebula and E. Mnkandla. A big data architecture for learning analytics higher education. IEEE Africon, 2017, pp.951-956.
8. Anindya Ojha, Upasana Pattnaik and Sasi Rekha Sankar Data Analytics on Placement Data in a South Asian University, ICECDS-2017
9. Senthil Kumar Thangavel,Divya Bharathi P and Abijith Sankar. Student Placement Analyzer: A Recommendation System Using Machine Learning, ICACCS -2017.
10. Munirah Alghamlas, Reham Alabduljabbar. Predicting the Suitability of IT Students' Skills for the Recruitment in Saudi Labor Market, 2019,IEEE.

AUTHORS PROFILE

Sarvesh T S is a post graduate student pursuing his M.Tech degree from Vidyavardhaka College of Engineering, Mysuru

Dr. T H Sreenivasa is working as professor in the department of Computer sciemce and engineering. His research interests are Internet of Things, Networking.

Prof. Ambika V is working as an assistant professor in the department of Computer sciemce and engineering. Her research interests are Internet of Things, Networking. .