

Analysis of Campus Placement Data: A Case Study

Aditi Indoori¹, Sathvika Patwari² and T Prathima³

¹Chaitanya Bharathi Institute of Technology/Information Technology, Hyderabad, India

Email: ugs18001_it.aditi@cbit.org.in

²⁻³Chaitanya Bharathi Institute of Technology/Information Technology, Hyderabad, India

Email: {Sathvika Patwari, T Prathima} ugs18016_it.sathvika@cbit.org.in tprathima_it@cbit.ac.in

Abstract—One of the main goals of the educational institution is the placement of students. An institution's reputation and annual enrolment rely inevitably on their students' placements. Analyzing the proportion of campus placements is one of the primal factors that students consider when selecting and applying to colleges. That is why all the institutions work hard to develop their organization in its entirety by strengthening the placements department. Any help in this specific field would have a positive effect on the campus placements. The students as well as the institution will benefit from this. In this context, the strategy is therefore to analyze and forecast the importance of campus placements, which not only help strengthen an institution's reputation but also optimize placements for students. Data pertaining to the study was collected from the institution for which suitable Data pre-processing methods like data blending, data Transformation, removing noise and inconsistencies are applied in order to analyze and mine useful insights which are then visualized in the form of an interactive dashboard. The charts surfaced in different formats offer concise ways to express the change in the placements in the past 11 years. To further aid this process, previous data has been used for future placement prediction using Machine Learning. Models are developed using the classification algorithms - Support Vector Machine, K- Nearest Neighbor, Decision Tree, Random Forest and Naive Bayes algorithm to validate the approaches.

Index Terms— Campus Placements, Data Pre-processing, Data Analysis, Data Visualization, Machine Learning, Support Vector Machine, K- Nearest Neighbours, Decision Tree, Random Forest.

I. INTRODUCTION

The main objective of this paper is to analyse Placements Data pertaining to the College and mine trends and patterns to understand better the factors influencing the Placements of students. Data relating to the Companies offering Placements is studied as well, in order to draw various inferences such as - Salary Trends, Branch Wise Recruitments etc. The paper also aims at developing a prediction model to determine Placement of Students, which enables students to gain insights into what aspects are important for College Placements and how they can improve their chance of being placed. Predicting the probability of students getting placed helps in uplifting their skills before the recruitment process starts. Performing such a comprehensive analysis would help both recruiters as well as students during College Placements and related activities.

The paper includes Data Analysis [1], Data Visualization, Training, building a model and testing. The user can enter details like Gender, CGPA, No. of Backlogs, Branch and the model predicts whether he/she can be placed or not. The model is deployed to a Flask Application.

A. Prediction System

In this paper, placement prediction is done by developing and training various machine learning models using the scikit-learn library. Models are developed using the classification algorithms SVM (Support Vector Machine), KNN (K- Nearest Neighbour), Decision Tree Random Forest and Naive Bayes algorithm to validate the approaches. The algorithms are applied on the data set and attributes used to build the model.

II. RELATED WORK

In [2], a model based on different algorithms like Naive Bayes, Decision Tree, Random Forest and KNN (K-Nearest Neighbour) is proposed, to know which one has high accuracy. This model will predict whether the student gets placed or not in campus recruitment. For this the data considers the academic history of students like overall percentage, skills, CRT training (Campus Recruitment Training) etc. After all the comparisons they concluded that Random Forest and KNN (K- Nearest Neighbour) showed high accuracy for their dataset.

In [3], a model based on the best algorithm from three Weka-based classification algorithms is proposed. They were Multilayer Perceptron, Naive Bayes Classification, and Decision Tree. The best algorithm based on the data is Naive Bayes Classification with an accuracy of 86.15%. This approach found an enhanced evaluation method for predicting the placement for students. This model can be used to assess the relationship between a student's academic success and their placement in the campus selection process.

In [4], the aim is to use linear regression techniques to construct a model that predicts the success of engineering students. Based on the data collected, the model's indicator or independent variables are the number of hours spent on the internet doing various activities. The estimation of end-of-semester examination grades, i.e., C.G.P.A., is the performance or dependent variable (Cumulative Grade Points).

In [5], the decision tree algorithm was used to predict the placement of students. They used Decision Tree (DT) algorithms like C4.5, ID3, and CHAID, which were created with the help of the Data Mining Rapid Miner software tool. The three algorithms are validated, and their accuracy is determined. Based on the collected placement results, the best algorithm from C4.5, ID3, and CHAID is ID3, which has 95.33% accuracy.

In [6], a model to classify the performance of the placement of students is built. The error rate for classifying validation data, result prediction classification tree, and validating placement prediction classification tree was found to be 38.46 percent and 45.38 percent, respectively.

In [7], a study on Student Placement Prediction was made using Machine learning models which included Naive Bayes Classifier and K-Nearest Neighbours [KNN] algorithm. The study highlighted the efficiency of the algorithms which used the data of the previously passed students and aimed to predict the placement probabilities of the current students. It presents a recommendation system that predicts whether the current student will be placed or not, if the student is placed the company is also predicted based on the data of previously placed students. Here two different machine learning classification algorithms, namely Naive Bayes Classifier and K-Nearest Neighbours [KNN] algorithm are used.

In [8], a model using linear regression is proposed where they have used the R-squared method to calculate the standard error rate. Linear Regression, a supervised machine learning method, can help with this by contributing its features. It is an effective method helping in predicting future trends of student placement based on advance placement practice test marks. The results will help students better understand their weak areas and where they should focus their efforts. Working on these areas will let students achieve a higher number of placements in an institution.

III. METHODOLOGY

The Placement Prediction Model is proposed with the intention of making the important Placement proceedings a smooth sail for College Students. It also has the potential to help the Company Recruiters by allowing them to investigate important statistics such as year wise and branch wise trends that will help them make informed decisions with respect to their recruitment process. The proposed system consists of 4 main stages - Data Analysis, Data Visualisation, building a Prediction Model, Deploying the Dashboard and Prediction Model to a Web Application.

A. Data Analysis and Data Visualisation

Data Analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

To perform Data Analysis, the first step is to obtain the Datasets. The Datasets used have been procured from the Placements Department of our college.

B. Description of the Datasets

The Datasets contain data pertaining to the last 11 years, that is from 2010-2021 in a Branch Wise and Class Wise format. Additionally, Company Wise, and Branch Wise data has also been obtained for the last ten years. The data is organized as shown in figure 1. The students' datasets are stored in a hierarchical manner. Root directory consists of folders in a year wise manner. Each year folder consists of the .csv files for the sections of each branch. In total there are 153 .csv files. Each dataset consists of data that gives deep insights into the students' placements. Columns such as CGPA, Gender, Branch, No. of placements are present. The **companies' datasets** consist of 11 .csv files, year wise from 2010 to 2021. Each file has data such as the name of the company, salary per annum (in lakhs), the number of placements for a particular company in each branch as well as the grand total of the number of placements. The Datasets undergo extensive pre-processing.

Once the Datasets are cleaned, exploratory data analysis is done in order to extract insightful information. This information is then displayed in the form of interactive charts and graphs.

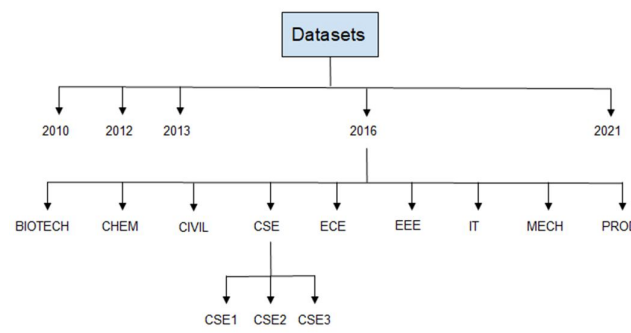


Figure 1. Students Dataset Structure

CLEANED DATASET

ALL							
	Gender	CGPA	No. of Placements	No. of Backlogs	Section	Year	Status
0	female	6.98	0	1.0	1	10	Biotechnology Not Placed
1	female	5.40	0	NaN	1	10	Biotechnology Not Placed
2	female	6.60	0	NaN	1	10	Biotechnology Not Placed
3	female	5.10	0	1.0	1	10	Biotechnology Not Placed
4	female	7.10	0	NaN	1	10	Biotechnology Not Placed
...	...	...	...	...	...	...	...
35	male	6.10	0	NaN	1	21	Production Not Placed
36	male	6.20	0	NaN	1	21	Production Not Placed
37	female	5.70	0	NaN	1	21	Production Not Placed
38	female	5.80	0	NaN	1	21	Production Not Placed
39	male	5.30	0	NaN	1	21	Production Not Placed

9532 rows x 8 columns

Figure 2. Cleaned Dataset for students' data

	Name of the Organization	Salary Per Annum (Rs. In Lakhs)	ECE	CSE	EEE	IT	MECH	PROD	CIVIL	CHEM	ME/EEE	ME/ArCadCam	ME/CIVIL	MBA	TOTAL	Year	ME/VLSI	ME/THERMAL
0	ASHWINARJA TELECOM	1.80	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	2	10	0.0	0.0
1	COROVIS	3.50	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	3	10	0.0	0.0
2	ANARJA	4.00	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	3	10	0.0	0.0
3	VINISOF	1.44	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	...	0.0	0.0	0.0	3	10	0.0	0.0
4	HONEYWELL	3.25	0.0	2.0	3.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	5	10	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
65	Sanghi Automotive Pvt Ltd	3.80	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	1	21	0.0	0.0
66	Godrej & Boyce Mfg	5.50	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	1	21	0.0	0.0
67	Modis Analytics	4.00	11.0	3.0	0.0	2.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	16	21	0.0	0.0
68	IndusMART InterMESH Ltd	3.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	1	21	0.0	0.0
69	Reputation.com	13.20	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	4	21	0.0	0.0

Figure 3. Cleaned & Merged Dataset for Companies data

C. Selection of Algorithm

To build a perfect Model with high accuracy, the first step is to analyse what all algorithms are suitable for the given Dataset. For our Data the classification algorithms used are, the details of these can be found in [9]:

1. SVM (Support Vector Machine)
2. KNN (K- Nearest Neighbour)
3. Decision Tree
4. Random Forest
5. Naive Bayes

D. Working of Machine Learning Algorithm

After thorough observation we found that Naive Bayes Algorithm is the best for our Dataset. The features used in our model are Gender, CGPA, Number of Backlogs, Branch. The data is converted to categorical type. It is then encoded using the .cat.codes method. The features are stored in variable X and the response vector is stored in variable y. The models are developed and are trained with 70% of the Dataset and tested with 30% of the Dataset.

E. Deploying to a Web Application

Once a prediction model is built, the next step typically entails deployment. The model is deployed using Flask to integrate it into an existing production environment which can take in any input and return an output which can be used by any student to know whether they have a chance of being placed or not. The data analysis and visualisation are presented in the form of a dashboard which is built using a python library called Dash. This dashboard is integrated to the existing flask application containing the predictor. Once the flask application is developed and all bugs are resolved, it is deployed to a Heroku web link.

IV. RESULTS

Random Forest Classifier is used to predict the importance of each parameter in the dataset with respect to Placements. It is apparent that CGPA has the most effect on Placements, followed by Branch, No. of Backlogs and Gender.

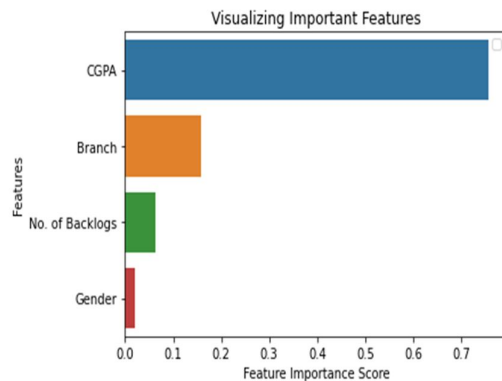


Figure 4. Visualising important features

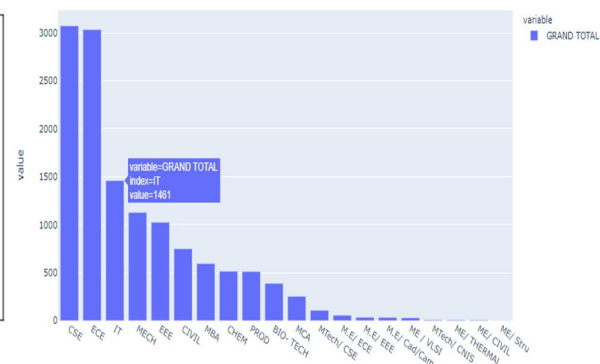


Figure 5. Branch Wise No. Of Placements

From Fig 5, it is evident that ECE and CSE are the top branches in terms of No. of Placements. From Figure 6, it is clear that a linear relationship exists between CGPA and No. of Placements. The higher the CGPA of a student, the more No. of Placements they receive. To have a better chance of being placed, the best bet would be to improve the CGPA.

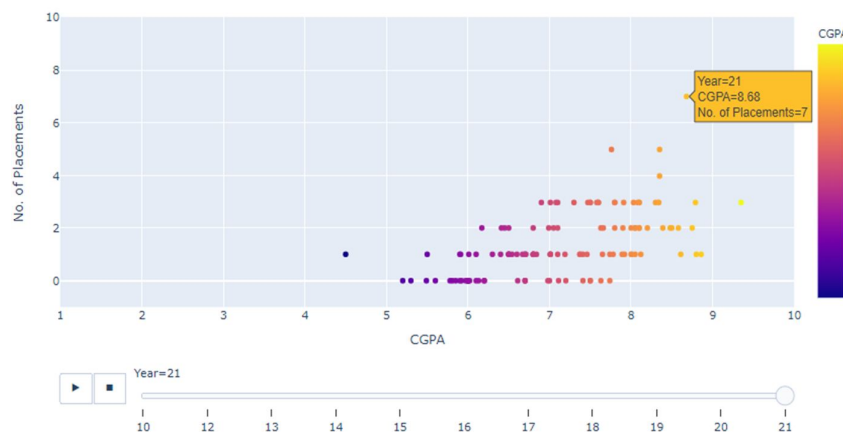


Figure 6. Relationship between CGPA and No. of Placements

TABLE I. ACCURACIES OF PROPOSED ALGORITHMS

Algorithms	Accuracy (%)
SVM (Support Vector Machine)	70.46
KNN (K- Nearest Neighbour)	67
Decision Tree	69
Random forest	70.43
Naive Bayes algorithm	72.8

V. CONCLUSIONS

It's a known fact that placements are what essentially determine and impact the number of students willing to join an Institution. Hence, with regard to this, a solution has been offered in the form of a Placement Analyser and Predictor. To improve the student's performance, previous data has been analysed, visualised and future prediction using a Machine Learning Model has been done. Models are developed using the classification algorithms SVM (Support Vector Machine), KNN (K- Nearest Neighbour), Decision Tree Random Forest and Naive Bayes algorithm to validate the approaches. The algorithms are applied on the data set and attributes used to build the model. The accuracy obtained after analysis for SVM (Support Vector Machine) is 70.46%, for KNN (K- Nearest Neighbour) is 67%, for Decision trees is 69% Naive Bayes is 72.8% and for the Random Forest is 70.43%. Hence, from the above said analysis and prediction we can conclude that Naive Bayes has the highest accuracy and is the best algorithm for our Dataset. Out of these Models, the best one is deployed to a Flask Web Application which takes in inputs such as CGPA, Gender, Branch and No. of Backlogs and returns an output that determines whether a student will be placed or not.

The accuracies of all the algorithms used ranges from 65-75, this is because the data was very noisy, there was a lot of missing data. Even after pre-processing of the datasets, the accuracy of the model was compromised due to insufficient data for some of the features.

The scope of this paper is not just confined to the existing state of art, there are many more tweaks that can be incorporated to make it a lot more beneficial on the whole. The Prediction Model is one of the most important parts of the project and it can be further improved to get better accuracy. It can also be enhanced to take in many more parameters, excluding the standard parameters, such as a student's dream company and predicting their chance of getting placed in that company along with feedback on how they can increase that chance. Predictions on how the average package of the institution can be boosted also will certainly prove to be extremely helpful to the Institution.

ACKNOWLEDGMENT

The authors would like to express their deep and sincere gratitude to their guide, Smt.T.Prathima for giving the opportunity to carry out the research and for providing invaluable guidance throughout the process. The authors would also like to thank the Head of their Department, Dr.K.Radhika, for providing the opportunity to improve their skills, and for providing them with the sources that were required for the completion of the paper. Sincere thanks to the Principal, Prof. G P S Varma, as well as the management for giving a good opportunity to improve skills with a good learning atmosphere.

REFERENCES

- [1] William McKinney, "Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython", Second Edition, O'Reilly Media, 2017
- [2] Ankita Mahalle, Divisha Samrit, Nishigandha Wagh, Rucheeka Gothe "campus placement predicting using data science", ijetemr, Vol. 7, June 2020
- [3] Ajay Kumar Pal, Saurabh Pal "Classification Model of Prediction for Placement of Students", I. J. Modern Education and Computer Science, Published Online in MECS (2013).
- [4] P.Natesan , N.Krishnamoorthy, S.Ponni,R.R. Rajalaxmi, "Regression Model for Predicting Engineering Students Academic Performance", researchgate, April 2019.
- [5] T. Jeevalatha,N. Ananthi,D. Saravana Kumar, "Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms", International Journal of Computer Applications (0975 – 8887),Volume 108 – No 15, December 2014

- [6] Neelam Naik and SeemaPurohit, "Prediction of Final Result and Placement of Students using Classification Algorithm", Interna-tional Journal of Computer Applications (0975–8887), Volume56–No.12, October 2012
- [7] Shreyas Harinath, Aksha Prasad, Suma H S, Suraksha A and Tojo Mathew, Student placement prediction using machine learning, (IRJET), e-ISSN: 2395-0056, pISSN: 2395-0072, Volume: 06 Issue: 04 | Apr 2019.
- [8] G.Gautami, Monam Hayat,"Prediction of Student Placement using Machine Learning Algorithms",IJCRT | Volume 8, Issue 5, May 2020
- [9] Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining: Concepts and Techniques", The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, 3rd Ed, ISBN 978-0123814791, (2011)