Journal of Engineering and Applied Sciences
MEDWELL PUBLICATIONS

Recommendation System for Predicting the Placement Percentage for an Educational Institute

Varul and Shubham Tiwari
Department of Computer Science, Galgotia University, Greater Noida, Uttar Pradesh, India

Key words: Data mining, machine learning, neural network, data cleaning, data set, data distribution, Training, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Artificial Neural Network (ANN), T-Distributed Stochastic Neighbor Embedding

Corresponding Author:
Varul
Department of Computer Science, Galgotia University, Greater Noida, Uttar Pradesh, India

Page No.: 2817-2826
Volume: 15, Issue 14, 2020
ISSN: 1816-949x
Journal of Engineering and Applied Sciences

Abstract: Engineering students are dubious about what they need to pursue after graduation. With extensive options available, starting from campus recruitments to Masters, students are perplexed, adding factors like salaries and different job opportunities makes it even worse. There aren't any reliable platforms where a student can predict the outcomes from the beginning of engineering and take action to bridge this gap for a far better future. Placement of students is one of the vital activities in academic establishments. Admission primarily depends on placements. Admission is directly proportional to the offer letters received by an institute. Hence all institutions strive to strengthen the placement department. Students studying in engineering colleges feel the difficulty to understand where they substitute comparison to others and what quite a placement they might get. The training and placement offices are available to an image when a student enters the final year but they're of no use to a student planning for future studies. Prediction about the student's performance is an integral part of an education system because the overall growth of the scholar is directly proportional to the success rate of the scholars in their examinations and extra-curricular activities. Therefore, there are many situations where the performance of the scholar must be predicted for instance in identifying weak performing students and taking actions for his or her betterment. The students do not have any platform to see their current position and repose on their strengths. The platforms currently available haven't been trained on real and complete data sets and don't learn from their wrong predictions which reduces the accuracy within the future. To realize far better efficiency and a system that determines with every wrong prediction it's made, so it uses algorithms like Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) which can cause endless accuracy growth. The model is going to train on a real data set and a massive number of

qualitative as well as quantitative parameters are going to consider. This study aims to study the previous year's student's data and predict the placement possibilities of current students and aids in increasing the situation percentage of the institutions. This study presents a recommendation system that predicts whether the present student is going to be placed or not with a percentage value. This study helps the placement cell at intervals to identify potential students and concentrate on and improve their technical and social skills.

INTRODUCTION

Campus placement of a student plays a significant role in a college. Campus placement is a process where companies visit colleges and identify students who are talented and qualified before they complete their graduation. Therefore, taking a career decision regarding the placement after completing a particular course structure is crucial in a student's life always. A high placement rate may be an essential entity in building the reputation of an academic institution. Hence, such a system features an essential place within the educational system of any higher learning institution. An educational organization contains a large number of student records. Therefore, finding patterns and characteristics in this vast pool of data will help find parameters that are the most important for this placement procedure.

The prediction of engineering students, about where they can place from the second year and onwards will help to improve the efforts of students for proper progress. It will help teachers to take appropriate attention towards the growth of the student over time. It will help to build a reputation of the institute for having such a sophisticated system in place which allows the students to train and practice for campus placements. The present study concentrates on helping the students, bridging the gap between the industry and the curriculum and showing them the path to a better future. So, using machine learning and data mining operations to understand the potential of the scholar.

One of the useful ways to approach the challenges for developing the condition is to provide new experience related to the institutional processes and entities to the managerial method. With the machine learning techniques, the information often extracted from operational and historical expertise that resides at intervals the academic organization's database exploitation. Machine learning is a sub-domain of computer science which developed from the knowledge of pattern identification in data and also from the computational learning system in artificial intelligence. It is a superior ticket to the most impressive careers in data analytics.

Data mining suggests to deriving or mining valuable patterns from a vast database. It is knowledge discovery in a large amount of data. Machine learning algorithms are a part of soft computing, that works well with low-level computing, gaining experience and knowledge from its mistakes and the later works well with the irregularities and the incompleteness of the data. While data mining tools can be applied to find patterns in the broad set of data that can help understand business requirements, market analysis and management. Machine learning algorithms can be of two types, supervised learning and unsupervised learning. Supervised learning is so named because the info expert acts, since, a guide to show the algorithm what results from it should come up among. It's almost like the way a toddler might learn arithmetic from an educator. Unsupervised machine learning is more aligned with what some call true AI the thought that a computer can learn to spot complex processes and patterns without a human to guide the way. So, it is going to use unsupervised machine learning techniques to guide the students.

In this technological world, many different techniques are used by humans for various purposes. There are many software's and applications that develop for reducing human effort. Data mining techniques and machine learning can use to find patterns in large databases and to guide decisions about future activities. It's supposed that by using machine learning, the model will learn on its own and work efficiently even with minimal input from the user to recognize. The model can be useful to understand the unexpected and provide an analysis of data followed by decision-making which measure and it ultimately leads to strategic decisions and business intelligence. The most straightforward word for knowledge extraction and exploration of volume data is very high and the more appropriate term is "Exploring the hidden knowledge of the database". This process includes the preparation and interpretation of results.

Literature review

ID3 classification algorithms: It identified relevant attributes based on quantitative and qualitative aspects of a student's profile such as CGPA, academic performance, technical and communication skills. It designed a model that can predict the placement of a student using ID3[1].

Classification algorithms: It implemented an empirical analysis on predicting academic performance by using classification techniques or mapping of data items into predefined groups and classes using supervised learning. They compared five classification algorithms, namely Decision Tree, Naive Bayes, Naive Bayes Tree, k-Nearest Neighbour and Bayesian Network algorithms

for predicting student's grades, particularly for engineering students using a four-class prediction problem[2].

State of the art regression algorithms predicted the student marks (pass and fail classes) using the regression methods and available previous data. The scope of this work compares some of the states of the art regression algorithms in the application domain of predicting student's marks. Several experiments have executed with six algorithms which trained using datasets provided by the Hellenic Open University. Student Placement Guidance Review of Literature[3].

Logistic regression: It applied the logistic regression method on the examination result data. They analyzed the data under the University Grant Commission sponsored project entitled-Prospects and problems of educational development (Higher Secondary Stage) in Tripura-An in-depth study.

Decision tree algorithm C4.5: It used decision tree algorithm C4.5 to establish a classification rule and an analysis-forecasting model for student's marks. They described how the analysis- forecasting result can be used to find out the factors which can affect student's scores, so, some negative learning habits or behaviors of students can be revealed and corrected in time. The teaching effect of the teacher can monitor, the teaching management can also be supportive.

ID3 and C4.5 classification algorithms: It analyzed the data set containing information about students such as gender, marks scored in the board examinations of classes X and XII, marks and rank in entrance examinations and results in the first year of the previous batch of students. By applying the ID3 and C4.5 classification algorithms on this data, they have predicted the general and individual performance of freshly admitted students in future examinations-student Placement Guidance Report on the Present Investigation (Existing Systems)[4].

Existing systems: A lot of research has already done on the topic of placement prediction in the past decade. The different researchers used different methods to produce the intended results. It was used as a classification algorithm to predict the outcome and placement of the students[1]. They used data mining techniques for creating knowledge about students of the Master of Computer Application (MCA) course before admitting them to the class. The overall error occurred to classify validation data using the MCA result prediction classification tree was 38.46% while for validating placement prediction classification tree, it was 45.38%. It was used as a logistic regression model to create a Placement Predictor System (PPS). They generated results from an open-source GNU Octave programming tool which brings about 83.33% accuracy 22[5]. Another approach for placement prediction used ID3 Decision Tree Algorithm. While predicting the placement, they incorporated both qualitative and quantitative parameters of a student to achieve better results. It used the machine learning model of the k-Nearest Classifier to predict the probability of an undergrad student getting placed in an IT company. They compared the results of the same against the results obtained from other models like logistic regression and SVM and proved that KNN produces better results.

Based on the above research, it proposed the usage of Artificial Neural Network for placement guidance which will provide higher accuracy compared to other algorithms. Though attempts made to create such system taking into consideration both qualitative and quantitative parameters; the number of qualitative factors considered for the same was very less which is intended to change by using >50 qualitative parameters which constitute a vital role in the placement of a student consequently improving the accuracy of the system.

Problem statement: Students studying in engineering colleges feel the difficulty to know where they stand in comparison to others and what kind of placement they would get. The training and placement offices come in the picture when a student enters the final year but they are of no use to a student planning for future studies. The students have no platform to check their current position and build on their strengths.

The platforms currently available have not been trained on real and complete data sets and do not learn from their wrong predictions which reduces the accuracy in the long term.

Planning for future role constitutes a vital role in any engineering student's life. The model compels a system to assist the academic planners in designing a strategy to improve the performance of students that will help them in getting placed at the earliest.

In all of the previous systems, placement prediction of a student done in terms of binary values, i.e., 0 and 1 which does not represent a clear picture to the user. Creating a system that will guide a user in a better way, like showing probability ranging from 0 to 1 is essential in such cases.

During placement prediction, various attributes of a student play a vital role in whether that particular student will get selected or not. These attributes constitute both qualitative and quantitative parameters. Previous systems considered only qualitative parameters of a student overlooking personal aspects of a candidate such as his confidence, ability to work on a problem, etc. Taking this into consideration, a system incorporating both parameters will provide better guidance to the students. Some of the earlier systems used decision trees such as

ID3 to provide placement prediction. But such methods are computationally too heavy and bound to break when a large number of datasets present. A self-adaptive Artificial Neural Network will overcome this significant drawback of previous systems.

Proposed system for project: Students are most benefited by this application. The students can manage their profile and give tests about programming languages, logic building and other such topics. The college has the student's quantitative data like CGPA, marks, internships, projects and certifications. The test data which gives the qualitative parameters and the quantitative parameters aid the predictive model that uses the maximization of entropy. Once the prediction graph is generated, all it needs is to fit a curve to map the data and apply the entropy maximization algorithm, so that, prediction can be done accurately. The students get the statistical data that will help with analytics and knowing how to improve themselves to get a better package. Statistical analytics also help the TPO to verify the data and if incorrect, TPO can change the data to maintain the accuracy.

Neural network: A standard neural network has nothing, but from few dozen to hundreds, thousands or even numerous artificial neurons described units provided during a series of layers, each of which equates to the sheets on either side. A variety of them stated as input units are designed to receive various kinds of information from the skin world that the network will attempt to study, recognize or otherwise process. Other groups sit on the opposite side of the system and signal how it responds to the information it's learned those start as output units. In between the input units and output, units are one or more layers of hidden groups which together form the majority of the bogus brain. Most neural networks are fully connected which suggests each hidden unit and each output unit connect to every unit within the layers either side. The connections between one group and another are represented by a variety called weight which can be either positive (if one unit excites another) or negative (if one unit suppresses or inhibits another). The upper the burden, the more influence one unit has on another. (This resembles the way genuine brain cells trigger one another across tiny gaps termed synapses). Data flows through a neural network in two ways. While it's learning (being trained) or acting normally (after being trained), models of data fed into the net via. the input layers which trigger the layers of hidden units and these successively hit the production units. This standard-design called a feedforward network. Not all groups "fire" all the time. Each system receives inputs from members to the left and hence, the data do multiply by the weights from connections they travel. Every unit adds up all the contributions it receives during this manner and (in the most effective reasonably network) if the sum is over a selected threshold value, the unit "fires" and triggers the groups it's connected to (those on its right). For a neural network to seek out, there has to be a part of feedback involved-just as children learn by being told what they're doing right or wrong. We all use feedback all the time. Remember to once you initially learned to play a game like ten-pin bowling.

As you picked up the heavy ball and rolled it down the alley, your brain watched how quickly the ball moved and thus the road it followed and noted how close you came to demolition the skittles. Next time it was your turn, you remembered what you'd done wrong before, modified your movements accordingly and hopefully threw the ball barely better. So, you practiced feedback to match the top result you wanted with what happened, revealed the difference between the two and used that to change what you most likely made subsequent time ("I should throw it harder," "I should roll lightly more to the left," "I should forsaking later," then on). The more extensive the difference between the intended and actual outcome, the more radically you'd have altered your moves. Neural networks learn things in precisely the same way, typically by a feedback process called backpropagation (sometimes abbreviated as "backdrop"). This involves comparing the output of a network produces with the production is meant to create and using the difference between them to modify the weights of the connections between the units in the system, working from the output units through the hidden groups to the input units-going backwards in other words. In time, backpropagation lets the network to detect, reducing the disparity between actual and expected output to the point where the two exactly coincide, so, the system figures things out precisely as it should.

MATERIALS AND METHODS

Practical terms may be calculations, technical details, data manipulation and processing and other specific functionality that define what a system is supposed to accomplish. Our web Application has three modules, which are design for three different users. First is students Portal, where they can log in through their portal page or registered themselves. Students can create their profiles using a personal dashboard. An administrator portal can log in to his account and he/she will send emails regarding placement and companies and verify the details and apply filters on the data. View the placement prediction analysis reports generated by our model. The company Portal use the generated report where the interviewer can view student profiles while taking the interview and will be prompted questions based on the student resume. The main backend of our system is our logistic Model. This mathematical model/machine

continually learns with every student's test data from the database and it will process this information. It will give the final numeric value/probability of success or getting placed.

The model uses the Agile methodology. This method promotes continuous iteration of development and testing throughout the software development lifecycle of the project. During the life cycle of the product, iterations were built simultaneously, providing efficient and quality output. Implementation plan comprised of the following significant steps.

Gathering data: This step is crucial because the quality and quantity of data that you pick will directly determine how good your predictive model can be. In this case, data that is collected consisted of student's marks across all semesters.

Data preparation: Data preparation where our data is loaded into a suitable place and prepare it for use in our machine learning training. This is also an excellent time to do any pertinent visualizations of your data to help you see if there are any relevant relationships between different variables you can take advantage of as well as show you if there are any data imbalances. This step comprised of converting data from various formats to Excel and perform different data visualization techniques to get insights about the features.

Then it also needed to split the data into two parts. The first part used in training our model will be the majority of the dataset. The second part will be used for evaluating our trained model's performance. It is mandatory to not use the same data that the model was trained on for evaluation.

Choosing a model: The next step in our workflow is selecting a model. There is 60+ predictive modelling algorithms to choose from. So, it is mustunderstand the type of problem and solution requirement to narrow down to select a few models which can be evaluated. The algorithms that are considered are as follows:

- Logistic regression
- Support Vector Machine (SVM)
- K-Nearest Neighbour (KNN)
- Decision tree
- Artificial Neural Network (ANN)

After getting a confidence score from each model, so, it was must to ranked our evaluation of all the models to choose the best one for our problem. So, it was evaluated to go forward with ANN because it can handle much more variability as compared to traditional models (Fig. 1).

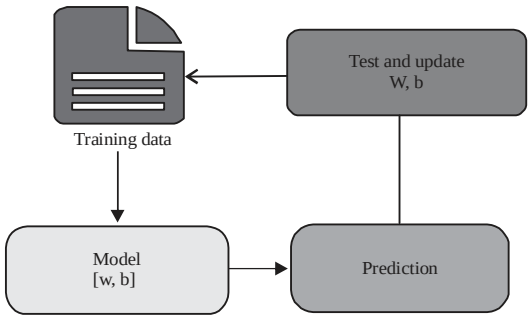


Fig. 1: Traditional modles

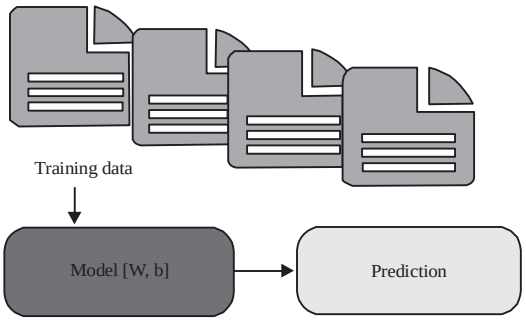


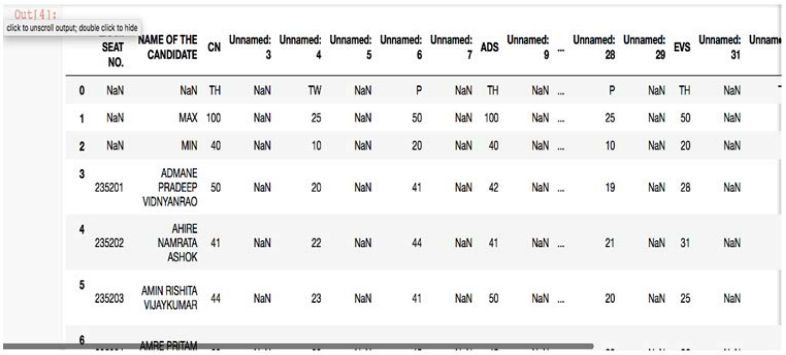
Fig. 2: Parameters tunning

Training: In this step, it will use our record to incrementally improve our model's ability to predict the probability of a student being placed. The training model consists of Weights (W) and biases (b) where weights are nothing but a collection of features.

The training process involves initializing some random values for W and b and attempting to predict the output with those values. So, it cancompare our model's predictions with the production that it should produce and adjust the values in W and b such that it will have more correct predictions. This process then repeats. Each iteration or cycle of updating the weights and biases is called one training "step." student placement guidance methodology evaluation.

Evaluation: Once training is complete, it's time to see if the model is any good using assessment. This is where that dataset that was set aside earlier comes into play. The evaluation allows us to test our model against data that has never used for training. This metric allows us to see how the model might perform against data that it has not yet seen.

Parameter tuning: Further, improvement of the model is made using parameter tuning. There were a few parameters which implicitly assumed when training was completed and in this step, it is to go back and test those assumptions and try other values (Fig. 2).



Out[4]:
click to unscroll output; double click to hide

	SEAT NO.	NAME OF THE CANDIDATE	CN	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	ADS	Unnamed: 9	...	Unnamed: 28	Unnamed: 29	EVS	Unnamed: 31	Unnam
0	NaN	NaN	TH	NaN	TW	NaN	P	NaN	TH	NaN	...	P	NaN	TH	NaN	
1	NaN	MAX	100	NaN	25	NaN	50	NaN	100	NaN	...	25	NaN	50	NaN	
2	NaN	MIN	40	NaN	10	NaN	20	NaN	40	NaN	...	10	NaN	20	NaN	
3	235201	ADMANE PRADEEP VIDNYANRAO	50	NaN	20	NaN	41	NaN	42	NaN	...	19	NaN	28	NaN	
4	235202	AHIRE NAMRATA ASHOK	41	NaN	22	NaN	44	NaN	41	NaN	...	21	NaN	31	NaN	
5	235203	AMIN RISHITA VIJAYKUMAR	44	NaN	23	NaN	41	NaN	50	NaN	...	20	NaN	25	NaN	
6		AMRE PRITAM														

Fig. 3: Pandas data frame used for cleaning

	CN	CN_TW	CN_Practical	ADS	ADS_TW	ADS_Practical	MP	MP_TW	MP_Practical	TCS	TCS_TW	WE
count	166.000000	169.000000	169.000000	166.000000	169.000000	169.000000	163.000000	169.000000	169.000000	161.000000	169.000000	163.000000
mean	53.722892	21.467456	42.112426	47.795181	20.520710	39.100592	41.220859	21.041420	21.384615	41.888199	21.213018	48.613497
std	13.579229	1.721833	3.050043	8.761844	1.942901	2.953297	12.340351	1.943717	1.349603	16.340897	2.418101	12.625824
min	4.000000	12.000000	35.000000	8.000000	16.000000	18.000000	3.000000	12.000000	16.000000	3.000000	16.000000	13.000000
25%	45.000000	20.000000	40.000000	41.000000	19.000000	38.000000	33.000000	20.000000	20.000000	32.000000	19.000000	40.000000
50%	54.000000	21.000000	42.000000	47.500000	20.000000	39.000000	42.000000	21.000000	22.000000	41.000000	22.000000	50.000000
75%	62.000000	23.000000	44.000000	54.000000	22.000000	40.000000	48.000000	23.000000	22.000000	51.000000	23.000000	58.500000
max	84.000000	24.000000	48.000000	69.000000	23.000000	47.000000	76.000000	24.000000	23.000000	89.000000	24.000000	72.000000

Fig. 4: Cleaned data frame

One example is how many times we run through the training dataset during training. It can "show" the model our full dataset multiple times rather than just once. This leads to higher accuracies.

Another parameter is the "learning rate." This defines how far it can shift the line during each step, based on the information from the previous training step. These values all play a role in how accurate our model can become and how long the training takes.

Prediction: Machine learning is using data to answer questions. So, prediction or inference is the step where it gets to answer some questions. In this step, our model is used to predict the probability of a student getting placed.

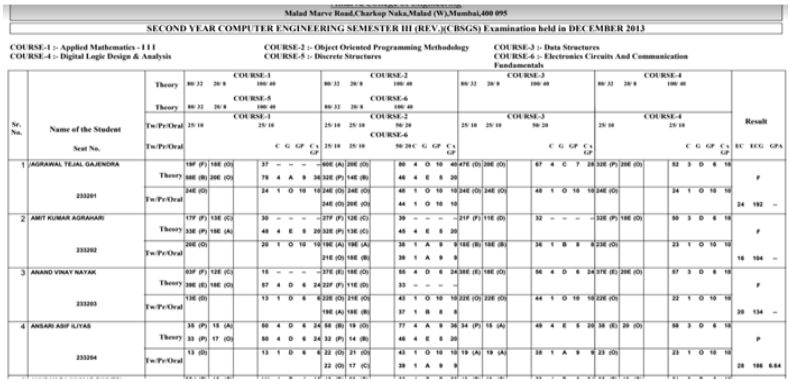
Dataset and connectivity: The models were trained on the bases of previous data of the CMPN branch of Galgotias university. The placement data and all the results were taken between the 2015 Passout batch to the 2018 Pas1ut batch. The data includes all the University gazettes from Sem 3 to Sem 8. The periodicals include all the details like:

- Internal assessment marks
- Grades
- Pointers
- Practical marks
- Teamwork marks and
- Semester theory marks

The data on which it is working was in Pdf format, which was then converted into excel using python scripts and then imported the excel sheets as panda's data frame wherein it was analyzed merged. Performed transformations upon the data to get insights and make it ready for the model training. The data consisted of only quantitative parameters, so, it had to infer qualitative parameters based on those quantitative parameters. By keeping seat number and Name as a primary key to merge student records and got one single track record of the student which was the entire history of the student in the college. So that, it must join the previous placement results in the same to train our model in the right manner, thus creating a complete and reliable centralized database. These processed data were stored in the excel formats which can be easily converted to MS-SQL format. The database tied with the helper functions using cufflinks, so, that whenever the data changes, the graphs reflect the desired changes (Fig. 3 and 4).

RESULTS AND DISCUSSION

Data cleaning results: Data cleaning is the process of detecting and correcting inaccurate records from a



Malad Marve Road,Charkop Naka,Malad (W),Mumbai,400 095

SECOND YEAR COMPUTER ENGINEERING SEMESTER III (REV.)(CBSGS) Examination held in DECEMBER 2013

COURSE-1 :- Applied Mathematics - I I I
COURSE-2 :- Object Oriented Programming Methodology
COURSE-3 :- Data Structures
COURSE-4 :- Digital Logic Design & Analysis
COURSE-5 :- Discrete Structures
COURSE-6 :- Electronics Circuits And Communication Fundamentals

Sr. No.	Name of the Student / Seat No.		COURSE-1	COURSE-2	COURSE-3	COURSE-4	Result
1	AGRAWAL TEJAL GAJENDRA	Theory	19F (F) 18E (O) 37 – – – – 60E (A) 20E (O)	80 4 O 10 40 47E (D) 20E (O)	67 4 C 7 28 32E (P) 20E (O)	82 3 D 6 18	F
	233201		59E (B) 20E (O) 79 4 A 9 36 32E (P) 14E (B)	46 4 E 5 20			
		Tw/Pr/Oral	24E (O) 24 1 O 10 10 24E (O) 24E (O)	48 1 O 10 10 24E (O) 24E (O)	48 1 O 10 10 24E (O) 24E (O)	24 1 O 10 10	24 192 –
				24E (O) 20E (O)	44 1 O 10 10		
2	AMIT KUMAR AGRAHARI	Theory	17F (F) 13E (C) 30 – – – – 27F (F) 12E (C)	39 – – – – 21F (F) 11E (D)	32 – – – – 32E (P) 18E (O)	50 3 D 6 18	F
	233202		33E (P) 18E (A) 48 4 E 5 20 32E (P) 13E (C)	45 4 E 5 20			
		Tw/Pr/Oral	20E (O) 20 1 O 10 10 19E (A) 19E (A)	36 1 A 9 9 14E (B) 18E (B)	36 1 B 8 8 23E (O)	23 1 O 10 10	16 104 –
				21E (O) 18E (B)	39 1 A 9 9		
3	ANAND VINAY NAYAK	Theory	03F (F) 12E (C) 15 – – – – 37E (E) 18E (O)	55 4 D 6 24 38E (E) 18E (O)	56 4 D 6 24 37E (E) 20E (O)	57 3 D 6 18	F
	233203		39E (E) 18E (O) 57 4 D 6 24 22F (F) 11E (D)	33 – – – –			
		Tw/Pr/Oral	13E (D) 13 1 D 6 6 22E (O) 21E (O)	43 1 O 10 10 22E (O) 22E (O)	44 1 O 10 10 22E (O)	22 1 O 10 10	20 134 –
				19E (A) 18E (B)	37 1 B 8 8		
4	ANSARI ASIF ILIYAS	Theory	35 (P) 15 (A) 50 4 D 6 24 55 (B) 19 (O)	77 4 A 9 36 34 (P) 15 (A)	49 4 E 5 20 36 (E) 20 (O)	56 3 D 6 18	P
	233204		33 (P) 17 (O) 50 4 D 6 24 32 (P) 14 (B)	46 4 E 5 20			
		Tw/Pr/Oral	13 (D) 13 1 D 6 6 22 (O) 21 (O)	43 1 O 10 10 19 (A) 19 (A)	36 1 A 9 9 23 (O)	23 1 O 10 10	28 186 6.64
				22 (O) 17 (C)	39 1 A 9 9		

Fig. 5: Representation of data in intial format

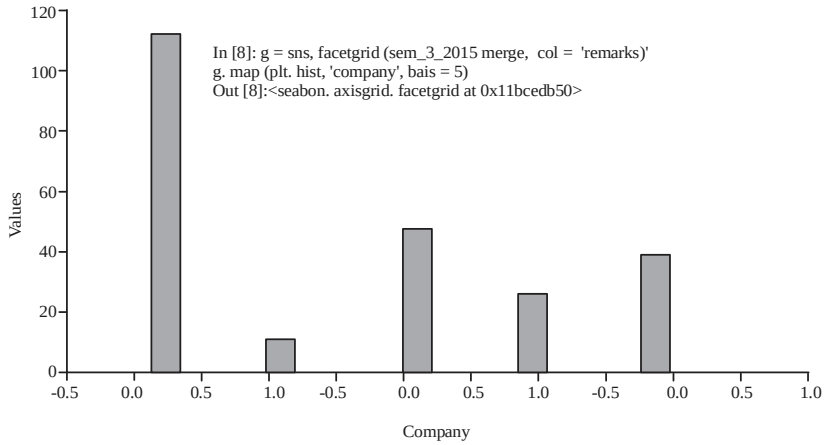


Fig. 6: Comparison of fail/pass students with placement

database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools or as batch processing through scripting. Here, it was used different scripts for data cleaning purposes. After cleansing, a data set was consistent with other similar data sets in the system (Fig. 5).

Exploratory data analysis results: Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. Numerous data visualization techniques are used here to find meaningful insights into the dataset (Fig. 6).

From the graph, it can be concluded that unplaced student contains a high percentage of FAIL students. This validates our assumption that failing a student in a certain exam has a direct impact on his placement.

The above correlation graph provides insights into the importance of each subject from a placement perspective. Insights like these proved to be useful in the selection of features for model prediction (Fig. 7).

Student tracking: Data visualization proves to be useful for tracking of different students. Using cleaned data, it becomes easier to track the progress of each student over each semester. Understanding of the data becomes easier when it is represented in the form of graphs (Fig. 8 and 9).

The placement phase in a student's life is stressful and proves a challenging experience. Not only candidates, placement coordinators, teachers but also alumni describe it as a brainstorming period. As the students stagger through days of interviews and test, their peer group becomes simultaneously a cause for stress and if the student doesn't feel ready for it then his thinking might take him down in competition, so, there is a strict need of a proper platform to guide and boost up the placement preparation. Thus, the market is already flooded with various placement prediction platform but their inability to learn from the wrong predictions is leading to a reduction in the accuracy in the long race. The main drawback in the earlier model is they either used fuzzy logic or ML algorithm like classification algorithm, data mining techniques, logistic regression model, open-source

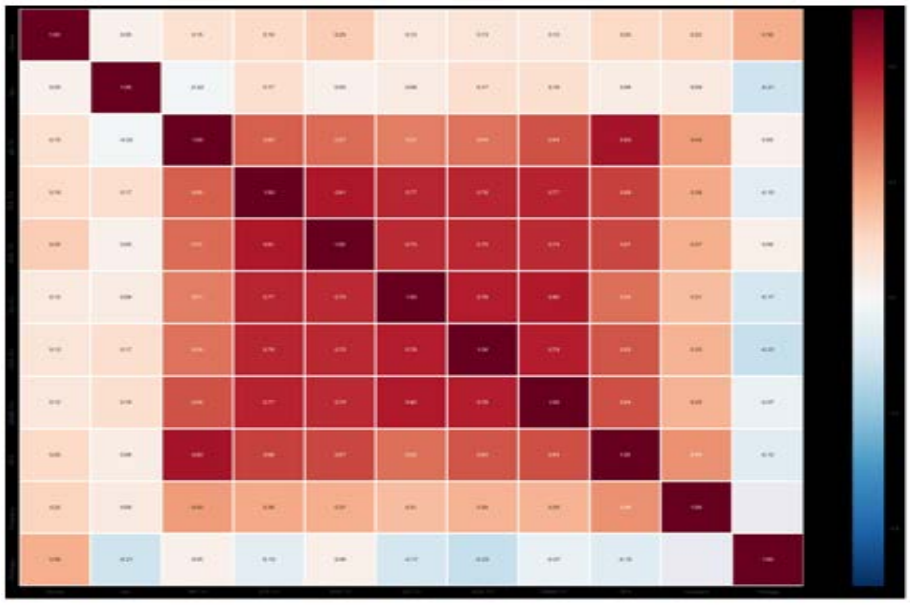


Fig. 7: Correlation of each subject with one another

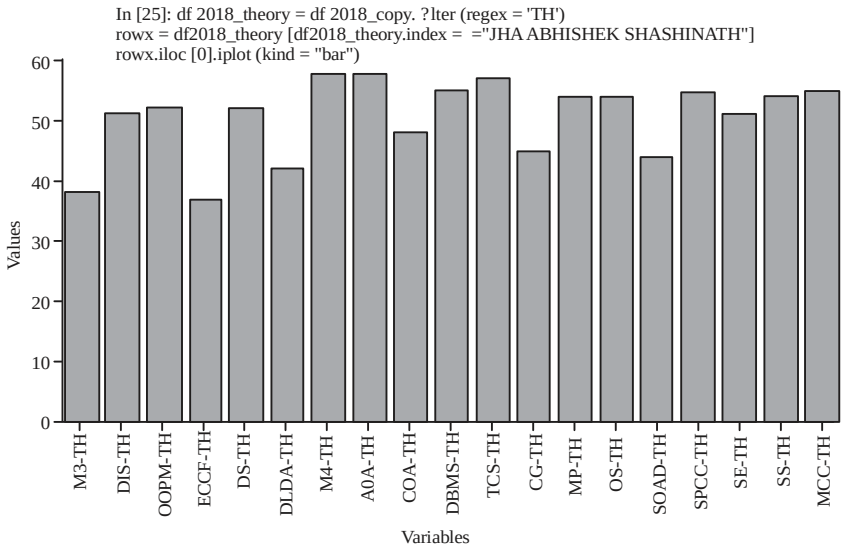


Fig. 8: Subject wise marks of a student

GNU octave, SVM, etc. which aren't efficient to work single-handedly in giving promising prediction (s) of real-world amendments of student's data results. So, if the concept of a neural network is infused in a placement prediction model then the efficiency and accuracy of the results will leap the average percentage of 35-97 percentile which lays the basic foundation of this research paper. The described model is incorporated with the special feature of choosing the ML algorithms among various choices individually or the mix-ups of algorithms per the compatibility of the data sets.

The graphs shown in the discussion section are derived from PCA and TSNE concepts which are widely used in prediction models. PCA is a method based on statistics that uses an orthogonal transformation for converting a set of correlated variables to a set of uncorrelated variables. Causes Euclidean distances to derive the components, thus the input variable needs to be numeric then box-plots for each numerical variable are generated separately just like in graphs here. To better visualize them a 3D plot is constructed using the three characteristic variables which show significant difference

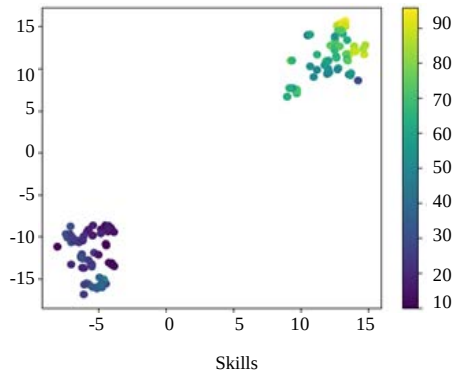


Fig. 10: Higher dimensions to 2D using TSNE

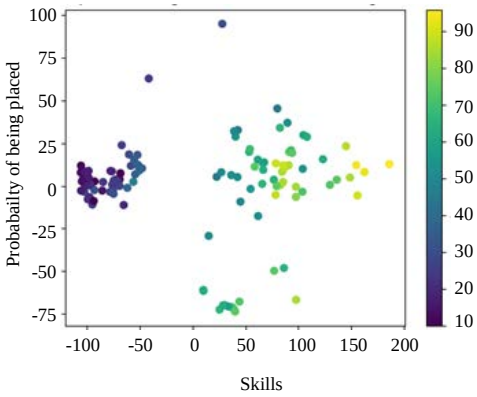


Fig. 11: Higher dimensions to 2D using PCA

among ratings It is used abundantly in exploratory data analysis and machine learning for predictive models and also examines the interrelation among a set of variables. It is also known as a general factor analysis where regression determines a line of best fit.

T-SNE is a technique in machine learning for dimensionality reduction which helps identify a relevant pattern. It is the profound merit of T-SNE to preserve local structure, i.e., that points which are close to one another in the high dimensional data set will tend to be close to one another in the chart. It also producers pretty looking visualizations. During the process of scanning raw data and calculating basic statistics can lead to some insights. Thus, t-SNE comes in the role of fitting multiple dimensions of data into a simple chart.

Principal Component Analysis (PCA) is used to visualize Higher Dimensional data into Lower Dimensions. In this method, the parameters are combined by using the TSNE method (sklearn decomposition PCA) to obtain a 2D plot. _Graph is as shown: _!(2D Visualization of 16D data) (Fig. 10-12).

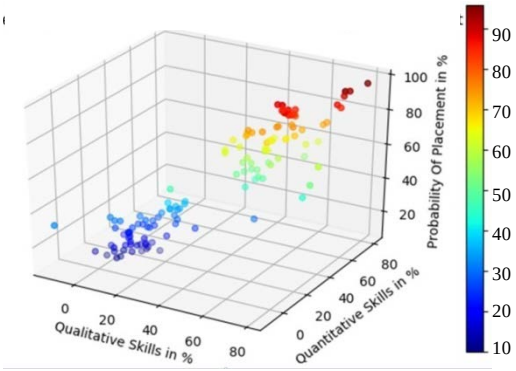


Fig. 12: Averaging the value of the parameters and represent it as a 3D plot. Graph is as shown_! (3D clour-coded plot of data)

T-Distributed Stochastic Neighbor Embedding (TSNE) is used to visualize Higher Dimensional data into Lower Dimensions. In this method, the parameters are combined by using the TSNE method (sklearn. manifold. TSNE) to obtain a 2D plot. Graph is as shown: _! (2D Visualization of 16D data). Averaging the values of the parameters and represent it as a 3D plot. Graph is as shown: _! (3D colour-coded plot of data).

CONCLUSION

As it has been seemed throughout our studies that the problem statements it was approached are student, college and corporate centric. The solution to all of these problem statements is based on the model it is going to build, the output of which will be a number between 0-1 which will determine, the prediction of a student being placed. During this process, a lot of other dependent variables will be predicted which will help solve the problem statements. The expected outputs of the system for student end are the prediction about their placement and the statistics of how they can fair well. The exactness that serves with this design is 97%. College end will have an analysis of every student and will have the opportunity to focus more on the improvement of students. Also, because of the system, the college will have one platform to manage the data of the students, thus, solving another issue. The corporates will be able to apply filters, compare students and download the resume of the students they're interested in also they will get student-related questions that they can ask in the interview.

REFERENCES

01. Bhatt, H., S. Mehta and L.R. D'mello, 2015. Use of ID3 decision tree algorithm for placement prediction. Int. J. Comput. Sci. Inf. Technol., 6: 4785-4789.

02. Taruna, S. and M. Pandey, 2014. An empirical analysis of classification techniques for predicting academic performance. Proceedings of the 2014 IEEE International Advance Computing Conference (IACC), February 21-22, 2014, IEEE, Gurgaon, India, pp: 523-528.
03. Kotsiantis, S.B. and P.E. Pintelas, 2005. Predicting students marks in hellenic open university. Proceedings of the 5th IEEE International Conference on Advanced Learning Technologies (ICALT 2005), July 5-8, 2005, IEEE, Kaohsiung, Taiwan, ISBN: 0-7695-2338-2, pp: 664-668.
04. Adhatrao, K., A. Gaykar, A. Dhawan, R. Jha and V. Honrao, 2013. Predicting students performance using ID3 and C4. 5 c lassification algorithms. Intl. J. Data Mining Knowl. Manage. Process, 3: 39-52.
05. Giri, A., M.V.V. Bhagavath, B. Pruthvi and N. Dubey, 2016. A placement prediction system using k-nearest neighbors classifier. Proceedings of the 2016 2nd International Conference on Cognitive Computing and Information Processing (CCIP), August 12-13, 2016, IEEE, Mysore, India, pp: 1-4.