International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Undergraduate Student's Campus Placement Determination Using Logistic Regression Analysis for Predicted Probabilities on Uncertain Dataset

Chandra Sekhar K[1*], K Santhosh Kumar[2]

Submitted: 14/08/2022 **Accepted:** 16/11/2022

***Abstract*—**Undergraduate students in technical institutions aim to secure a placement within four years of the Course. Many factors impact student placement chances. Analysis and understanding of the factors influencing the student chances could change the orientation of coming generations towards education. This work is a clear understanding of the placement factors affecting students' chances and an illustration of Logistic Regression. In this paper, Logistic Regression accurately predicts which factors influence graduate placement opportunities. Using a uncertain dataset that comprises more than 1000 students' information to find predictions. We have shown the predicted probabilities of each student who can secure a job along with actual job status. The prediction probability calculations and machine learning model predictions are compared and found to be equal. The approach can be used to guide the coming generation of students to have a note of factors that could influence their placement chances.

Keywords*—Logistic Regression; Undergraduate; Placement; Prediction; Probabilities; Coefficients; Features*

1. Introduction

Educational institutions perform further analysis and predictions for the placement outcome of students. Whether a student has a good chance of attaining a job opportunity or not can be predicted by different machine learning algorithms. Since the use of logistic regression has increased over the years, the use of logistic regression could yield good results in this particular problem of student placement. The logit is the mathematical concept that causes logistic regression. In this paper, logistic regression impact of Coefficients on the Placement, chances are determined to predict student placements. The Prediction probability calculation is done and compared with the actual values of data. Identifying the factors to find the impact of each factor upon Placement is critical. The probability calculations are done, and a detailed comparison is made with model-generated values.

2. Literate Survey

The desired pattern for the use of logistic regression techniques, as well as an instance of logistic regression applied to data for gender and recommendations for remedial reading instruction. For logistic regression reports, a new set of Suggestions is required. According to the observation to predictor ratio, only the most relevant data are taken into account. This work gauged the usage & clarification of the logistic reversion. This work is mainly comprised of 5 different sections logistic regression models, Design of logistic regression models, Guidelines and recommendations, Evaluation, and a final model summary. It has been decided to use logistic regression to assess 189 referrals for the Remedial Class Reading Program. Statistical tests of individual predictors are limited to a few parameters. Predicted probabilities are validated only for variables named Gender [1].

This research forecasted the difficulties that students may encounter in the course design process. This approach helps identify the student dropping before final examinations and helps teachers handle scenarios like this. With the use of artificial intelligence, we examined the log files from an educational software programme called Digital Electronics Education and Design Suite. Along with artificial neural networks, machine learning algorithms like support vector machines, logistic regression, naïve Bayes classifiers and decision trees are used. This method allows the student to give different inputs depending on the session and parameters. There are two metrics used to assess the model's performance: receiver operating characteristic and root mean square error. The systems grades are generated by doing k fold cross validations. The instructor's involvement and assessment of grades by collecting data from the system has proven to be a hard task[2].

Job performance is considered an essential parameter in assessing institutional and educational quality. Linear and logistic regression are the models employed to find job performance. To accurately predict the experience on the Educational Evaluation Index, linear regression is used. Used logistic regression to find the level of EEI sustained by educationists. The institutional, educational quality is assessed by student feedback, one of the concerning factors. Overall few parameters from the feedback of students are used to conclude. Organizational DSS in evaluation is less than 60 percent, so some measures must be taken to improve performance[3].

Online student learning platforms and student educational data analytics have developed personalized learning platforms. Due to that, the prediction has become a favourable research field. Online learning systems that include the whole learning process are the focus of this study. Used an optimized logistic regression algorithm to analyse student behaviour and predict performance. The Hstar teaching platform was chosen as a case study for the course's cloud

[1]*Research Scholar, Department of Information Technology, Annamalai University, Annamalai Nagar, Chidambaram-608002, India*
ORCID ID: 0000-0001-7052-3008
[2]*Assistant Professor, Department of Information Technology, Annamalai University, Annamalai Nagar, Chidambaram-608002, India*
**Corresponding author: sekharonemay@gmail.com, santhosh09539@gmail.com@gmail.com*

computing in student education section. It used enhanced Logistic Regression to predict whether or not a student is exceptional based on their behavior while reading course materials in Hstar. [4].

3. Proposed System

Processes involved in determining coefficients using Logistic Regression on undergraduate data are initial Data cleansing and making necessary modifications to the raw data. Then, identifying the necessary columns and making them factors. The different influencing factors are represented below in figure 2. In the next step, a student with Job and without Job representation is done. 80% of the data will be used for training, and 20% will be used for testing, resulting in two separate sets of data. Logistic regression Impact of CGPA and all other parameters' influence in securing a Job is noted. Probability Calculations are done and compared with model-generated values. The entire work is represented using a block diagram in figure 2. In this paper, the predictor, the logistic regression model, is fitted to the uncertain dataset of student placement, and the comparisons are made. Accuracy prediction and model building are done. The probability calculations are compared with manual calculations and found to be equal.

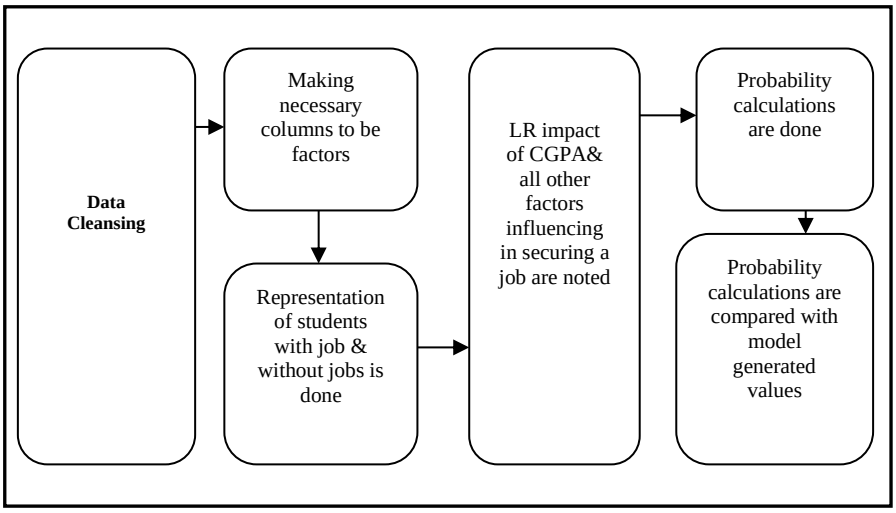

Figure. 1. A complete outline of linear regression application on the Graduate Placement uncertain dataset.

The above figure shows a block representation that explains the steps involved in determining the coefficients using logistic regression to predict probabilities [5]. The steps involved are data pre-processing and probability calculations compared with model-generated values.

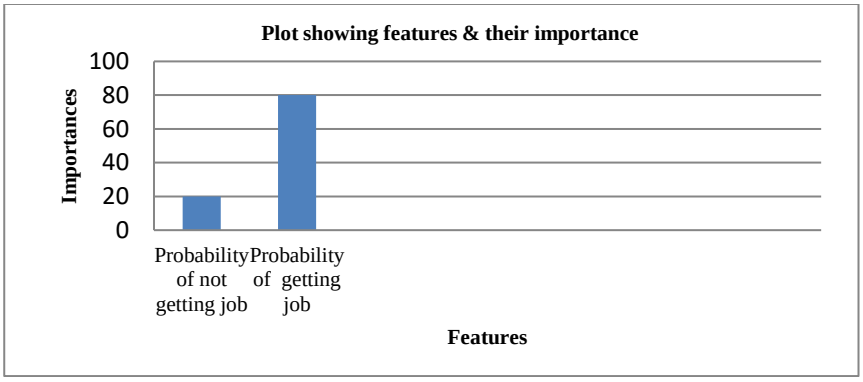

Figure. 2. Figure showing Weights of all features and their importance.

Identifying features and their importance is crucial in applying machine learning algorithms to the given uncertain dataset. For example, in the above figure, the features like EAMCET rank and B.E% have high importance compared to other features [6]. This means these factors will play an essential role in determining the placement chances of graduate students.

4. Data set for Logistic Regression

The table 1 shows the different parameters taken into account to construct student data.

Table 1. The data set comprises student data for four consecutive years

S.No	Features	Description
1	S.No	This column defines the serial number of the student.
2	Gender	This column defines the Gender of the student.
3	Branch	This column defines the student's branch in his/her Btech graduation.
4	X^{th} %	This column defines the 10^{th} Grade percentage of marks of the student.

5	**YOP**	This column defines the year of pass of every graduate.
6	**Inter %**	This column defines the inter marks percentage of the student.
7	**BTech %**	This column defines the Btechcgpa of the student.
8	**Backlogs**	This column defines the number of backlogs for the student.
9	**Selected Company**	This column defines the company for which the student got placed.
10	**Eamcet**	This column defines the Eamcet (Engineering entrance exam) rank of the student.
11	**College**	This column defines the college of the student.
12	**University**	This column defines the university of the student.
13	**DOB**	This column defines the date of birth of the student.

The parameters include the student data from their class Xth CGPA to Engineering (Graduation), i.e., Btech CGPA. Their Engineering entrance examination rank, Gender, college, and university. Using the data collected from the institute, we are proceeding with applying machine learning techniques[7].

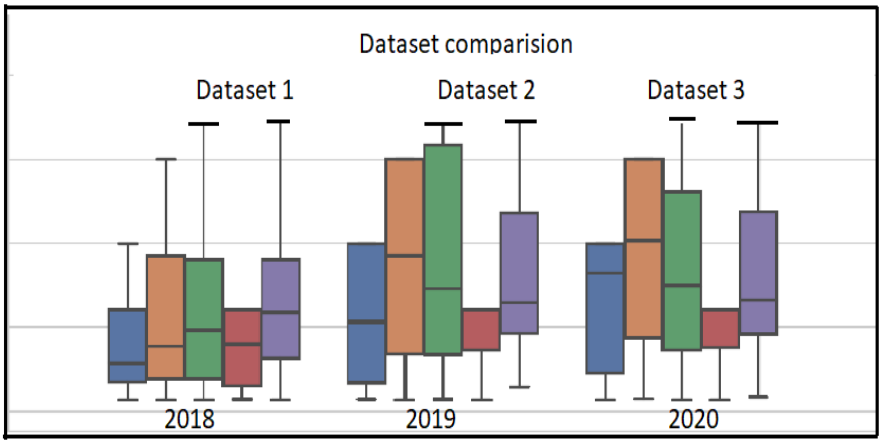

Figure. 3. Showing uncertain dataset Comparison of 2018, 2019 and 2020

Using the box plot method, we tried to explain the spread of student placement data for four consecutive years through their quartiles [8]. The box plot contrasts the Eamcet rank of the student distribution on the x-axis with the student branch on the y-axis. We have shown different student departments on the y-axis.

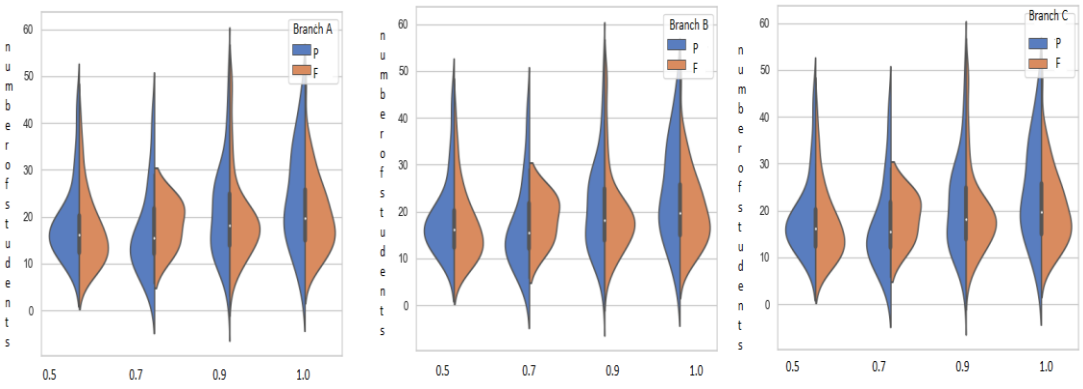

Figure. 4. Violin plot checks Graduation CGPA with Branch.

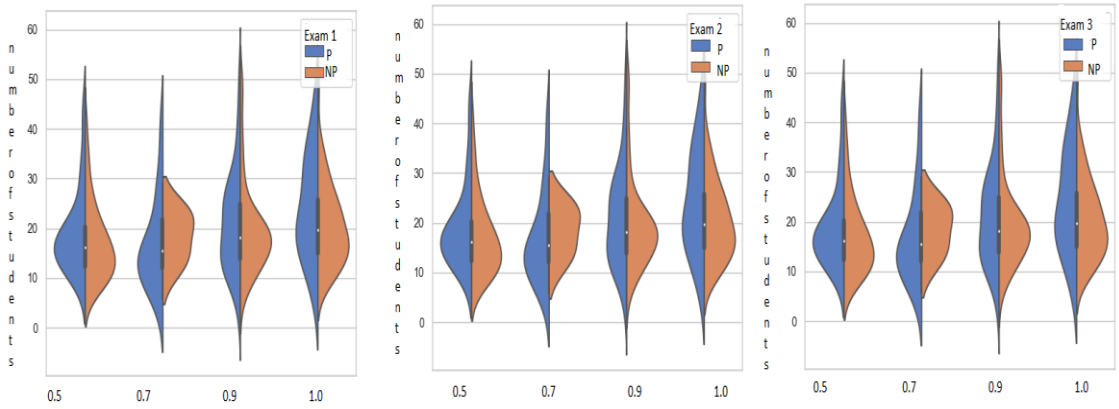

Figure. 5. Violin plot checks the Entrance exam rank with the Placement of the student.

The above figures are violin plots that represent the median and interquartile range. The first representation, figure 4, represents the kernel density estimation which shows the distribution of Graduate CGPA concerning different branches(departments). The box plot elements show that the median weight for some branches is less when compared with others, and it varies with different years[9]. Similarly, figure 5 represents the kernel density estimation of the entrance exam rank to the Placement of the students. This violin shows the median weight of the students with Placement is around the higher rank region when compared with placed students whose rank is below the higher rank range. Also observed was that over the years, the students placed have an Eamcet rank(entrance exam rank) in the range below the median[10].

5. Regression

Logit functions are the central mathematical function behind the working of Logistic regression. Consider our problem of placements depending on factors like CGPA, Backlogs, and Xth percentage. The Placement chances are dependent on factors like CGPA, Backlogs, and percentages. These factors are the predictor variables. The logistic regression is ideal for explaining the association between a categorical variable and a continuous predictor variable. Using linear regression, it is not easy to describe the multiple factors, each corresponding to different outcomes. Logistic regression can be used to solve these challenges by applying logit transformation to the dependent variable.

Regression models observe the relationships among variables by sizing a line to the given data. If they are Linear Regression models, they use a straight line, while the models are said to be logistic and nonlinear representations; they use up a curved line. To model the relationship between two variables.

The logistic model has the below form

$$Y = mX + b; \quad (1)$$

Y is treated to be the dependent variable
X is treated to be the independent variable
Where is y-intercept

$$m = \frac{(n(\sum(X \cdot Y)) - (\sum X \cdot \sum Y))}{(n \cdot \sum X^2 - (\sum X^2))} \quad (2)$$

$$b = \frac{\left((\sum Y \cdot \sum X^2) - (\sum X \cdot \sum(X \cdot Y))\right)}{((n \cdot \sum X^2) - (\sum X^2))} \quad (3)$$

$$Y = m_1 \cdot X_1 + m_2 \cdot X_2 + m_3 \cdot X_3 + b \quad (4)$$

Where m is the slope defined in Equation 2, b is the y-intercept, b is defined in Equation 3, and Y is the outcome of interest Y is categorical and is defined in Equation 4. X_1, X_2 … are a set of predictors in Equation 4.

6. Undergraduate Placement Logistic Regression Analysis

A specific year's data was used to create a predictive logistic model. Then we evaluated our study hypothesis, which was that there is a connection between a specific factor and the likelihood of a student being placed. The Logistic regression [11] was carried out using python. Initially, we found the weight of the features and then can calculate the predicted values of the new record when an entry was given manually. After calculating the features' weights, we checked the results manually by giving random student values to the model [12].

Probability Calculation for Logistic Regression

$$\text{Logit}(Y) = \ln\left(\frac{P}{1-P}\right) = b_0 + b_1X_1 + b_2X_2 \ldots + b_nX_n \quad (5)$$

P is the Probability of Accepting(Probability of Student Getting Placed), and (1-P) is the Probability of Rejection(Probability of Student Not Getting Placed) in equation 5

$$Y = \ln\left(\frac{P}{1-P}\right) \quad (6)$$

$$P = \frac{1}{(1+e^{-Y})} \quad (7)$$

P is the Probability Student Getting Placed in equation (7)

The Probability calculation for the given student data is done in the below approach

$$Y = m_0 + m_1X_1 + m_2X_2 + m_3X_3 + m_4X_4 + m_5X_5 + m_6X_6 + m_7X_7 \quad (8)$$

Here m_0 is intercept, $m_1, m_2, m_3, m_4, m_5, m_6, m_7$ are the Feature Weights and $X_1, X_2, X_3, X_4, X_5, X_6, X_7$ are the values of Features from equation 8. By using the feature values and weights, we calculate the probability. The value of probability will be Between 0 and 1.

Herewith is our data set; we have calculated our seven weights and intercept value.

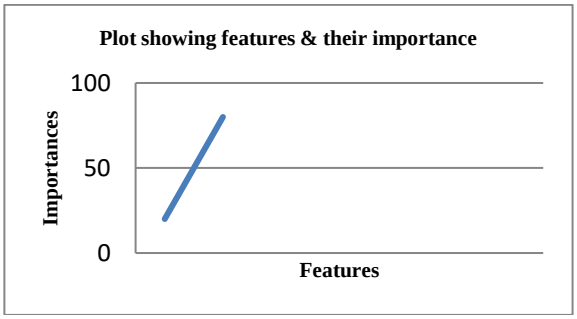

Figure. 6. Showing the Intercept value for the student data set of 2019, the value is -0.9770

```
# model training
Step 1:  reg_std = linear_model.LogisticRegression()
# next train the present model using the training sets
Step 2:  reg_std.fit(x_train_std, y_train_std)
Step 3:  LogisticRegression()
# model prediction and accuracy calculation
Step 4: y_pred_std = reg_std.predict(x_test_std)
# and then comparing actual response values (y_test)
with predicted response values (y_pred)
Step 5: print("Logistic Regression model accuracy(in
%):",metrics.accuracy_score(y_test_std,
y_pred_std)*100)
Logistic Regression model accuracy(in %):
78.87700534759358
# finding the intercept value
Step 6: m0 = reg_std.intercept_[0]
Step 7: m0= -0.9770799319454345
```

Figure 7. Weights of all the features and all features are noted, and their values are shown below

The above figure 7 represents different features and based on the features the values for the Weights are retrieved were $m_1 = -0.08769014$, $m_2 = 0.02659039$, $m_3 = -0.06504804$, $m_4 = -1.11514301$, $m_5 = 1.10440895$, $m_6 = 0$, $m_7 = -1.66028645$. The Backlogs are the least influential feature, which have less impact on placement chances. Now substituting the above values in equation 8

$$Y = -0.977080 + -0.08769014 * X_1 + 0.02659039 * X_2 + -0.06504804 * X_3 + -1.11514301 * X_4 + 1.10440895 * X_5 + 0 * X_6 + -1.66028645 * X_7 \quad (9)$$

Here features and their relations are mentioned $X_1 = \text{Gender}$, $X_2 = \text{Branch}$, $X_3 = \text{Xth Percentage}$, $X_4 = \text{Inter Percentage}$, $X_5 = \text{Btech Percentage}$, $X_6 = \text{Backlogs}$, $X_7 = \text{Eamcet Rank}$ from the given uncertain dataset

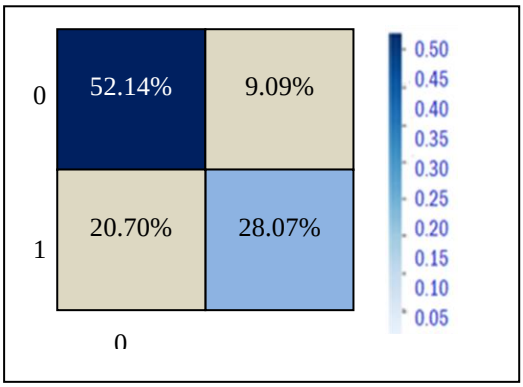

Figure. 8. The confusion matrix representation for the 2019 uncertain dataset

This implementation of the logistic regression model is described by employing the above confusion matrix. This confusion matrix allows the visualization of logistic regression[14] (seen in figure 8).

```
# Defining the function to make the calculation
Step 1: import math
Step 2: def sigmoid(x):
Step 3: return i / (1 + row (math.e, -x))
Step 4: Take result = 0
Step 5: result += m0
Step 6: for i in range (0, 7):
Step 7: result += custom_data_std_tran[0] [i] *
m  [i]
#prinitng the calculated values of z and y of the
formulas
Step 8: print (' value of Z ') and print (result)
Step 9: result = sigmoid(result)
Step 10: print (' y predicted value ') and print (result)
Step 11: value of z = 1.470498792541516
Step 12: Y predicted value is 0.8131331882026058
```

Figure. 9. The outcome of Interest for the predicted variable is derived from the above process.

Considering a student entry with features like Gender = 1 (male), branch = 0 (civil), Xth percentage = 9.5, Inter percentage = 97.5, btech percentage = 82.6, Backlogs = 0, eamcet rank = 9295.0. Defined functions to make calculations and found the values of z and Y which is the dependent value (seen in figure 9).

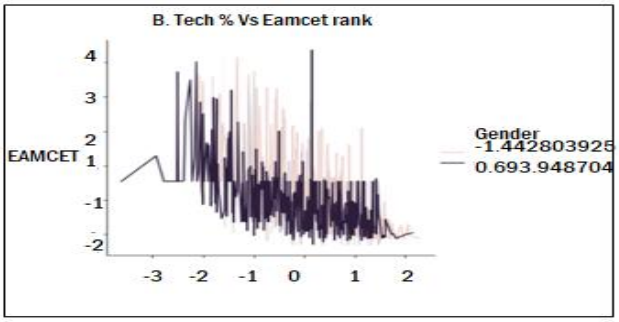

Figure. 10. Code taking the same person's data to find the probability of Placement.

Convert this data as a NumPy array. We transformed this data with a standard scaler using the st. transform method. Using reg_std. Predict () method to predict the data with our previously trained model. The model gives one as output, which means it will predict that person will get the Job [15] (seen in figure 11).

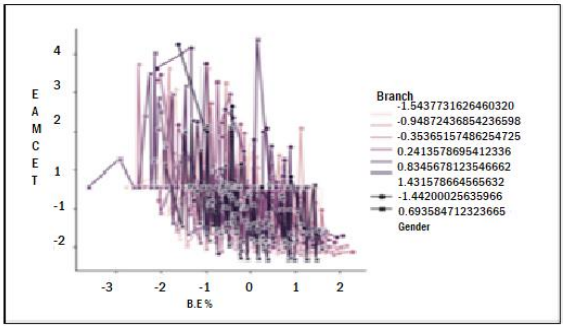

Figure. 11. Graduation Cgpa to Entrance exam rank distribution explained with respective to Gender.

Now substituting these values in equation 9

$$Y = -0.977080 + -0.08769014 * 1 + 0.02659039 * 0 + -0.06504804 * 9.5 + -1.11514301 * 97.5 + 1.10440895 * 82.6 + 0 * 0 + -1.66028645 * 929 \quad (10)$$

The resultant value is $Y = 1.470498792541516$ by placing the value of Y in 7.

$$P = \frac{1}{(1+e^{-1.4704})} \quad (11)$$

The final value for P =0.8131 from equation 11
Here the probability of getting placed is 81 percent. let us prove the same with python code taking the same persons data Gender = 1 (male), branch = 0 (civil), Xth percentage = 9.5, Inter percentage = 97.5, BTech percentage = 82.6, Backlogs = 0 , EAMCET rank = 9295.0 as an input (seen in figure 10).

```
#predicting a value by taking a sample data entry
Step 1: nn = [[1,0,9.500,97.50,8.26,0,9295.0]]
Step 2: custom_data_std=np.array([[1,0,9.500,97.50,8.26,0,9295.0]])
# transforming the taken data with fitted standard scaler
Step 3: custom_data_std_tran=std. transform(custom_data_std)
Step 4: custom_data_std_tran =  array([[ 0.69309487, -1.54377318, 0.51533173, -0.89463018]])
Step 5: custom_data_prediction_std=reg_std.predict(custom_data_std_tran)
# predicted value by the model
Step 6: custom_data_prediction_std
Step 7: array([1], dtype=int64)
# predicting the probability of getting the placement of particular person
Step 8: y_reg_val5 = reg_std.predict_proba(custom_data_std_tran)[0]
Step 9: print(y_reg_val5[1])
Step 10: 0.8131331882026058
# probability of getting the placement and not getting one
Step 11: y_reg_val5
Step 12: array( [0.18686681, 0.81313319])
```

Figure. 12. Plots Showing Transform () and predict () function output and distribution of different data features.

```
#feature_names = ['Gender', 'Branch', 'X th%', 'Inter %', 'B.E%', 'Backlogs', 'EAMCET', 'Placement']
Step 1: nn_v = [[1,0,9.200,95.50,8.5,6,15000.]]
Step 2: custom_data_std_v=np.array([[1,0,9.200,95.50,8.5,6,15000.0]])
# transforming the taken data with fitted standard scaler
Step 3: custom_data_std_tran_v=std.transform(custom_data_std_v)
Step 4: custom_data_prediction_std_v=reg_std. predict (custom_data_std_tran_v)
# predicting the probability of getting the job of that person
Step 5:  y_reg_val5_v = reg_std. predict_proba(custom_data_std_tran_v)[0]
Step 6: print(y_reg_val5_V[1])
# Probability of getting the job and not getting the job
Step 7: y_reg_val5_V
Step 8: my_dict = {u'Probabilty of not getting job': y_reg_val5_v[@]*100 , u'Probability of getting job':
y_reg_val5_v[1]*100}
Step 9: print(my_dict)
Step 10: plt.bar("zip("my_dict.items()))
plt.show() 0.8107861853964691
{'Probabilty of not getting placement': 18.921381460353093, 'Probability of getting placement':
81.07861853964691}
```

Figure 13. Deriving probability of Placement

The above plot is the distribution of the features like EAMCET rank, B.E % (or Btech percentage graduation Cgpa), Branch of the student, and Gender (seen in figure 12).
Now we have to check with what percentage the model was sure at its prediction. For that, we use reg_std.predict_proba method to see which probabilities the model is classifying the predictions[16] [17].

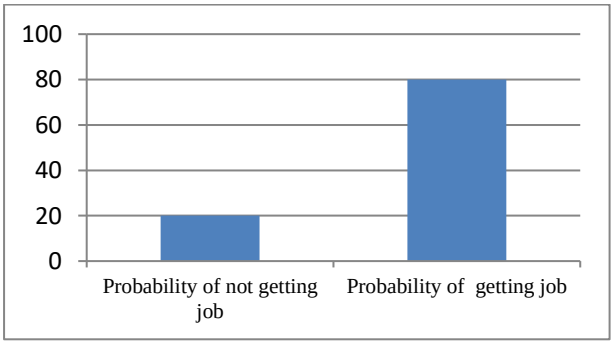

Figure 14. Showing Final probability of placed and unplaced students

We get an array of [[0.19 0.81]] using the predict_proba() method [18][19]. The model shows showing 81 percent probability of getting the Job which is accurately equal to our calculation with the formula as it also gave us the same 81 percent probability. Features and their values are collected and transformed the taken data into a standard scaler. They are finally predicting the chances of Placement [20].

7. Conclusion

In this paper, we have explained that Logistic regression is an important analytical technique for use in Graduate placement analysis. The importance of the logistic model is shown by the test of predictor and predicted probabilities. The use of logistic regression could yield better results in the field of education. The working results of the logistic regression model and calculations of probabilities are found to be accurate. Among the features of the uncertain dataset taken, the EAMCET rank is proven to be the most effective feature in determining student placement. Along with this feature, other influential coefficients are determined.

8. Acknowledgment

This project would not have been feasible without the support of my supervisor, k Santhosh Kumar. I could not have completed this project without his help and suggestions.

References

[1] Peng, C. Y. J., & Lee, K. L. 8c Ingersoll, GM (2021). An introduction to logistic regression analysis and reporting. Thejournal ofEducational Research, 96(1), 3-14.

[2] Saa, A. A. (2016). Educational data mining & students' performance prediction. International Journal of Advanced Computer Science and Applications, 7(5).

[3] Arora, S., Agarwal, M., & Kawatra, R. (2020, March). Prediction of educationist's performance using regression model. In 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 88-93). IEEE.

[4] Zhang, W., Huang, X., Wang, S., Shu, J., Liu, H., & Chen, H. (2017, June). Student performance prediction via online learning behavior analytics. In 2017 International Symposium on Educational Technology (ISET) (pp. 153-157). IEEE.

[5] Ko, C. Y., & Leu, F. Y. (2020). Examining Successful Attributes for Undergraduate Students by Applying Machine Learning Techniques. IEEE Transactions on Education, 64(1), 50-57.

[6] Perez, B., Castellanos, C., & Correal, D. (2018, May). Applying data mining techniques to predict student dropout: a case study. In 2018 IEEE 1st colombian conference on applications in computational intelligence (colcaci) (pp. 1-6). IEEE.

[7] Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. IEEE Journal of Selected Topics in Signal Processing, 11(5), 742-753.

[8] Mirzaei, S., Sidi, T., Keasar, C., & Crivelli, S. (2016). Purely structural protein scoring functions using support vector machine and ensemble learning. IEEE/ACM transactions on computational biology and bioinformatics, 16(5), 1515-1523.

[9] Botto-Tobar, M., Vizuete, M. Z., Torres-Carrión, P., León, S. M., Vásquez, G. P., & Durakovic, B. (Eds.). (2020). Applied Technologies: First International Conference, ICAT 2019, Quito, Ecuador, December 3–5, 2019, Proceedings, Part I (Vol. 1193). Springer Nature.

[10] Y. Liu, W. Cao, Y. Liu, and W. Zou, "A Novel Ensemble Learning Method for Online Learning Scenarios," in 2021 IEEE 4th International Conference on Electronics Technology, ICET 2021, 2021, pp. 1137–1140, DOI: 0.1109/ICET51757.2021.9451004.

[11] R. Miceli, "A coefficient of determination for logistic regression models," TPM - Testing, Psychom. Methodol. Appl. Psychol., vol. 14, no. 2, pp. 83–98, 2007, [Online]. Available: https://www.tpmap.org/wp-content/uploads/2014/11/14.2.2.pdf.

[12] Buenaño-Fernández, D., Gil, D., & Luján-Mora, S. (2019). Application of machine learning in predicting performance for computer engineering students: A case study. Sustainability, 11(10), 2833.

[13] Golding, P., & Donaldson, O. (2006, October). Predicting academic performance. In Proceedings. Frontiers in Education. 36th Annual Conference (pp. 21-26). IEEE.

[14] Guezzaz, A., Asimi, Y., Azrour, M., & Asimi, A. (2021). Mathematical validation of proposed machine learning classifier for heterogeneous traffic and anomaly detection. Big Data Mining and Analytics, 4(1), 18-24.

[15] D. Kumar, C. Verma, P. K. Singh, M. S. Raboaca, R. A. Felseghi, and K. Z. Ghafoor, "Computational statistics and machine learning techniques for effective decision making on student's employment for real-time," Mathematics, vol. 9, no. 11, 2021, DOI: 10.3390/math9111166.

[16] Mengcan, M. I. N., Xiaofang, C. H. E. N., & Yongfang, X. I. E. (2021). Constrained voting extreme learning machine and its application. Journal of Systems Engineering and Electronics, 32(1), 209-219.

[17] Saikumar, K., Rajesh, V., Babu, B.S. (2022). Heart disease detection based on feature fusion technique with augmented classification using deep learning technology. Traitement du Signal, Vol. 39, No. 1, pp. 31-42. https://doi.org/10.18280/ts.390104

[18] Kailasam, S., Achanta, S.D.M., Rama Koteswara Rao, P., Vatambeti, R., Kayam, S. (2022). An IoT-based agriculture maintenance using pervasive computing with machine learning technique. International Journal of Intelligent Computing and Cybernetics, 15(2), pp. 184–197

[19] Saikumar, K. (2020). RajeshV. Coronary blockage of artery for Heart diagnosis with DT Artificial Intelligence Algorithm. Int J Res Pharma Sci, 11(1), 471-479.

[20] Saikumar, K., Rajesh, V. (2020). A novel implementation heart diagnosis system based on random forest machine learning technique International Journal of Pharmaceutical Research 12, pp. 3904-3916.