STUDENT PLACEMENT ANALYZER

Using Machine Learning and Data Analytics

Nikhil Sonavane
Electronics and Telecommunication Department
Vishwakarma Institute of Technology
Pune, India
nikhil.sonavane16@vit.edu

Harshit Sarna
Electronics and Telecommunication Department
Vishwakarma Institute of Technology
Pune, India
harshit.sarna16@vit.edu

Shoumik Nandi
Electronics and Telecommunication Department
Vishwakarma Institute of Technology
Pune, India
shoumik.nandi16@vit.edu

Shravan Sriram
Electronics and Telecommunication Department
Vishwakarma Institute of Technology
Pune, India
shravan.sriram16@vit.edu

Rishikesh Devarmani
Electronics and Telecommunication Department
Vishwakarma Institute of Technology
Pune, India
rishikesh.devarmani17@vit.edu

Prof. Medha Wyawahare
Electronics and Telecommunication Department
Vishwakarma Institute of Technology
Pune, India
medha.wyawahare@vit.edu

***Abstract*—** One of the biggest challenges that higher learning institutions face today is to improve the placement performance of students. The placement analyzer and prediction is more complex when the complexity of educational entities increase. Educational institutes look for more efficient technology that assist better management and support decision making procedures or assist them to set new strategies. One of the effective ways to address the challenges for improving the quality is to provide new knowledge related to the educational processes and entities to the managerial system. With the machine learning techniques, the knowledge can be extracted from operational and historical data that resides within the educational organization's databases using. The dataset for system implementation contains information about past data of students who may have passed out from that educational institute. These data are used for training the model for rule identification and for testing the model for classification. This paper presents a recommendation system that predicts the students to have one of the six placement classes namely Class A, Class B, Class C, Class D, Class E and Class F. This model helps the placement cell within an organization to identify the prospective students and pay attention to and improve their technical as well as interpersonal skills. It will let the institutes know on which parameters are the companies selecting students. Furthermore, the students in pre-final and final years of their B. Tech course can also use this system to know their individual placement status that they are most likely to achieve. With this they can put in more hard work for getting placed in to the companies that belong to higher hierarchies.

I. INTRODUCTION

The primary aim of students who join professional courses in higher learning institutions is to secure a wellpaid job in a reputed organization. Professional education can be either completely technical or it can be managerial as well. Bachelors of Technology (B. Tech) provides technical education to students in various fields such as Computer Science and Engineering, Electronics and Communication Engineering, Civil Engineering Mechanical Engineering, etc. This degree is aimed at making students experts in state of the art conjectural as well as practical knowledge in various engineering branches. The prediction of placement status that B. Tech students are most likely to achieve will help students to put in more hard work to make appropriate progress in stepping into a career in

various technical fields. It will also help the teachers as well as placement cell in an institution to provide proper care towards the improvement of students in the duration of course. A high placement rate is a key entity in building the reputation of an educational institution. Hence such a system has a significant place in the educational system of any higher learning institution. We have used random forest classifier, decision tree classifier, Support Vector Machine, Kernel Support Vector Machine, K-Nearest Neighbours and Naïve Bayes within Scikit-learn-a machine learning module in spyder having simple and efficient data mining and data analytics capability-for the implementation of the system.

II. Materials and Methods

A. Machine Learning

Machine Learning deals with the development, analysis and study of algorithms that can automatically detect patterns from data and use it to predict future data or perform decision making [1]. Machine learning does its functionality by creating models out of it [2]. Machine Learning has become widespread and has its applications in the field of bioinformatics, computer vision, robot locomotion, computational finance, search engine etc.

B. Decision Tree

In real world problems, observations are made on entities associated with a problem so as to make inferences on the target value of those entities. This mapping is encompassed in a predictive model with the help of decision trees. This method of learning is referred to as Decision Tree Learning. This is one of the predictive modeling methods that can be found in the fields of data mining [3], machine learning and statistics. In this model we have made use of classification trees, a typical decision tree in which the predictor variable (target variable) takes on finite set of categorical values only. In this type of trees, the leaves represent the class labels and the branches represent the splitting path through which a decision travels from root to the leaf of the tree.

C. Naïve Bayes

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. [4]

D. Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. [4]

E. K-Nearest Neighbour

The k-nearest-neighbors algorithm is a classification algorithm, and it is supervised: it takes a bunch of labelled points and uses them to learn how to label other points. To label a new point, it looks at the labelled points closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point (the "k" is the number of neighbors it checks). [4]

F. Support Vector Machine

Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. Effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient. The algorithm does not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

G. SCI-Kit Learn

Scikit-learn is an open source machine learning module in python [4] that is comprised of wide range of classification, clustering and regression algorithms in machine learning. Major algorithms featured are Naive Bayes, Decision Tree, Random Forests, Support Vector Machines, Logistic Regression, Gradient Boosting, kMeans and DBSCAN. This module is primarily aimed at solving supervised and unsupervised problems. It aims at making machine learning accessible to novices by providing an abstraction using a general-purpose high-level language. Ease of use, documentation, performance and API consistency are the key features of this module. [5]

H. Background and Related Work

Machine Learning techniques has a significant role in deriving innovative knowledge in the educational field so as to help students for their better performance in placement. Many scientists across the world has done considerable amount of work in determining the methodologies for performance analysis and placement. Few of the relevant works in this field are listed out so as to obtain an idea on what has been done so far and what further growth is expected in this area of work. Hijazi and Naqvi [11] conducted a study to find the factors affecting the academic performance of students. They made use of questionnaires to elicit information from students highlighting factors such as income factor, parents' educational background, size of the family, regularity of teachers, subject interest created by the teachers and student's interest in co-curricular activities. They used Pearson Correlation Coefficient

to highlight the important factors and they found that mother's education and family income played an important role in students' academic performance. Pal and Pal [6] conducted a study on student data that have information on their academic records and proposed a classification model to find an efficient method to predict student placements. They concluded that Naïve Bayes classifier is the best classification method for use in placements in comparison with Multilayer Perceptron and J48 algorithms. Ramanathan, Swarnalatha and Gopal [7] conducted a study using sum of difference method for students' placement prediction. They used the attributes such as age, academic records, achievements etc. for the prediction. They concluded that based on their results higher learning institutions can offer its students a superior education. Arora and Badal [8] conducted a study to predict student placements using data mining. They made predictions on MCA students in Ghaziabad in UP, considering parameters such as MCA result, Communication skills, programming skills, co-curricular activity participation, gender, 12th result and graduation result. They concluded that their model based on decision tree algorithm can assist the placement cell and faculties in identifying set of students that are likely to face problem during final placements. Elayidom, Idikkula and Alexander [9] designed a generalized data mining framework for placement chance prediction problems. They considered the students' Entrance Rank, Gender, Sector and Reservation Category to predict the branch of study that is Excellent, Good, Average or Poor for him/her using decision trees and neural networks. Naik and Purohit [10] made a study to use prediction technique using data mining for producing knowledge about students of MCA course before admitting them.

I. Data Preparation

The dataset used for training as well as testing was obtained by creating a google form and getting the responses by asking them to fill it via phone calls. The data sample is of 400 plus entries of students who passed out of the institution in the academic year 2016-2017.

J. Procedure

A google form was sent to the students of the institution of various departments. After getting responses the data was split into various columns some being CPI, domain of project, domain of internship. Some columns were again then spitted into different domains like in column of domain of project. We have a column for machine learning/deep learning/artificial intelligence, another for embedded systems and image processing, one for VLSI, signal processing, the other being for Internet of things and android, also columns for mechanical and research. Say if a student has his domain of project as machine learning in third year and signal processing in fourth year, then we will assign '1' to columns of machine learning/deep learning/artificial intelligence, '1' to column of signal processing and VLSI and '-1' to rest all of the columns.

After doing this to all columns wherever necessary, we move forward to applying different classifiers and predicting which classifier is the best for our data. P value's were used to determine which column played the most crucial rule and had the most weightage in determining the class of the company.

Description	Possible Values
GR Number	Integer
Department	Comp, E&TC, Electronics, Mechanical, IT, Instrumentation
Domain of Project	ML, DL, AI, Signal Processing, Image Processing, Research, IoT, Android, Mechanical, Data Mining, Data Analytics, Communication, Theory of Machines, Gears
Domain of Internship	ML, DL, AI, Signal Processing, Image Processing, Research, IoT, Android, Mechanical, Data Mining, Data Analytics, Communication
CPI	Integer

Table 1- Parameters

K. Implementation of Machine Learning Model

Python is one of the best data analytics language used widely in the industry. It has sophisticated data mining and machine learning capabilities and is the best practical language for building products. This makes python a favorite in data processing. In data processing, there's often a trade-off between scale and sophistication, and Python has emerged as a compromise. Anaconda Navigator (Spyder) notebook and NumPy can be used as a scratchpad for lighter work, while Python is a powerful tool for medium-scale data processing. Python also has lot of advantages like rich data community, offering vast amounts of toolkits and features. Sci-kit learn is a greatly a sophisticated module available in python, which comprises of almost all major machine learning algorithms. The training dataset which contains the placement data of 2014 pass out batch is loaded to the python code, macros are attached to the variables for their easy processing and is fit to a decision tree classifier model using Scikit libraries. Once the modelling is completed, the test data is uploaded to the python, the variables are read using macros and is provided to a predict function available in Scikit learn. This produces the macro result corresponding to the placement status class with respect to the training data. Finally, the macro result is mapped back to the placement status variable used in the model.

III. EXPERIMENTS AND RESULTS

Algorithm	Accuracy
Random Forest	**81.03 %**
Decision Tree	67.24%
K-NN	56.8%
SVM	55%
Kernel SVM	67.3 %
Naïve Bayes	46.5%

Table 2: CS + IT + E&TC Results

Algorithm	Accuracy
Random Forest	**75.3%**
Decision Tree	70 %
K-NN	60 %
SVM	53.42 %
Kernel SVM	56.1%
Naïve Bayes	19.1 %

Table 3: CS + IT + E&TC + Elex + Instru results

Algorithm	Accuracy
Random Forest	**79%**
Decision Tree	72%
K-NN	53%
SVM	60%
Kernel SVM	61%
Naïve Bayes	27%

Table 4: CS + IT + E&TC +Elex + Instru + Mech Results

IV. Conclusion

V. References

[1] Kohavi, R. and F. Provost (1998) *Glossary of term, Machine Learning* 30:271-274.

[2] Bishop, C.M. (2006) *Pattern Recognition and Machine Learning. Springer,* ISBN0-387- 31073- 8.

[3] Rokach, L and O. Maimon (2008) "Data mining with decision trees: theory and applications*," World Scientific Pub Co Inc. ISBN 978-98127717711*.

[4] Pedregosa, F , G. Varoquax, A. Gramfort, V. Michel, B. Thrion, O. Grisel, M.Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos and D. Cournaeau (2011) "Scikit-learn: Machine Learning in Python*," Journal of Machine Learning Research 12: 2825-2830.*

[5] Pal, A.K. and S. Pal (2013) "Analysis and Mining of Educational Data for Predicting the Performance of Students," *(IJECCE) International Journal of Electronics Communication and Computer Engineering, Vol. 4, Issue 5*, pp. 1560-1565, ISSN: 2278-4209, 2013.

[6] Ramanathan, L., P. Swarnalathat and G.D. Gopal (2014) "Mining Educational Data for Students' Placement Prediction using Sum of Difference Method," *International Journal of Computer Applications 99(18): 36-39*

[7] Arora, R.K. and Dr. D. Badal (2014 ")Placement Prediction Through Data Mining," *International Journal of Advanced Research in Computer Science and Software Engineering. Volume 4, Issue 7.*

[8] Elayidom, S., S. M. Idikkulaand J.Alexander (2011) "A Generalized Data mining Framework for Placement Chance Prediction Problems," *International Journal of Computer Applications(0975– 8887) Volume 31– No.3*.

[9] Naik, N. and S. Purohit (2012) "Prediction of Final Result and Placement of Students using Classification Algorithm," *International Journal of Computer Applications (0975 – 8887) Volume 56- No.12.*

[10] Hijazi, S.T. and R. S. M. M. Naqvi (2006) "Factors affecting student's performance: A Case of Private Colleges, Bangladesh e-Journal of Sociology*," Vol. 3, No. 1 A Case of Private Colleges, Bangladesh e-Journal of Sociology,* Vol. 3, No. 1