An Analytical Approach to Predict Employability Status of Students

Bhavna Saini\textsuperscript{1}, Ginika Mahajan\textsuperscript{1}, Harish Sharma\textsuperscript{1} and Ziniya\textsuperscript{1}

\textsuperscript{1}Manipal University Jaipur, Jaipur-Ajmer Express Highway, Dehmi Kalan, Near GVK Toll Plaza, Jaipur, Rajasthan 303007

E-mail: ginika.mahajan@jaipur.manipal.edu

Abstract. One of the major concerns of the students after graduation is the job opportunities offered to them. Not only students, but also the universities are inclined towards maximizing the job offers for their students through campus recruitment drives. Against this background, the scope of this study is to gauge the performance of top four known classification techniques of data mining, which are, Decision tree, Random forest, Naive Bayes & KNN. These machine learning algorithms are applied on students’ data, collected from the university database of Manipal University Jaipur and student models are created which will predict the employability status of students in future and discover factors which will significantly contribute to their employability. After applying and studying the accuracies of these algorithms, we have found that Random forest behaves better than the rest of the algorithms with 89\% accuracy.

1. Introduction

Employability means that the employee recognize and brace the skills and knowledge of the employees. Graduates who have developed the ability to work in any organization are employable. Information Technology companies or any organization looks forward to some set of technical skills along with non-technical skill set according to the company’s necessity or the market demand. These demands fluctuate but certain skills remain in the constant requirement list of the companies. Since the students are aware with their weaknesses which they can work upon and ameliorate them but unaware of the skills to be refined which employers are demanding for their respective job position \cite{1}\cite{2}\cite{3}. The current study has focused on the classification of students based on attributes considered to be important requirements of companies who come for on-campus recruitment. Classification is the most commonly applied data mining technique. It is a type of supervised learning which directs the data element by employing a set of pre-classified attributes to a class. It is very useful when the output has finite and discrete values \cite{4}. Classifier employs set of attributes and generates a model which uses training data. One attribute is taken as the class \cite{5}. The model which is produced should fit well on the basis of training data and accordingly predict class of test data. Test data is not involved to generate the classifier \cite{6}. This study shows the relation between student’s academic performance and placement in campus recruitment drives through data mining algorithms. Data mining often called knowledge discovery in database (KDD), is known for its powerful role in uncovering hidden information from large volumes of data \cite{7}. The campus recruitment drives by companies presents excellent opportunities to college folks where they are tested for their skills and abilities for employment \cite{8}. It is a widely held view that those whose performance is better in the academics, have better knowledge and work more efficiently in the employment work place \cite{9}.
In the current market scenario where high level of competition exists, one should possess employability skills considered important for any job position. Campus recruitment drives by companies provide job opportunities to the graduates based on their technical and non-technical skill set. Here, we have taken into consideration of technical knowledge. Among hundreds of students, employers choose, selectively, 5 to 10 according to their requirement. This study aims to discover what the qualities of students are that company expects.

We have collected the data of 262 student data records. After collection of data, dataset was merged, cleaned and transformed into seven set of attributes which includes their aptitude skill and technical skill set, Gender, CGPAs, no of time they appeared for the drives etc. Other factors might also affect recruitments but here we have considered only these factors. There are two kinds of prediction done. First is whether the students can be employed or not and second is in what range of salary they can receive which is below 4 lakhs, above 4 lakhs or will be unemployed. Accuracies of all algorithms were evaluated and ranked. Among all, Random Forest performs best with 89% accuracy. It has been proved that Random Forest [10] generates the best efficiency model for prediction as compared to performing decision tree and C4.5. Evaluations have been done by confusion matrix. Confusion matrix is a table which divides and tells how many instances are correctly predicted and how many are not.

2. Related Work
Ajay Kumar Pal and Saurabh Pal [4] conducted a study for analysis for training and placement on dataset of size 65 collected from VBS University, Jaunpur (Uttar Pradesh). They attempted to present a model for enhanced evaluation method. Data mining techniques such as J48 and Naïve Bayes were implemented and 3-fold validation was done using WEKA tools. Z. J. Kovacic [3] explored in the study about the limit to which the factors i.e., enrolment data of students of open polytechnic of New Zealand help in predicting their success. Two algorithms implemented were CHAID and CART which gave accuracies of 59.4 and 60.5 respectively.

Minaei-Bidgolim, [11] classified students by using generic algorithms to predict their final grade based in features taken out from logged data in an education web-based system. Using the regression methods, Kotsiantis and Pintelas [12] predicted a student’s marks (pass and fail classes). R.R. Kabra and R.S. Bichkar [13] conducted a study to predict student’s academic performance. Data was collected via enrolment form filled by 346 engineering students of the college S.G.R Education and Foundation’s College of Engineering and Management. Decision tree J48 was applied to classify whether they would fail or pass [14].

Al-Radaideh [15], predicted the final grades of the student of Yarmouk University, studying C++ course, by implementing a decision tree model. Among other classification models applied, such as ID3, Naïve Bayes, to speculate the student performance of the students of Industrial Engineering Universities Islam Indonesia. Before prediction, feature selection was performed to discover those attributes which have great impact on the student performance. Two important attributes were Student attendance and Grade point Average. According to result showed Naïve Bayes outperformed decision tree in this case.

Pedro Strecht [16], conducted a study to predict students’ result pass or fail and grades in their work. Classification algorithms such as Decision tree and SVM predicted students’ result while Regression models such as SVM, Random Forest, and ADA- Boost.R2 were used to for regression analysis. Data was collected from University of Porto, Result showed that classification models discovered some useful patterns while regression models were of no use in this case.

3. Methodology
3.1. Data collection and pre-processing
Data collection refers to the initial phase where raw data is collected. We have collected data of 262 Computer Science and Information Technology students (2014-2018) of Manipal University Jaipur (MUJ) who attended campus recruitment process shown in table 1. Data was provided by placement cell and university database of MUJ. Source attributes. Range of values for few variables used in the study defined for the present study as follows:
A. **Aptitude:** Companies test undergraduate fresher’s aptitude knowledge and is mostly first test conducted in the placement drives process. Students have been given average score between 0-1 on the basis of their aptitude skills.

B. **Technical:** Technical knowledge of the student which includes the marks obtained by him/her in core subjects and lab works in the respective subjects such as JAVA, JAVA Lab, RDBMS etc.

C. **CGPA_12 & CGPA_btech:** Aggregate CGPA scored by the students in class 12 and in B.Tech up to 6th semester shows the academic success. Most companies set a threshold CGPA which must be achieved to attend their on-campus placement drives.

D. **No_skills:** Technical skills that student have such as certification and training done on CLOUD computing, android development etc., which were not included in the course structure of engineering. Some companies are based upon a particular technology and they prefer students who possess such knowledge.

Table 1. Attributes for Employability Prediction.

| S. No. | Attribute          | Value                                      |
|--------|--------------------|--------------------------------------------|
| 1.     | Aptitude           | 0-1                                        |
| 2.     | Technical          | 0-1                                        |
| 3.     | CGPA_12            | 1-10                                       |
| 4.     | CGPA_btech         | 1-10                                       |
| 5.     | No_skills          | 1-8                                        |
| 6.     | Gender             | 0(female) / 1(male)                        |
| 7.     | No_chances         | 1-26                                       |
| 8.     | Placed             | 0(no) / 1(yes)                             |
| 9.     | Category           | 0(unplaced) / 1(0- 4L) / 2(> 4L)           |

E. **Gender:** Whether the student getting placed is a female or male. ‘0’ denotes female while ‘1’ denotes male.

F. **No_chances:** Total number of times a particular student has appeared for the campus drives. Maximum up to 26 companies visited MUJ campus for placement drives. Student gets familiar with the pattern and types of questions.

G. **Placed:** If the student has job offer or not after completing his/her engineering. Possible values are 0 if unplaced and 1 if placed.

H. **Category:** Undergraduate with given set of skills and knowledge can be placed in companies which provide package greater than 4LPA or below 4LPA. Possible values are 0 if unplaced, 1 if placed with package less than 4LPA and 2 if placed with package greater than 4LPA.

Values were then transformed into units or range of values which were suitable for the statistical analysis. For example, technical skills consists of marks scored by students in 5 subjects which are asked generally in campus recruitment drives by companies and lab scores of those subjects as well. Subjects
include JAVA Language programming, Analysis Lab, Data Structures Lab. Soft skills data are not included in this study. This study purely judges student upon his/her technical and aptitude skills.

4. Implementation and Results
Our goal was to predict and classify students based on their skills, first into two categories, whether they can be employed or not and second, if their employment lies in the salary range of below 4L, above 4L or unemployed. Four classification data mining techniques are applied and their accuracies are ranked accordingly. Python was used to predict and test the results. Results obtained, separately, were compared using confusion matrix.

4.1. Decision Tree.
A decision tree is tree-like model for representing a sequential classification process which includes chance event outcomes. It contains conditional control statements. Decision tree predicts the placement status of student with the accuracy of 87.78%. Also, it is used to predict the salary range in which student can get placed with accuracy of 82.44%. Figure 1 shows the partial view of the one of the two decision trees.

![Decision Tree for Placement Prediction](image_url)

Table 2 displays the confusion matrix modelled for decision tree algorithm, predicting the placement status of students as “Yes” (1) or “No” (0). According to the table, there are 37 students who are actually unplaced and are predicted by decision tree as to un-placed too. While 7 students are actually unplaced but the algorithm predicts them to be placed.
Table 2: Confusion Matrix

| (Actual No) 0 | (Predicted Yes) 1 |
|---------------|------------------|
| (Predicted No) 0 | 37               | 7               |
| (Actual Yes) 1     | 9                | 78              |

Figure 2 shows the decision tree for salary range prediction.

4.2. Random Forest
It can be used for both classification and regression. It builds multiple trees from different samples and different initial variables, then merges them to get more explicit and firm result. For first kind of prediction the accuracy was 89.31% while for the second it achieved 83.20%. It shows better performance than CART. Figure 3 is a graph which illustrates the importance of attributes in the prediction. According to it, Technical and Aptitude are the two top most important features affecting the placement of students in on-campus drives.

4.3. Naïve Bayes Classifier
A Naïve Bayesian classifier is a simple probabilistic classifier. It is based on Bayesian theorem (from Bayesian statistics) with an assumption of independence among predictors which simply means that it assumes that, if a particular feature is present in a class then it is unrelated to the presence of other feature [4]. Bayesian Network employs conditional dependencies using direct graph [17]. For each class, membership probabilities is predicted. The class with maximum scored probability is said to be the most likely. It has achieved 85.49% and 80.15% accuracy in the two predictions.
4.4. **KNN**

K-Nearest neighbours is a classification supervised algorithm, which is extensively used in real life scenarios. KNN sees the majority label of labelled k-nearest neighbours data and learn a function which it uses to produce appropriate output when given new unlabelled data. KNN is based upon the assumption that similar things exist in close proximity. Hence, its performance is proportional to the distance metric employed to figure out the nearest neighbours [18].

Table 3 compares the four algorithms on the basis of their accuracies. Here random forest works best with 89% accuracy.

| Sr. No | Algorithm                           | Accuracy (%) |
|-------|-------------------------------------|--------------|
| 1     | Decision Tree                       | 87.78        |
| 2     | Decision Tree for salary range      | 82.44        |
| 3     | Random Forest                       | 89.31        |
| 4     | Random Forest for salary range      | 83.20        |
| 5     | Naïve Bayes                         | 85.49        |
| 6     | Naïve Bayes for salary range        | 80.15        |
| 7     | K-NN                                | 81.01        |
| 8     | K-NN for salary range               | 77.21        |
5. Conclusion
This study has investigated the performance of machine learning techniques for the early prediction of student’s employability along with their tentative packages. The comparison of various classification algorithms is done which reveals that random forest gives best results. In future, using appropriate software extension methods a proper management system can be built for non-technical person. Overall education data mining opens new doors for educators as well as students to enhance the education system

6. References

[1] Bharambe, Y., Mored, N., Mulchandani, M., Shankarmani, R., & Shinde, S. G. (2017, September). Assessing employability of students using data mining techniques. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2110-2114). IEEE.

[2] Khasanah, A. U. (2017, June). A comparative study to predict student’s performance using educational data mining techniques. In IOP Conference Series: Materials Science and Engineering (Vol. 215, No. 1, p. 012036). IOP Publishing.

[3] Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data.

[4] Pal, A. K., & Pal, S. (2013). Classification model of prediction for placement of students. International Journal of Modern Education and Computer Science, 5(11), 49.

[5] Asif, R., Merceren, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students’ performance using educational data mining. Computers & Education, 113, 177-194.

[6] Arora, Amandeep Singh, Linesh Raja, and Barkha Bahl. "Data centric security approach: A way to achieve security & privacy in cloud computing." Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), 2018.

[7] Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with Java implementations.

[8] Fell, A., & Kuit, J. A. (2003). Placement learning and the code of practice: rhetoric or reality. Active learning in higher education, 4(3), 214-225.

[9] Ferris, K. R. (1982). Educational predictors of professional pay and performance. Accounting, Organizations and Society, 7(3), 225-230.

[10] Khongchai, P., & Songmuang, P. (2016, November). Implement of salary prediction system to improve student motivation using data mining technique. In 2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS) (pp. 1-6). IEEE.

[11] Minaei-Bidgoli, B., & Punch, W. F. (2003, July). Using genetic algorithms for data mining optimization in an educational web-based system. In Genetic and evolutionary computation conference (pp. 2252-2263). Springer, Berlin, Heidelberg.

[12] Kostopoulos, G., Kotsiantis, S., & Pintelas, P. (2015, October). Estimating student dropout in distance higher education using semi-supervised techniques. In Proceedings of the 19th Panhellenic Conference on Informatics (pp. 38-43).

[13] Kabra, R. R., & Bichkar, R. S. (2011). Performance prediction of engineering students using decision trees. International Journal of computer applications, 36(11), 8-12.

[14] Quinlan, J. R. (1987, August). Generating production rules from decision trees. In IJCAI (Vol. 87, pp. 304-307).

[15] Al-Radaideh, Q. A., Al-Shawkaf, E. M., & Al-Najjar, M. I. (2006, December). Mining student data using decision trees. In International Arab Conference on Information Technology (ACIT’2006), Yarmouk University, Jordan.

[16] Strecht, P., Cruz, L., Soares, C., & Mendes-Moreira, J. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. International Educational Data Mining Society.

[17] Almarabeh, H. (2017). Analysis of students' performance by using different data mining classifiers. International Journal of Modern Education and Computer Science, 9(8), 9.
[18] Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. *In Advances in neural information processing systems* (pp. 1473-1480).