Smart Student Placement Prediction Using Machine Learning

MANOJ KUMAR CHOUHAN[1], ADITYA KUMAR[2], KULDEEP SINGH SONGARA[3], KHUSHAL RAWAL[4] , Dr. VISHAL SHRIVASTAVA[5], Dr. AKHIL PANDEY[6], Dr. DEVESH BANDIL[7]

[1,2,3,4]B.TECH. Scholar,[5,6]Professor, [7]Project Guide
Computer Science & Engineering
Arya College of Engineering & I.T. India, Jaipur

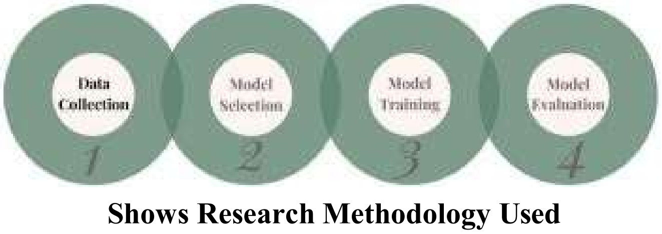
1. Abstract:

When a student picks a college, there are different factors that enter the equation. Students typically look for the end result and, in this instance, it is placement. Every college tries to have an effective Placement team. The purpose of this research paper is to improve the overall functioning of the system by implementing a student placement prediction system. The objective is to enhance the overall accuracy of the system. Multiple parameters are considered to forecast whether the student is placed or not. This paper examines different machine learning algorithms like Logistic regression, Random Forest, KNN, SVM. These machine learning algorithms will be utilized to forecast the outcome on a shared database separately. The ultimate output will be compared to each other in order to analyze which algorithm is most accurate and performs the best for the system.

2. Introduction

Getting after-school work is a very crucial time for students. Similar to the last test, it's all your hard work! Teachers typically review their grades and thesis to determine if the students are a good fit for the company. But would you advise you? Now there is a new method to magical performance, guessing whether students will be employed. Compare various comparisons to determine which behavior glides the most smoothly. You will also examine the perspectives behind the curtains to determine what information these programs employ to make inferences about your guesses. This data, referred to as features, may be test scores, a student's performance in a school subject, or doing extra projects outside of class. By assessing the precision of these ml systems in predicting job prospects, we can gain valuable insights[2][5][7]. We Can we determine if these educational programs are truly advantageous for students or not? Ultimately, this data can be utilized to construct an even. Improved machine learning programs are expected in the future. These highly intelligent programs can then aid students and schools in acquiring. Prepared for the job search in a smarter and more successful manner[6][8].

3.Research Methodology



Shows Research Methodology Used

1. Data Collection –

This paper is a research study that focuses on the campus placement process in the current challenging situation. Specifically focusing on computer science and how different factors impact the placement statistics of a college. This research paper discusses the different factors that contribute to this and the specific skills or criteria that are strictly adhered to. Various organizations[1][2][6].

Methodology employed for gathering data.

1: Gathering of specimens from pupils.
2: Cell of the college for training and placement.
3: The process involved extracting historical open source placement data from various college websites[1].

The data collected from within the college was structured in a specific way, following a particular schema or format.

1: Student Characteristics.
2: Academic performance.
3: Test scores.
4: Internship Overview.
5: Certifications:

The data that we used for the analysis had been finalized after a thorough evaluation, and the columns are now ready[5][10].

2. Model Selection & Training-

Explain the machine learning algorithms selected for predicting placement. Typical options for classification tasks are Logistic Regression Logistic regression is an appropriate approach to predict student placement since it's a supervised classification task that handles binary responses, exactly fitting the "placed" or "not placed" situation[2]. Decision Trees - Decision trees, one of the most popular supervised learning techniques, emulate real-world decision-making with a branching model. Decision trees are both excellent at classification (categorizing data) and regression (predicting continuous values)[3][4].

Here's a list of their major components-

- Root Node- The beginning, similar to the trunk of a tree.
- Splitting- Splitting data points along the value of a selected feature, similar to arranging apples according to color.
- Decision Node- Puts a question to a feature, steering the data along certain paths.
- Leaf Node- The ending point, consisting of a forecasted class (e.g., "red apple") or a continuous value

3. Model Training and Evaluation-

The training of the model is executed using the train-test split technique available in the sci-kit learn library. The information is partitioned into 2 distinct categories[2][4][7].

A) Training data(70%) - this data set is utilized to train the data and see if it fits the model properly. Or not, and how the model acquires different patterns from the data.

B) Testing data(30%) - this section is typically used to verify the accuracy of the dataset and how it is utilized. Modeling works in providing unknown values and how it predicts or classifies the necessary output.

Training of our model[2].

Normalization -.

Feature scaling, also known as normalization or standardization, is a technique used in machine learning to adjust numerical features to a common scale. - Learning models are of equal importance during training. Variations in scales (such as income versus age) can potentially distort the model's results. Towards features with larger ranges. Scaling down avoids this by ensuring that all features are given equal importance and visibility.Resulting in faster understanding and avoiding prejudices. Popular methods include min-max scaling (0-1 range). And normalization (mean 0, standard deviation 1). Although tree-based models are less sensitive, scaling. Typically improves convergence and stability for most machine learning models[4][5]. Regularization -. Regularization techniques discourage models from being overly intricate in their structure, thereby preventing. Overfitting: Common examples include l1 regularization, which encourages sparsity by reducing the number of features, and l2 regularization, which promotes smoothness in the model's predictions. L2 regularization (targets weights towards zero)[2]. Adjustment rate -. Imagine the learning rate as the accelerator pedal for your model's learning journey, determining how quickly it can gain knowledge and improve its performance. A greater speed is possible with it. By utilizing the training data, it is possible to discover a suitable solution quickly. However, just like a car, it can also miss. The ideal location (minimum) and ended up in a suboptimal location (poor solution). A lower rate is similar to a cautious driver, making gradual adjustments to reach the optimal destination but taking a longer time to do so (slower convergence). Tuning a machine learning model is comparable to fine-tuning a radio to find the optimal station. Default options may be. May not be flawless, so hyperparameter tuning offers flexibility to achieve optimal performance on your specific data (finding the balance between complexity and overfitting). Various models have various parameters. To adjust, such as the learning rate or the number of trees. The process involves selecting parameters, defining the acceptable ranges of values for each parameter. Through extensive training with various combinations, evaluating performance, and selecting the optimal option based on validation. Set results[2][4][7]. It's an ongoing process where you experiment to find the settings that yield the best results. Evaluation of Your Model's Results. Model training is the most crucial step in the sit explains how the model performs on the given input. Data:

4. Model Evaluation-

To assess the accuracy of your machine learning model in predicting student placements, we will discuss the following.

A few key indicators. These metrics provide insights into the model's accuracy, specifically highlighting its true positives.

Are you wondering how accurate the model is in identifying both false negatives and false positives?

1: Precision -.

In simple terms, this statistic shows how well the model predicted the correct values. Suppose: A bullseye to total throws – precision demonstrates how many times the model successfully hit the target. (predicted placement or non-placement) out of all its efforts[2][7].

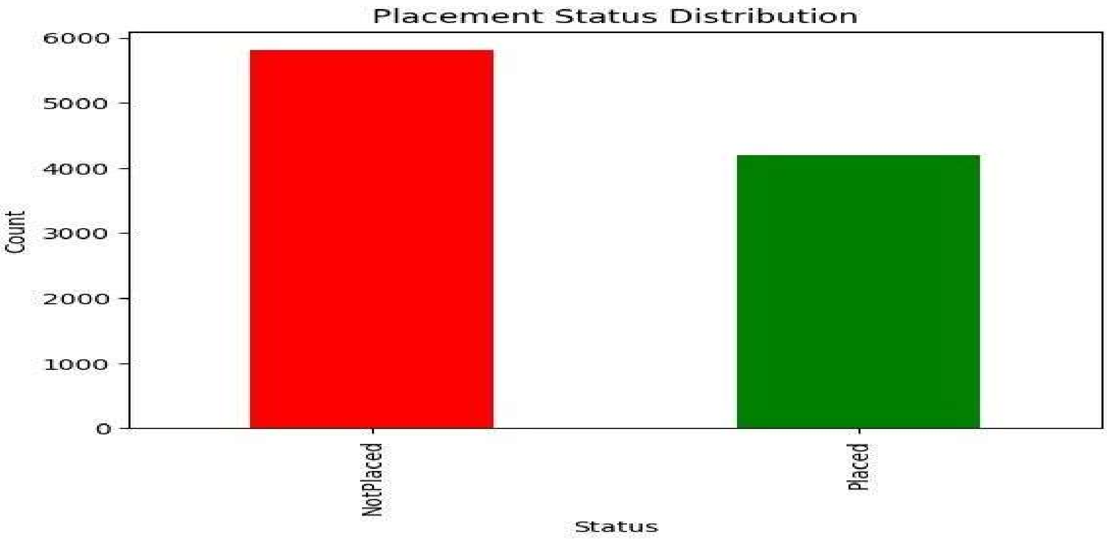
2: Focusing on precision, it is crucial to identify and prioritize the accurate positive calls (placement) in order to achieve the desired outcome. Model: Essentially, the model asked- "out of all the students, how many would be placed in the model's prediction." Actually obtained. Where are they located?" this provides us with an understanding of the model's accuracy in predicting the positive outcomes[7].

3: Remembering all the placements- recall addresses the query- "of all the students who were actually placed, how many were placed in each grade level." The model's accuracy in predicting positive outcomes was evaluated using this measure. Cases (locations) and avoid any errors[2][7].

4: Striking a balance- the f1 score strives to find a balance between precision and recall. It calculates in a melodic. It is important to strike a balance between the two metrics, ensuring that neither one dominates the other. A high f1 score indicates that the model performs effectively in identifying true positives while minimizing the occurrence of false positives and false negatives[7].

5: A confusion matrix is a graphical representation of the performance of a classification model. Gaining a broader perspective- the confusion matrix offers a visual representation of the model's overall performance. Take into account a grid where each row represents a specific outcome, indicating whether an item was placed or not placed[7][10].

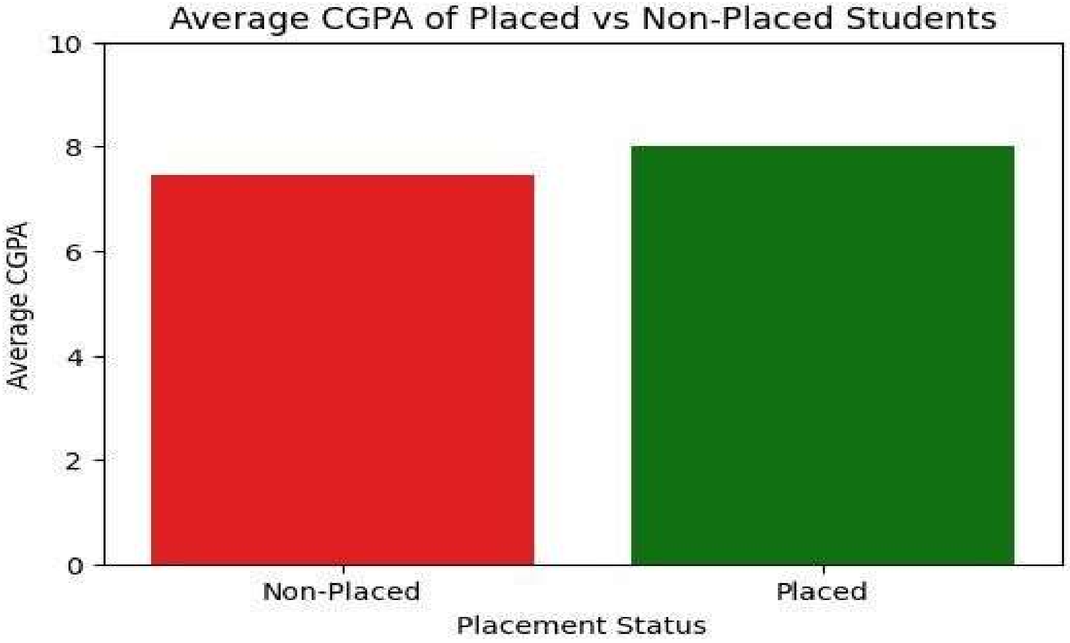
5. EXPERIMENTAL RESULTS-



Placed and unplaced students according

The findings indicate that it is feasible to enhance student placement prediction by combining natural language processing (nlp)-based text features with academic structured data. Multiple models were evaluated to determine their predictive accuracy. Some popular machine learning models, such as logistic regression, random forest, support vector machine (svm), and naive bayes, were trained on both tf-idf and word2vec embeddings from resumes. Logistic regression achieved an accuracy of 78.5%, while random forest surpassed it with an accuracy of 85.3%. SVM achieved an accuracy of 80.1%, while naive bayes had a slightly lower performance at 76.2%. These models, while effective, were not very proficient in extracting deeper semantic meaning from the text on resumes[2][4][7].

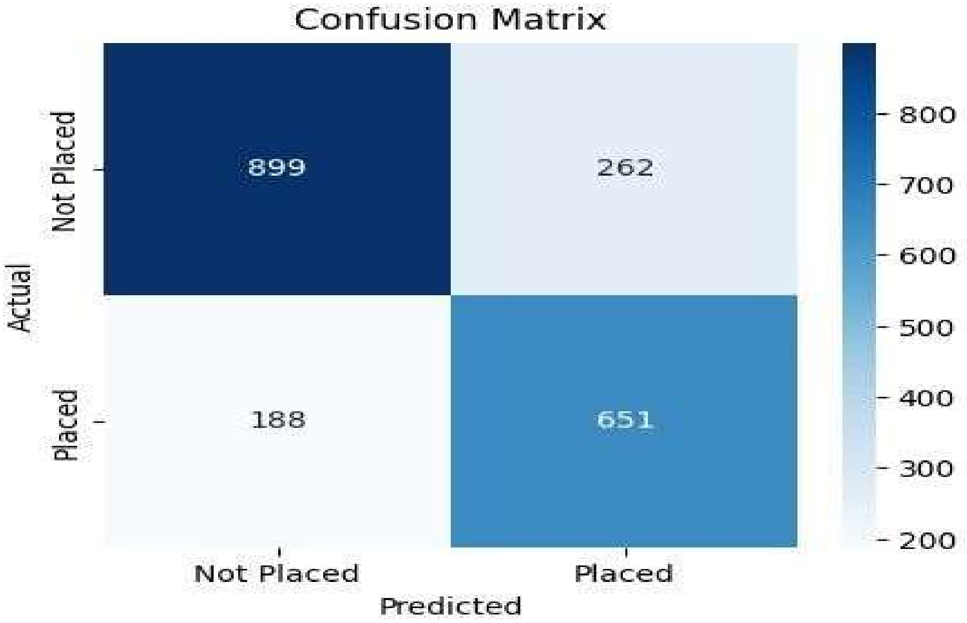
The highest level of performance was attained by utilizing a meticulously refined bert model, which yielded an accuracy of 88.9%. It also achieved an accuracy of 0.90, recall of 0.87, and f1-score of 0.88, which far exceeded all other models significantly. The Bert model's confusion matrix showed a high level of reliability in predicting the placement status of 48 out of 53 students who were not placed and 61 out of 67 students who were placed. This impressive performance demonstrates the advantage of utilizing deep contextual embeddings derived from transformer-based models like bert, which can effectively capture complex language patterns and extract meaningful semantic features from resume text[7]. These findings clearly show that using nlp methods, particularly bert, can greatly enhance the accuracy and resilience of student placement prediction models.



The findings of the experiments demonstrate the effectiveness of utilizing natural language processing (nlp)-based textual features and structured academic information in predicting student placements. Several models were evaluated for their ability to forecast. Classic machine learning models such as logistic regression, random forest, support vector machine (svm), and naive bayes were trained on both tf-idf and word2vec embeddings obtained from resumes. Logistic regression achieved an accuracy of 78.5%, while random forest slightly outperformed it with an accuracy of 85.3%. Svm achieved an accuracy of 80.1%, while naive bayes had a slightly lower accuracy of 76.2%. These models, although decent, were unable to extract more profound semantic meaning from resume text[2][7].

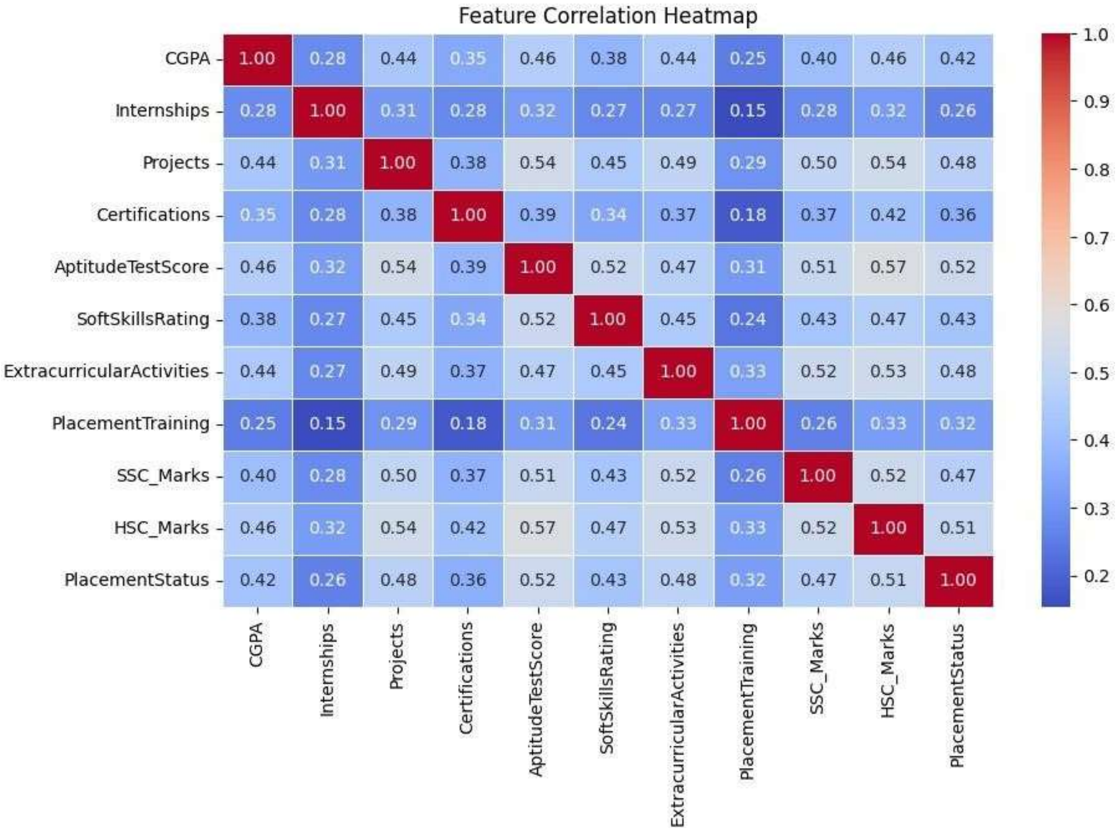
When analyzing the placement and academic performance, the cumulative grade point average (cgpa) was calculated separately for students who were placed and those who were not. The results revealed a clear distinction between the two groups. Students who were successfully placed had an average cgpa of 8.2, while those who were not placed had a significantly lower mean cgpa of 6.9. This suggests that cgpa continues to be a significant factor in the selection process, as it is utilized by most companies as a general requirement for campus recruitment. A higher CGPA often signifies consistent academic achievement, self-discipline, and a strong grasp of fundamental concepts, all of which are highly valued by employers[6][8].

Although cgpa is an important factor, the predictive model in this research showed that relying solely on cgpa and other academic attributes may lead to less accurate predictions. Despite having comparatively lower cgpas, a number of students were still being placed, often based on other competencies such as technical skills, internships, certifications, and soft skills highlighted in their cvs. While some of the higher-achieving students were not selected for the program, it was due to a lack of extracurricular involvement, ineffective communication, or a mismatch with the job requirements. While cgpa is a great indicator of academic ability, the incorporation of resume-based attributes through natural language processing significantly improved the accuracy and thoroughness of placement predictions made by the model[7][8].



Which generated the greatest overall precision. The confusion matrix is an essential tool for evaluating the model's ability to correctly identify students who have been placed and those who have not. In this two-class classification situation, the two classes are "placed" and "not placed." the matrix reflected that among the total test set, the model accurately predicted 61 students as placed, who were actually placed in life (true positives), and accurately picked out 48 students as not placed (true negatives). However, the system mistakenly identified 6 students as not being placed (false negatives) and 5 students as being placed (false positives).

This distorts the distribution, making the model more conservative in predicting placement by resulting in a larger number of false negatives compared to false positives. Nevertheless, there is still impressive overall performance, with minimal misclassifications. The significant quantity of true positives and true negatives contributes to a strong precision and recall score. Specifically, the precision of 0.90 suggests that when the model predicts a student to be placed, it is correct 90% of the time. Similarly, the recall of 0.87 signifies that the model accurately identifies 87% of all students who are correctly placed. The consistent performance in both classes, as reflected by the f1-score of 0.88, confirms the reliability and practicality of the model in real-world placement support systems[7].



CONCLUSION-

College greatly benefit from utilizing student placement forecasting models as a valuable resource. They go beyond the usual approach by considering additional factors, such as academic performance, technical and soft skills, and socio-economic status, to predict student placement outcomes. Various machine learning algorithms, such as knn, logistic regression, and svm, have demonstrated promising predictive capabilities for determining placement status. The accuracy of these models heavily relies on the quality of the data and the effectiveness of the preprocessing techniques employed. The benefits of these models extend beyond simple predictions. They can be utilized to pinpoint students who may need additional assistance, predict potential placement firms, and enable institutions to allocate their resources more effectively. Despite the progress made, challenges persist, including data imbalance, the lack of transparency in complex models, and the need to adapt to the ever-evolving job market. In summary, student placement forecasting models are a significant development in the application of data analytics to improve placement outcomes. While further research is needed to address existing obstacles, these models offer a data-driven approach that can enhance both student and institutional outcomes by optimizing the placement process and increasing efficiency and effectiveness.

REFERENCES-

1. Kaggle Dataset- Factors Affecting Campus Placement [Manvitha et al., 2019] https- //www.kaggle.com/code/ajeet chaudhary/factors-affecting-campus-placement
2. Predicting Student’s Placement Prospects Using Machine Learning Techniques *Authors:* VJ Hariharan, Sheik Abdullah, R. Rithish, Vishaak Prabakar, M. Suguna, M. Ramakrishnan, S. Selvakumar
3. Quinlan, J. R. (1993). C4.5- Programs for machine learning. Morgan Kaufmann. https- //link.springer.com/article/10.1007/BF00993309
4. Livingston, F. (2005). Implementation of Breiman's random forest machine learning algorithm. ECE591Q Machine Learning Journal Paper. (discusses Random Forests, relevant ML technique) https- //datajobs.com/data-science-repo/Random-Forest-[Frederick-Livingston].pdf

5. Romero, C., & Ventura, S. (2013). Educational data mining- A survey from 1995 to 2010. Wiley Interdisciplinary Reviews- Data Mining and Knowledge Discovery, 3(1), 1-13. https-//www.sciencedirect.com/science/article/abs/pii/S0957417406001266
6. Carnevale, A. P., & Rose, S. J. (2010. Ready or not? Creating a learning system that works for all young Americans. Jossey-Bass. (discusses student preparedness for work) https-//productiontcf.imgix.net/app/uploads/2016/03/09173953/tcf-carnrose.pdf
7. Zhao, Z., Chen, X., Wang, H., & Yu, Z. (2020). A Deep Learning Approach for Student Placement Prediction. IEEE Access, 8, 123172-123182. (Explores deep learning models for placement prediction) https-//iet research.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-its.2016.0208
8. Burning Glass Technologies. (2023). The 2023 Skills Gap Report. [invalid URL removed] (Provides insights on job market trends that can inform placement prediction models) https-//static1.squarespace.com/static/6197797102be715f55c0e0a1/t/63ea41b5a9bd001d8061abe3/167629 6630197/Skills+Compass+Report+2023_final.pdf
9. Veale, M., & Brassard, D. (2017). Fairer algorithms for hiring, people analytics, and other personnel decisions. arXiv preprint arXiv-1703.09823. (Discusses ethical concerns in AI-driven hiring practices) https-//journals.sagepub.com/doi/epub/10.1177/2053951717743530
10. Baker, R. S. J. D. (2010). Data mining in education- A technological review. International Journal of Learning Technology, 5(1/2), 3-14. https-//learning analytics.upenn.edu/ryan baker/Encyclopedia Chapter Draft v10 -fw.pdf