Salary Prediction Using Machine Learning

Krishna Gopal[1], Ashish Singh[2], Harsh Kumar[3], Dr.Shrddha Sagar[4]
[1,2,3]*B. Tech CSE, Galgotias University, Greater Noida 203201, India*
[4]*Professor, Galgotias University, Greater Noida 203201, India*

Abstract **- Our salary prediction system is aimed toward providing better assistance to the school students regarding the salary that they will aspect after completing their course. Not only they are going to be ready to get a thought of their deserving salary but also, they will get to understand about the talents that they have to satisfy their professional goals. this may enhance the motivation of scholars who are enrolled in education institutes and supply better assistance also. Through this paper we have tried to provide a system for salary prediction during which data processing technique is employed. during this system, the profile of student is going to be compared with graduated student. we have used data mining techniques for comparison as they perform best. we have also performed an experiment on student data set using 10-fold cross-validation.**

Index Terms **- Salary prediction, salary, raise, salary increment, job, professional growth.**

I.INTRODUCTION

Nowadays as we will see, data mining is becoming latest trend within the computer sector which involves finding beneficial patterns and knowledge from systems. Under data processing, educational data processing is booming also which involves extracting useful information from educational data system like student admission, course registration, course management system and lots of more systems regarding students at various levels of institute from school to college levels.

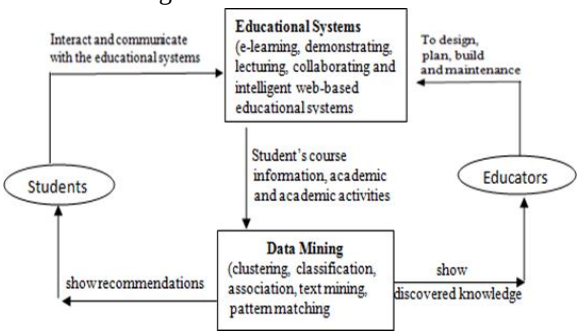

Figure.1

The major focus of educational data mining is to assist the institutes to arrange their students far better and aid them to reinforce their performance. Various machine learning concepts are applied to review the data collected from learning. Figure1 shows the picturing of application of educational data mining system.

As now seen, many students select their course on the idea of trend, or they select course on the idea of their peer suggestion. The impact of all such is that the performance of scholars goes to pot and this have a tendency to throwing in the towel from institutes. The performance of such students can be enhanced if we offer them with samples of their college graduates as they need already taken very same path and that they may need faced same issues. Their career and aid could help the scholars to be motivated for his or her courses and be focused.

The salary rate of passed out students will function an interesting point which can further help the scholars presently enrolled within the course to maneuver forward in their career. The salary rate however depends on the prestige of school name, student score in academics and other activities that he/she does during college times. So, to predict the salary, records of graduated people are used as a reference for salary rate. This model is often implemented as a tool to point out the way to achieve different salary brackets by taking reference of graduated students by considering their profiles as they were during their college durations.

II.LITERATURE REVIEW

K.Lakshmi and A.Parkavil [1], performed analysis in order that they might understand the extent of students (their knowledge) by means of associative rule mining. They applied association rule mining alongside linear regression to observe the group (or cluster) to which the student knowledge is that the closest to.

An important concept to provide efficient and effective learning is data mining for researchers [2]. Different kinds of recommendations are produced by recommender systems. The concept of data mining is primarily used for studying the patterns of learners. The efficiency of the system can be enhanced by personalizing the system for the students [3].

Pokpong Songmuang along with Pornthep Khongchai [4] proposes system of salary prediction to enhance the motivation of students in college. A seven-feature prediction model was generated by using the technique of decision tree. With more feedbacks from faculties and staffs the performance of the system was improved. Jobs which are related to the field of study of the students was also included in the paper to enhance the performance of the system. In [5], Bayesian network is used for classifying students based on marks scored by them. The cross validation uses for evaluating the model is 10-fold. In [6] Ordinary least squares regression model is used to create models for predicting salaries of students on the basis of profiles and family background.

A hierarchal linear regression is used by Karla et al [7] for building a model by taking students profile as fixed parameter and salary as output variable. However, this model has majorly two problems, one is this system is personalized for students as it predicts salary of students group, second is the output of the model requires immense statistical knowledge to understand the predictions.

In research [8] by Rajveer Singh conducted a study for salary estimation for entry-level Indian engineering graduates indicates that the academic performance in school and college, school affiliation, college reputation are important indicators for starting salary. C.-C. Hung, E.-P. Lim [9] in their paper proposed "Company, Occupation, Company" (COC) model to derive unbiased salaries by aggregating job review and job post data. This model can predict unbiased salaries, companies' inflation, and competitiveness effectively.

III. SALARY PREDICTION SYSTEM

This proposed model is directed to predict the salary of graduating students by comparing their academic performance and additional activities with the successful students who have graduates from same streams. If student has done any certification course, training, internship, or the other specialization then it'll make the resume of the scholar stronger form others. Through this technique students can focus the areas where they ought to focus to reinforce their skills for better opportunities. For prediction system data mining classification is that the core technique. For prediction data of graduated students along with their salary are going to be used for training and on the idea of it, our model will predict salary for current students.

A. System Design

This system needs 2 input data:

1-Profiles of student

2-Graduated student's profiles with their salaries.

This system will use k profiles of former graduates whose profiles are similar to students presently exemplifier. Number of features are scaled down by applying feature selection.

Following features are selected for comparison:

- Student Gender(M/F)
- Course department
- Program enrolled
- Certifications or equivalent
- Academics Grade
- Salary

B. Training of Model:

For salary prediction we will apply various data mining techniques, however we have selected 3 classification techniques for comparison which are as follows:

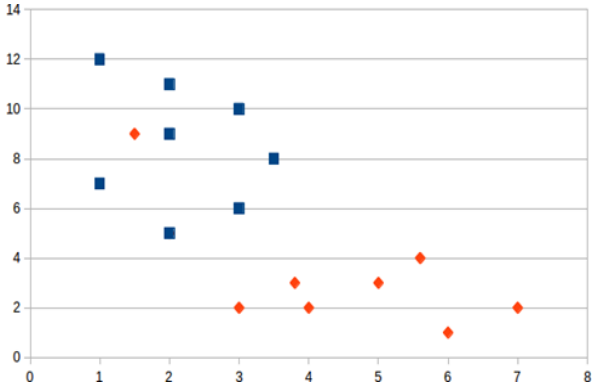

Figure.2

K-nearest neighbors, a classification algorithm uses distance vector for classifying object.

Let us take an example of data points consisting of 2 features shown as blue cubes and red diamonds in Figure 2. the class of nearest point is employed to

classify test data in Figure 3 by using triangles in yellow.

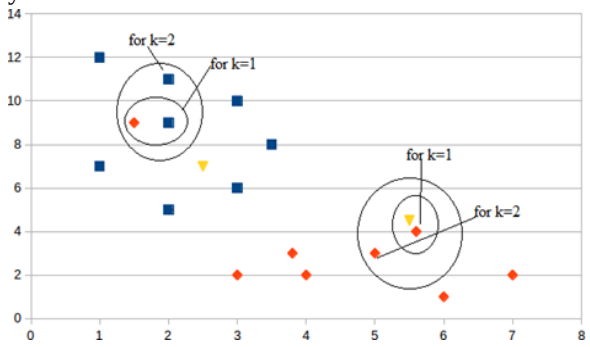

Figure.3

Naïve Bayes is kind of classifier technique which is based on probability. For classifying set it involves high dimensional set. Some examples where Naïve Bayes algorithm is use are classification of new articles, spam messages detection etc. An assumption is made by this algorithm is that some features is not dependent on the occurrence of other features. Such as identifying fruits on the basis of its color, taste, then yellow color, oval shape, sweet flavor fruit is most likely to be a mango. Fast and quick prediction can be made by using Naïve bayes algorithm.

The target of Naïve bayes is to compute conditional probability with vector x1, x2, x3…,xn which belongs to class Ci.

$$P(C_i|x_1,x_2,\ldots,x_n)=\frac{P(x_1,x_2,\ldots,x_n|C_i).P(C_i)}{P(x_1,x_2,\ldots,x_n)} \text{ for } 1\le i\le k \quad (1)$$

We can write the numerator of (1) as:

$$P(x_1,x_2,\ldots,x_n|C_i).P(C_i)=P(x_1,x_2,\ldots,x_n,C_i)$$

$$\begin{aligned}P(x_1,x_2,\ldots,x_n,C_i) &= P(x_1|x_2,\ldots,x_n,C_i).P(x_2,\ldots,x_n,C_i)\\ &= P(x_1|x_2,\ldots,x_n,C_i).P(x_2|x_3,\ldots,x_n,C_i)P(x_3,\ldots,x_n,C_i)\\ &= \ldots\\ &= P(x_1|x_2,\ldots,x_n,C_i).P(x_2|x_3,\ldots,x_n,C_i)\ldots P(x_{n-1}|x_n,C_i).P(x_n|C_i).P(C_i)\end{aligned} \quad (2)$$

By assuming that features are not dependent the conditional probability term become P (xj | Ci). Thus, by above calculation and assuming independence nature, the equation becomes:

$$P(C_i|x_1,x_2,\ldots,x_n)=\left(\prod_{j=1}^{j=n}P(x_j|C_i)\right).\frac{P(C_i)}{P(x_1,x_2,\ldots,x_n)} \text{ for } 1\le i\le k \quad (3)$$

As P is constant for all classes the equation above can be written as:

$$P(C_i|x_1,x_2,\ldots,x_n)\propto\left(\prod_{j=1}^{j=n}P(x_j|C_i)\right).P(C_i) \text{ for } 1\le i\le k \quad (4)$$

The NB identifies the class of feature set x1,x2,x3,…,xn by using (4).

Decision tree(J48) uses tree data structure to represent data set. In tree data structure, the decision nodes are represented by leaf node. Decision tree are of two types: REGRESSION and CLASSIFICATION TREE. The output of regression tree analysis is real number. Decision tree has advantage of providing theoretical framework so that it can take account of not only experimental data but also structural for providing getter capability [10]. Data is portioned into subsets that has similar value objects.

C. Salary prediction system usage:

The goal of this system of salary prediction is to motivate students so that they can be motivated for their course and can extract maximum benefits of this system.

This model motivates the by using

1. Prediction by using profiles of graduated student.
2. Showing example of top three graduated students and their handsome salaries.

Thus, users can understand the working of this model without much statistical knowledge to know his/her salary.

Students will have to fill their corresponding details as an input.

The output of the system is:

```
Enter department (engineering/MBA)
engineering
Enter Program (MECHANICAL/COMPUTER/ELECTRICAL/TEXTILE)
computer
 Wether do Job Training (yes/no)
yes
 Have Certificate (yes/no)
yes
 Enter the CGPA
7.8
Enter the number of the instances you want on your window k =
3

---- Results by KNN Algorithm-----

Department      Program          Job_Training   Certificate      CGPA    Salary

engineering     computer              yes            yes      7.8     32000.0
engineering     computer              yes            yes      8.1     40000.0
engineering     computer              yes            yes      7.0     23000.0
```

Figure.4

IV. EXPERIMENT AND RESULT

For identifying the best technique for graduate student's data let us perform an experiment in which we are comparing performance of three different classification algorithms: Decision tree, Naïve Bayes

and k-NN on data. This data contains department, program, certificate, CGPA and salary class.

For predicting salary, WEKA, a data mining tool is use for the comparison of different mining techniques.

To calculate the efficiency of the model, 10-fold cross validation was used.

Table 1 shows prediction by each of the mining technique.

Precision, F-measure and recall are the performance measures for comparison.

Precision - it is defined as division of number of positive predictions by total number of positive class value predicted.

Recall – it is defined as division of number of positive predictions by number of positive class value in test.

F-measures – it computes test accuracy and harmonic mean of precision and recall results into F-measure.

On observing the results, we get that decision tree(J48) is better than Naïve bayes and KNN as its throughput is better and efficiency is higher.

Table 1.1

CLASS	RECALL (%)		
	KNN	*NB*	*J48*
LOW	30.00	90.00	90.00
MEDIUM	10.70	93.30	93.30
HIGH	44.00	83.30	98.00

Table 1.2

CLASS	PRECISION (%)		
	KNN	*NB*	*J48*
LOW	75.00	81.80	98.00
MEDIUM	78.60	93.30	99.00
HIGH	56.00	83.30	90.00

Table 1.3

CLASS	F-MEASURE (%)		
	KNN	*NB*	*J48*
LOW	42.90	85.70	94.7
MEDIUM	75.90	93.30	96.6
HIGH	65.10	83.30	94.7

V. SUMMARY

This paper provides a system for salary prediction in which data mining technique is used. In this system the profile of student will be compared with graduated student. We have used data mining techniques for comparison as they perform best. We have also performed an experiment on student data set using 10-fold cross-validation. We concluded that decision tree(J48) provides the best result. But KNN will give better performance if featured attribute is less.

REFERENCES

[1] A. K. Lakshmi and A.Parkavil Predicting the course knowledge level of students using data mining techniques, IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials ,2017

[2] A.W. Husain, N.A Rashid and A.M. Shahiri, A Review on Predicting Students Performance using Data mining Techniques Procedia Computer Science 72:414-422, 2015.

[3] Richard A Huebner, A Survey of Educational Data-Mining Research, Academic and Business Research Institute,2013

[4] Pornthep Khongchai, Pokpong Songmuang, improving students' motivation to study using salary prediction system, 13th International Joint Conference on Computer Science and Software Engineering, 2016

[5] S. Anupama Kumar Vijayalakshmi M.N. Inference of Naïve Baye‟s Technique on Student Assessment Data, Communications in Computer and Information Science book series (CCIS, volume 270),2011

[6] John Jerrim, Do college students make better predictions of their future income than young adults in the labor force ?, Education Economics, Taylor & Francis Journals, vol. 23(2), pages 162-179, 2015

[7] Karlar Hamlen and William A. Hamlen, Faculty Salary as a predictor of student outgoing salaries from MBA programs, Journal of Education for Business, 2016

[8] Rajveer Singh, A Regression Study of Salary Determinants in Indian Job Markets for Entry Level Engineering Graduates, Masters Dissertation. Dublin Institute of Technology, 2016.

[9] C.C. Hung, E.-P. Lim, On Aggregating Salaries of Occupations from Job Post and Review Data”, IEEE Access, Volume 9, 2021

[10] Dr. Kamaljit I. Lakhtaria, Bhaskar Patel and S.G. Prajapati, Efficient classification of data using decision Tree, Bonfring international journal of data mining 2 (1), 06-12,2012.