Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
April 2024 5(4), 2082–2094

FAIR AND TRANSPARENT STUDENT PLACEMENT PREDICTIONS: A MACHINE LEARNING AND EXPLAINABLE AI (XAI) APPROACH

Anjali Jindia [1], Sonal Chawla [1]

[1] Department of Computer Science and Applications, Panjab University, Chandigarh, India

ABSTRACT

Career path prediction plays a vital role in guiding students towards successful careers by offering personalized recommendations based on academic performance, skills, interests, and market trends. Early career path identification allows students to develop relevant skills and gain necessary experience, increasing their competitiveness in the job market. This study aims to ensure fair and transparent predictions by leveraging Machine Learning (ML) and Explainable AI (XAI) techniques on student career dataset. Initially, ML algorithms were applied to predict placement status, followed by an assessment of the model for bias using XAI. Upon detecting bias, mitigation strategies were implemented to enhance fairness. The use of XAI techniques improved model transparency and trustworthiness, allowing stakeholders to understand and trust the decision-making process. The methodology involved identifying and addressing dataset imbalances that could skew predictions. By applying oversampling techniques, the dataset was balanced, leading to significant improvements in model performance. The initial model showed poor performance metrics due to data imbalance, but after oversampling, the F1 score improved. Further application of XAI techniques, such as Local Interpretable Model-agnostic Explanations (LIME) and SHapely Additive exPlanation (SHAP), provided deeper insights into the model's decision-making process. This analysis highlighted specific features that, when oversampled, further enhanced the F1 score. The study emphasizes the importance of using XAI to not only improve model performance but also provide a trustworthy framework for stakeholders. By evaluating models using metrics like recall, precision, accuracy, and F1 score, the study demonstrated that integrating fairness and transparency into predictive models is achievable and beneficial for ensuring equitable student placement outcomes.

Keywords: Career Path Prediction, Explainable AI (XAI), Machine Learning, Bias Mitigation, LIME, SHAP

Corresponding Author
Anjali Jindia, ajindia82@gmail.com

DOI
10.29121/shodhkosh.v5.i4.2024.5996

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

1. INTRODUCTION

In today's rapidly evolving job market, guiding students toward appropriate career paths is increasingly essential. Career path prediction leverages data-driven techniques to provide personalized recommendations, aiding students in making informed decisions about their education and future careers. By considering factors such as academic performance, skills, interests, and market trends, predictive models can reduce the uncertainty and anxiety associated with career planning. This empowers students to focus their efforts on developing relevant skills and gaining valuable experiences, thereby enhancing their employability and success. Ensuring fairness and transparency in these predictions is critical to providing equitable opportunities for all students. This study utilizes advanced Machine Learning (ML) and Explainable AI (XAI) techniques to achieve accurate and fair career path predictions. By assessing and mitigating bias in the predictive models and employing XAI to enhance model transparency, this research aims to build a trustworthy

framework that stakeholders can rely on. Through this approach, the study highlights the potential for integrating fairness and transparency into predictive modeling to improve student placement outcomes equitably.

2. REVIEW OF RELATED LITERATURE

Researchers Mukesh Kumar et. al in [1] predicted College Students' Placements on the basis of their Academic Performance by using various Machine Learning Approaches like Logistic Regression, Naïve Bayes, Random Forest, k Nearest Neighbor (kNN) and Support Vector Machine (SVM). Apoorva Rao et. al in [2] proposed a "Student Placement Analyzer" by using Naïve Bayes algorithm, which predicts Placement status of a student in 5 categories viz dream company, core company, mass recruiter, not eligible and not interested in placements. This system is also helpful to weaker students as the institutions can provide extra care towards them to improve their performance. Irene Treesa et. al in [3] proposed a Placement Prediction System for B.Tech students by employing various Machine Learning classifiers like kNN, Logistic Regression, Random Forest and SVM. Research in the field of Bias Mitigation and Explainability in Machine Learning has also seen significant advancements in recent years. Authors Kamishima, Akaho and Asoh in [4] focused on developing fairness-aware learning approaches using regularization techniques to address bias in Machine Learning models. The paper focuses on incorporating fairness constraints into the learning process to promote fair decision-making. By introducing regularization for fairness, the authors aim to address bias issues in Machine Learning algorithms. Ribeiro, Singh and Guestrin in [5] introduced a model-agnostic framework for explaining classifier predictions, aiming to provide insights into model decisions. By providing explanations for model predictions, the authors aim to improve the interpretability of Machine Learning systems. These research efforts collectively contribute to the ongoing discourse on Bias Mitigation and Explainability in Machine Learning, paving the way for more accountable and interpretable AI systems. It also proposes methods for generating counterfactual explanations to enhance the Explainability of AI systems while maintaining privacy and confidentiality.

3. AIM AND OBJECTIVES

The primary aim of this research paper is to develop a fair and transparent predictive model for student placement using Machine Learning and Explainable AI (XAI) techniques. This study focuses on identifying key features that enhance the Machine Learning model's ability to predict placement status efficiently. It conducts a prediction task and compare two Explainable AI techniques for predicting placement status and identifying significant features. This study seeks to provide insights into the decision-making process of predictive models.

The objectives of this study are:

- To use Machine Learning for the prediction of placement status
- To use Explainable AI techniques on the student placement dataset
- To investigate bias using Explainable AI

4. MATERIALS AND METHODS

4.1. RESEARCH DESIGN

The study proposed the following steps given in figure 1 as the methodology for investigating bias. It began by cleaning the dataset, followed by data analysis and conversion into a numerical format to facilitate prediction. The processed dataset was then fed into various Machine Learning algorithms, including Decision Tree, Logistic Regression, Random Forest, Naive Bayes and k-Nearest Neighbors (kNN). Subsequently, the model was developed and testing was conducted on an unseen dataset to obtain accuracy readings. Finally, post hoc explainability techniques using SHAP and LIME were applied to interpret the model's decisions.

Figure 1

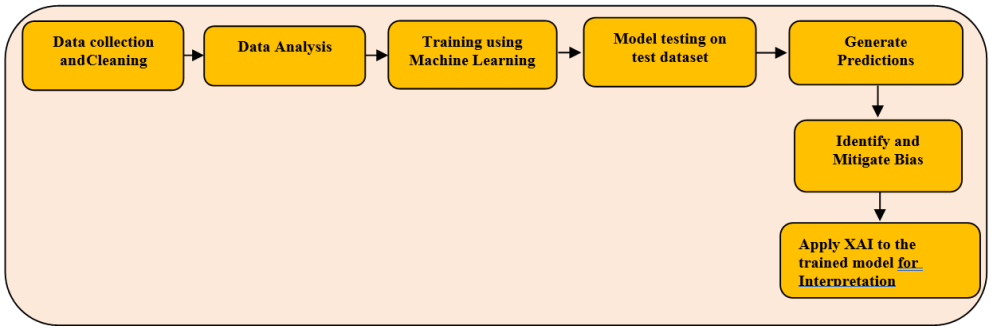

Figure 1 Research steps of the Proposed Method

4.2. DATA COLLECTION

The research data was obtained from Kaggle website and it comprised of 13 features (numerical, categorical and binary) and has 670 instances. The screenshot is shown below in figure 2:

Figure 2

gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	etest_p	mba_p	specialisat	workex	status
M	73.97697	Central	77.39269	Others	Science	73.62315	Sci&Tech	84.49831	67.20701	Mkt&HR	Yes	Placed
M	62	Others	62	Others	Commerc	60	Comm&M	63	52.38	Mkt&HR	Yes	Placed
M	80.96524	Others	67.10427	Others	Science	72.48262	Sci&Tech	91.34359	71.39231	Mkt&Fin	No	Placed
M	52	Others	65	Others	Arts	57	Others	75	59.81	Mkt&Fin	Yes	Not Placed
F	69	Central	62	Central	Science	66	Sci&Tech	75	67.99	Mkt&HR	No	Not Placed
M	78.76237	Others	78.66305	Others	Commerc	70.28107	Comm&M	67.44177	65.31805	Mkt&Fin	Yes	Placed
M	51.7945	Others	40.9856	Others	Science	62.12665	Others	65.86247	52.24975	Mkt&HR	No	Not Placed
M	59.80592	Central	61.45148	Others	Commerc	60	Comm&M	62.45148	57.02596	Mkt&HR	Yes	Placed
M	77.35132	Central	74.83467	Others	Science	73.48771	Sci&Tech	80.12399	68.00166	Mkt&HR	Yes	Placed
M	76.5	Others	97.7	Others	Science	78.86	Sci&Tech	97.4	74.01	Mkt&Fin	No	Placed
M	60.96136	Others	49.3827	Others	Science	54.50354	Others	70.3768	61.72612	Mkt&HR	No	Not Placed
F	73.80921	Central	81.21382	Central	Arts	58.65793	Others	59.23355	63.1802	Mkt&HR	Yes	Placed
M	80.91169	Others	60.9363	Others	Science	62.6819	Sci&Tech	89.78997	64.95734	Mkt&Fin	Yes	Placed
M	73.592	Central	78.39467	Others	Science	73.98667	Sci&Tech	87.368	66.7476	Mkt&HR	Yes	Placed
M	70.9782	Central	68.63551	Central	Science	72	Sci&Tech	57.38193	57.46757	Mkt&HR	No	Placed
M	60.08876	Others	44.52812	Others	Science	54.73006	Others	72.9946	58.51853	Mkt&HR	No	Not Placed
M	67	Others	61	Central	Science	72	Comm&M	72	61.01	Mkt&Fin	No	Placed
F	52	Others	52	Others	Science	55	Sci&Tech	67	59.32	Mkt&HR	No	Not Placed
F	48	Central	51	Central	Commerc	58	Comm&M	60	58.79	Mkt&HR	Yes	Not Placed
M	60.38914	Others	45.37504	Others	Science	54.22145	Others	72.53511	58.84182	Mkt&HR	No	Not Placed
M	79.39641	Others	61.39412	Others	Science	67.07126	Sci&Tech	92.93826	66.43687	Mkt&Fin	No	Placed
M	82	Others	63.56546	Others	Science	71.40667	Sci&Tech	94.98607	70.85456	Mkt&Fin	No	Placed

Figure 2 Screenshot of Dataset Used

4.3. DATA ANALYSIS

The data contained 670 entries; each field contained non-null values (all contained data) as shown in figure 3. 439 entries were positive for placement status outcomes (about 65.52%) and 231entries were negative for placement status outcomes (about 34.48%) as shown in figure 4. Moreover, out of 670 entries, 475 were from males and only 195 were from females as shown in figure 5. There were no duplicate records.

Figure 3

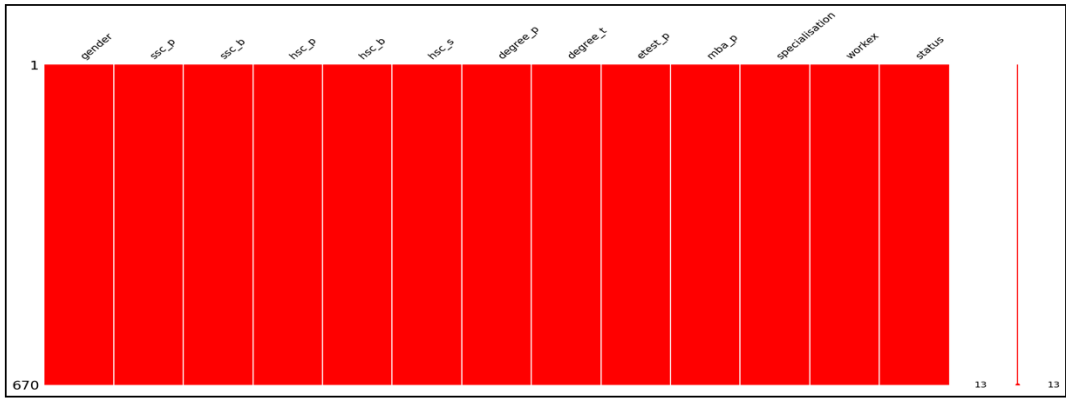

Figure 3 Graph Showing Absence of Missing Data

Figure 4

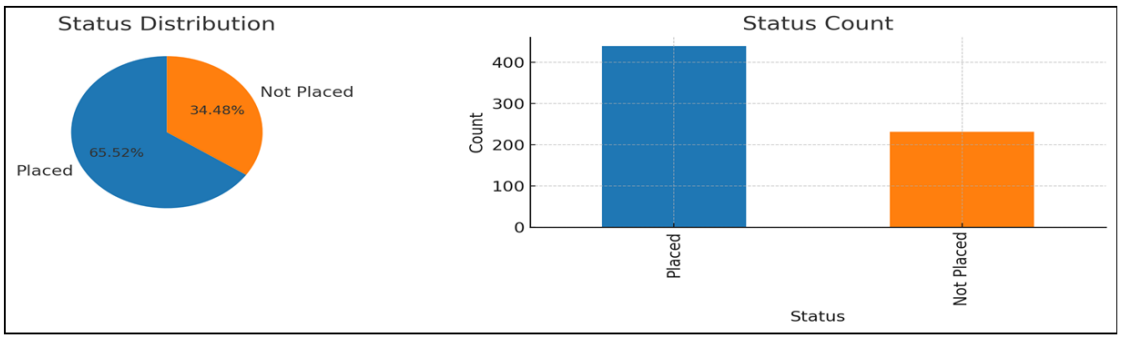

Figure 4 Graph Showing Placement Status Percentage

Figure 5

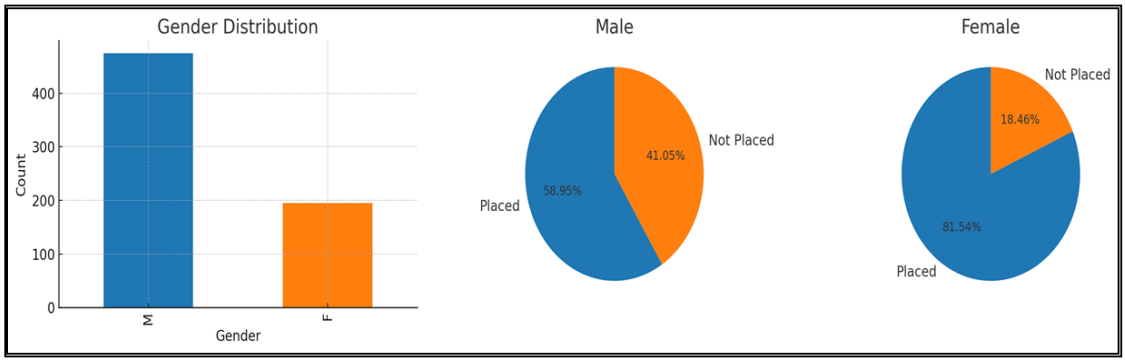

Figure 5 Graph Showing Gender Wise Placement Status Percentage

4.4. DATA PREPROCESSING

To make the dataset useful for Machine Learning algorithms, several data preprocessing steps are required to ensure that the data is clean, consistent, and in a format suitable for predictive modeling. Some of these steps include handling Missing Values, Encoding Categorical Variables i.e converting categorical variables into numerical format using techniques like one-hot encoding or label encoding, Normalizing/Scaling Numerical Features to a standard range, such as 0 to 1 or -1 to 1, using normalization or standardization to ensure that all features contribute equally to the model's learning process, data balancing in order to improve the model's performance & fairness and Feature Engineering. By

performing these preprocessing steps, the dataset gets transformed into a format suitable for training robust and accurate Machine Learning models. This ensures that the models can effectively learn from the data and make reliable predictions regarding student placement status.

4.5. DATA SPLITTING

In this study, the features required for predicting student placement status were identified as input features, and the actual placement outcomes were designated as labels. The dataset, excluding the placement status column, was stored in the features variable 'X', while the placement outcomes were stored in the label variable 'y'. An 80-20 split was then performed on the data, with 80% allocated for training the machine learning models and 20% for testing. This approach allowed the models to be trained on the majority of the data while reserving a separate, unseen test set for evaluating their performance.

4.6. MACHINE LEARNING MODELS USED ON THE STUDENT PLACEMENT DATASET

- **Logistic regression**

Logistic regression is particularly effective for binary classification problems. It predicts the probability by employing a logistic function to map values to probabilities, making it suitable for categorical response variables [6]

- **Naive Bayes**

Naive Bayes is based on Bayes' Theorem, assuming independence between predictors. It calculates the probability of each class and selects the class with the highest probability [7].

- **Decision Tree**

This algorithm splits the dataset into subsets based on the value of input features, forming a tree-like structure where each node represents a decision rule and each branch represents an outcome [8].

- **Random Forest**

Random Forest builds multiple decision trees and merges their results to improve accuracy and control overfitting. Each tree is trained on a random subset of the data and features, and the final prediction is made by majority voting for classification. This method enhances predictive performance and robustness [9].

- **kNN (k Nearest Neighbor)**

The k-Nearest Neighbors (kNN) algorithm works by identifying the 'k' closest training examples to a given input and predicting the output based on the majority class of these neighbors [10].

4.7. EVALUATION METRICS

To evaluate Machine Learning models, several common metrics are used: Accuracy, Recall, Precision, and F1 score. Accuracy measures the proportion of correctly predicted instances out of the total instances. Precision measures the proportion of correctly identified positive samples (placement status) out of all samples predicted as positive. Recall indicates the percentage of actual positive samples correctly identified. Precision and recall are crucial for assessing models, especially with imbalanced data. The F1 score, the harmonic mean of precision and recall, provides a balanced measure of the model's overall performance by combining these two metrics.

4.8. EXPLAINABLE AI TECHNIQUES

Explainable AI can be categorized into two types: post-hoc explainability and inherent explainability. Post-hoc explainability occurs after the model has been trained or a prediction made, typically used with complex models. Examples include LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapely Additive Explanation). Inherent explainability, on the other hand, is evident directly from the data without the need for additional models or libraries. This study implemented both SHAP and LIME techniques. The process of investigating bias using XAI is depicted in Figure 6 and table 1 shows the proposed methods to be used in Bias investigation using XAI.

Figure 6

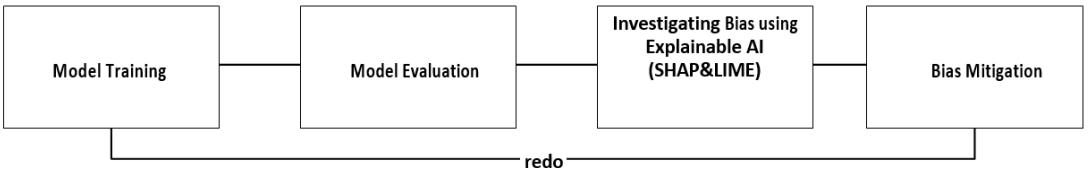

Figure 6 Process of investigation for bias using XAI

Table 1 Proposed Methods to be used in the investigation of Bias using Explainable AI (XAI)

Method	Dataset	XAI Technique
Model1 (M1)	Imbalance data	SHAP and LIME
Model2 (M2)	Balanced data	LIME
Model3 (M3)	Improved performance using explainable AI	LIME

M1 – The study will use M1 as the first model on the original dataset which is imbalanced and will use SHAP and LIME techniques to explain the data so as for bias to be observed.

M2 – The study will use M2 as the second model after balancing the original data and then use the LIME technique on the model. After using the Explainable AI technique to observe the features, the study will further investigate and mitigate bias to obtain a better model.

M3 – The study will use M3 as the third model to show how investigation using Explainable AI, further improved the model's performance.

5. RESULTS

The ML models were implemented using Python. To analyze the data, the study observed how the features correlate with the output (Placement Status) by obtaining the result as shown under in the figure 7.

Figure 7

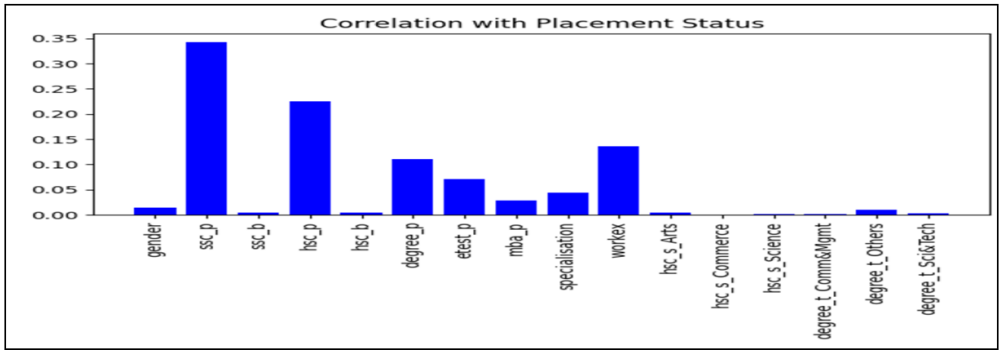

Figure 7 The feature correlation with our model

1) Model 1 (M1) Evaluation of the Imbalance Dataset

After training and testing the models, the following results were obtained as shown in figure 8. Logistic regression showed an accuracy of 92.86%, the random forest 93%, Decision Tree 86.43%, kNN 89.27% and the Gaussian naive Bayes 91.64%. Therefore, the Logistic Regression model was used in the Explainable AI analysis. The model performs well overall, with higher precision, recall, and F1-scores for class 1 compared to class 0. This indicates better performance in identifying positive instances, likely due to a higher number of positive samples in the dataset. Both SHAP and LIME techniques were used to investigate the imbalance in dataset.

Figure 8

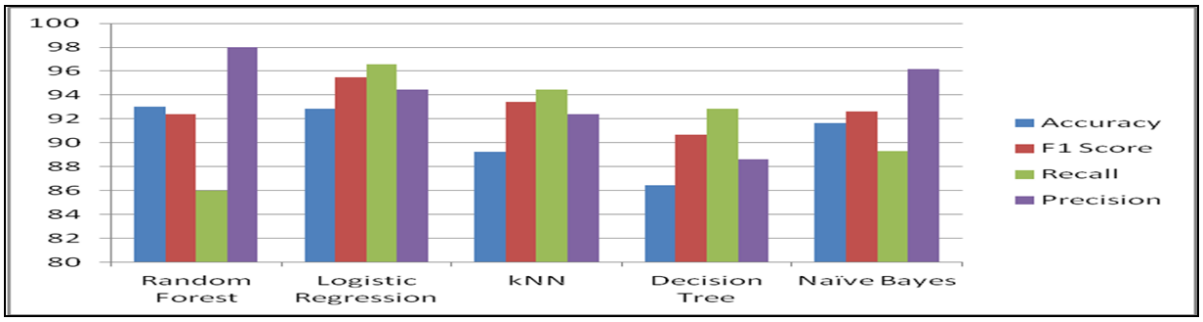

Figure 8 Evaluation Metrics

2) SHAP techniques on the imbalanced dataset

- **Feature importance**

The plot below in figure 9, displays the feature importance of the dataset calculated using SHAP values. SHAP values, derived from cooperative game theory, ensure a fair distribution of contributions among different factors working together to achieve a result. This method provides a clear understanding of the model by highlighting important features, which are those with high SHAP values.

Figure 9

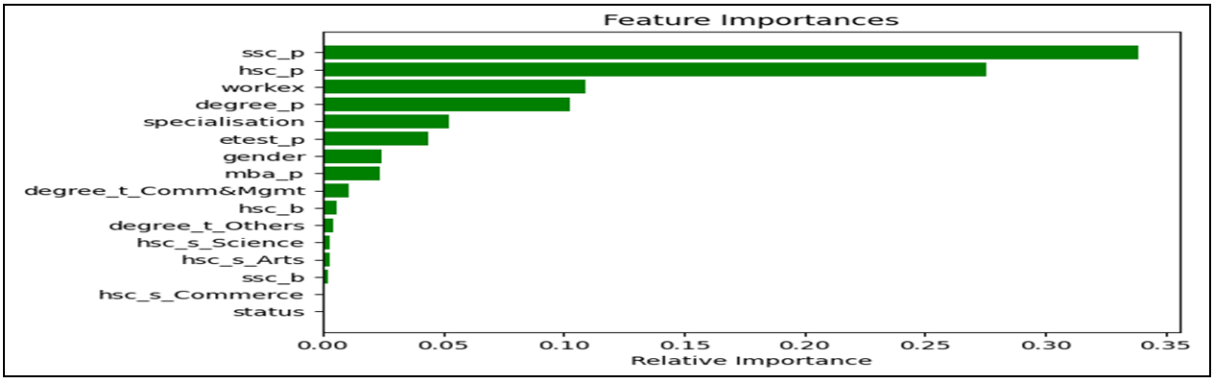

Figure 9 Feature importance based on SHAP values

- **Summary plot**

The SHAP summary plot in figure 10, visualizes the impact of various features on the model's output. Each dot represents a SHAP value for a feature in a specific instance, with colors indicating the feature value (blue for low, pink for high). The horizontal position shows whether the feature contributes positively or negatively to the prediction. The y-axis indicates the features, while the x-axis shows the SHAP value for each instance.

Figure 10

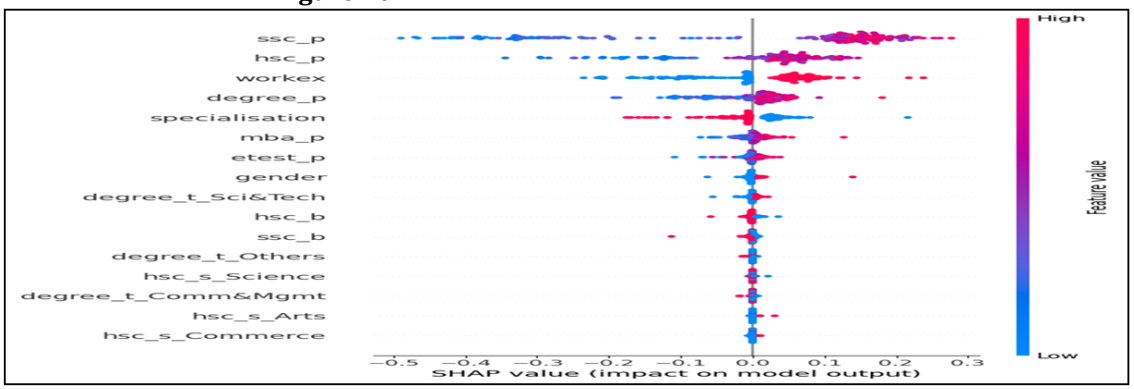

Figure 10 Feature importances with feature effects

- **The Waterfall Plot**

The SHAP waterfall plot in figure 11, illustrates the contribution of individual features to the model's prediction for a specific instance. The baseline value (E[f(x)]) is 0.75, and the plot shows how each feature affects this baseline to reach the final prediction. This waterfall plot breaks down the prediction by showing the cumulative effect of each feature starting from the baseline. Features like secondary education percentage and work experience have significant negative impacts on this particular prediction, while higher secondary percentage, gender, and degree percentage contribute positively, illustrating the detailed decision-making process of the model.

Figure 11

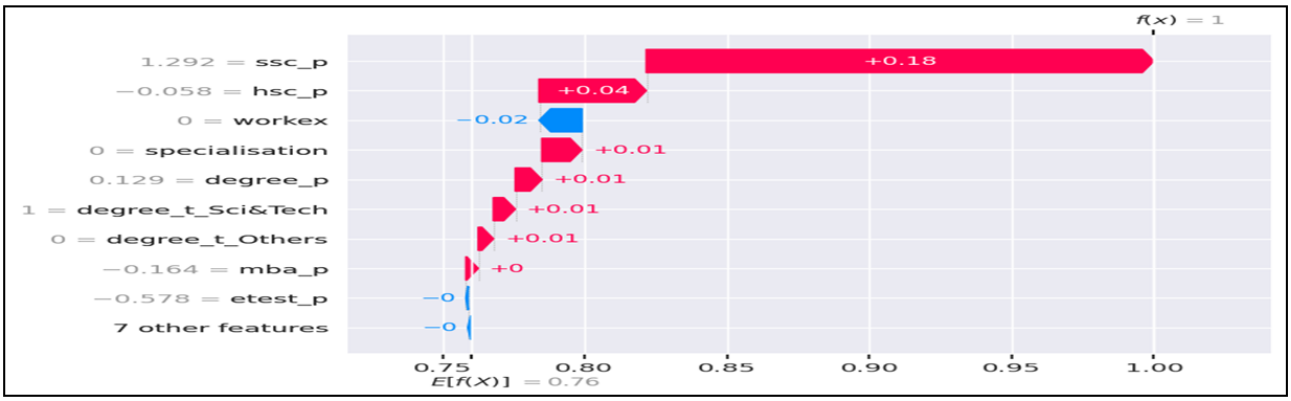

Figure 11 The Waterfall Plot

- **The Force Plot**

The force plot is a visualization tool used to explain the individual prediction of a machine learning model. It shows how different features influence a machine learning model's prediction, which is 1.00. In figure 12 below, starting from a base value of 0.68, the features ssc_p (+0.30), workex (+0.08), specialisation (+0.03), and degree_p (+0.02) contribute positively to the prediction. Conversely, hsc_p has a negative impact (-0.20). The combined effects of these features adjust the base value to reach the final prediction of 1.00.It shows how different features contribute to the final prediction.

Figure 12

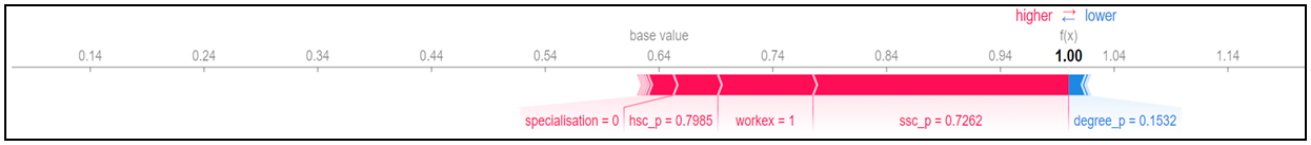

Figure 12 The Force Plot

3) LIME Technique for Explainable AI

- **Positive prediction**

The figure 13 below shows a correct prediction of student getting placed and shows the features that made the system come to that conclusion.

Figure 13

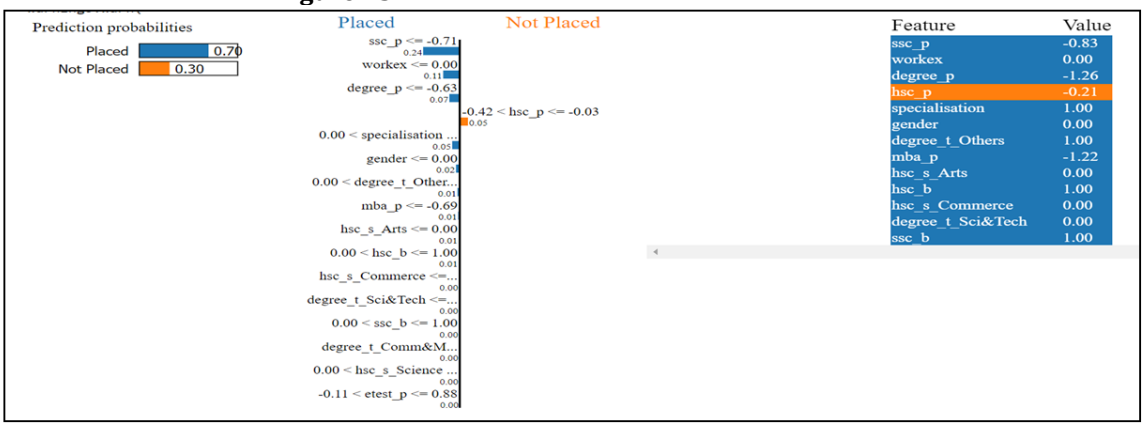

Figure 13 Positive prediction of LIME for M1

- **Negative prediction**

The figure 14 shows how an output that should be '1' (Placed) was explained as Not-Placed. This explanation shows the problem in the dataset or why it has that output which is one of the benefits of explainable AI.

Figure 14

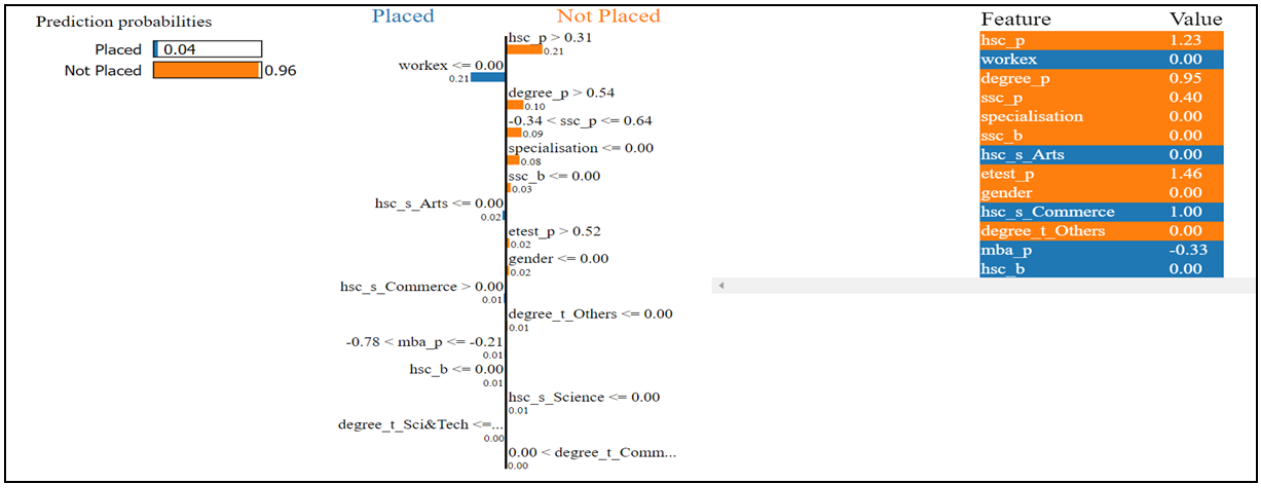

Figure 14 Negative Prediction of LIME for M1

This visualization can be used to identify bias in the dataset. The figure illustrates that 'hsc_p' was one of the probabilities the system relied on to reach its conclusion. This bias likely stems from an unbalanced dataset. A professional can use this visualization to detect and address the bias, potentially disregarding that particular model output.

4) Model2 (M2): Balancing the dataset and Model Evaluation

To solve the issue of the imbalance dataset, data was balanced by oversampling the negative ('0' for Not Placed) outcomes. This increased the dataset from 670 to 878. The negative values increased from 231 to 439. This gave a higher overall accuracy of 90%, precision score of 91%, recall scoreof 89.9% and higher f1 score of 90.37%. Now the model predicts the negative outcomes equally well. The study performed an XAI technique (LIME) on the new (balanced) data for analysis.

- **Positive outcome**

In the figure 15 below, the probability of having placed is 100% for that student's data. Based on this figure, it can be said that the specialization had an impact in making it '100', the patient has had a heart disease but it shows the other features had importance in this outcome.

Figure 15

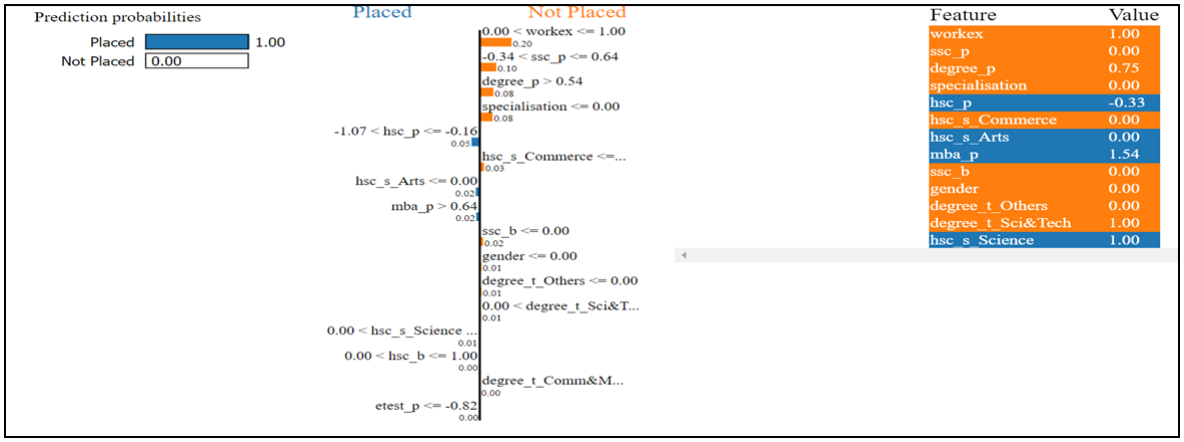

Figure 15 Positive prediction of LIME for M2

- **Negative outcome**

The figure 16 below explains how the system came to its prediction of not getting placed with a 96% probability, degree percentage greater than 55%. It explains that 'hsc_p', 'degree_p', and 'ssc_p' were also significant in making this decision.

Figure 16

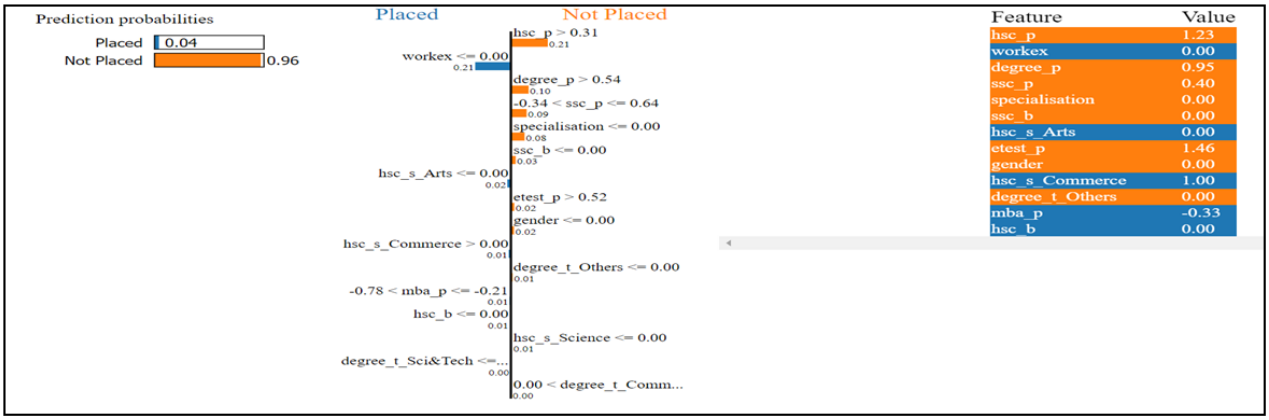

Figure 16 Negative prediction of LIME for M2

- **Observation after oversampling**

After observing the above data, we noticed that after performing oversampling of data, more accurate predictions of student not getting placed are shown.

5) Model3 (M3): Investigating Feature and Model Evaluation

Based on the previous task, we noticed the model used gender to lean towards the positive (Placed) decision when the student was female. To find out why, the study looked through the dataset and found that for gender, there were only 195 examples of females whereas 475 examples were of males. So, oversampling based on the gender was performed to have equal data for males and females. After this, new dataset size became 1268 as compared to 678 records in the original unbalanced and biased dataset. The new precision score was 88%, recall 88% and the f1 score also 88%.

- **Positive outcome investigation**

Due to oversampling the gender feature, we can see from figure 17 below that male/female is no longer a measure in leaning towards a positive output for placement status.

Figure 17

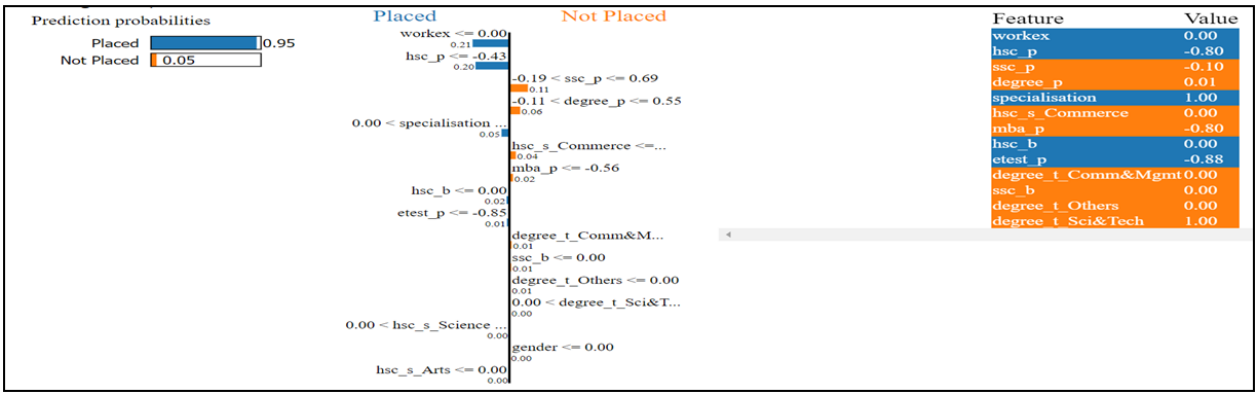

Figure 17 Positive outcome investigation of LIME for M3

- **Negative outcome investigation**

In the figure 18 below, we can see the improvement in the prediction

Figure 18

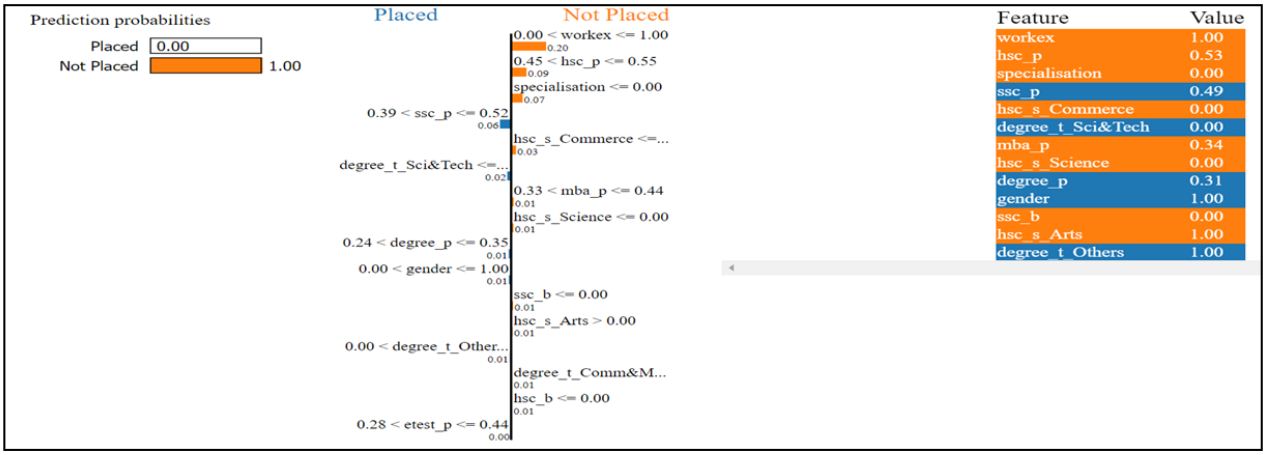

Figure 18 Negative outcome of LIME for M1

Table 2 Performance of the methods used for investigation of Bias

Method	Performance (f1 score)
M1	80%
M2	89.78%
M3	93.34%

M1 is the first model, this was used on the original dataset which was imbalanced. We used SHAP and LIME techniques to explain the data so the bias could be observed. The model classified all the data samples with an f1-score of 80%.

M2 is our second model, we oversample the original data and used Decision Tree, Logistic Regression, Naive Bayes, kNN and Random Forest and used the highest score (Random Forest) to perform the LIME technique on the model. After using the explainable AI technique, we found the model was using some features in the wrong way. We further investigated this and tried M3.

M3 is the third model that was used to show how our investigation using explainable AI, improved the model's performance. We were able to improve our score from 89.78% to 93.34% just by investigating the gender feature with explainable AI.

6. DISCUSSION

In the absence of a standard method to evaluate explainable AI, this research proposes a technique allowing career counselors to detect biases using explainable AI. This method highlights biases in AI predictions, such as gender-based disparities in dataset examples, which initially included disproportionately more males (634) than females (244). To address this, the study balanced the gender representation through oversampling (new size of dataset as 1268), enhancing the dataset's robustness and resulting in improved predictive accuracy (92.84) and fairness (precision score was 97.25 and recall was 98.51 and f1 score was 95.88). The outcomes suggest that while biases were reduced, gender imbalances still influenced placement predictions, indicating the need for further refinement of the data handling processes to minimize bias without adversely affecting other variables.

7. CONCLUSION

Our analysis revealed that the system, hindered by an unbalanced dataset, tended to predict "placed" status more accurately than "not-placed." By employing explainable AI, this study identified biases within the dataset, illustrating the technology's utility in revealing how specific features influence model outcomes. This insight allows developers to adjust features and address biases, as demonstrated by our use of oversampling to correct imbalances. Techniques like LIME further elucidated decision-making processes, showing, for instance, how certain feature values influence predictions, which aids practitioners in understanding and potentially following AI recommendations.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

"Predicting College Students' Placements Based on Academic Performance Using Machine Learning Approaches", Mukesh Kumar, Nidhi Walia, Sushil Bansal, Girish Kumar, Korhan Cengiz, International Journal of Modern Education and Computer Science (IJMECS), Vol.15, No.6, pp. 1-13, 2023.

"Student Placement Analyzer: A Recommendation System Using Machine", Apoorva Rao R, Deeksha K C, Vishal Prajwal R, Vrushak K, Nandini, JETIR, Volume 7, Issue 5, May 2020.

"Placement Prediction using Various Machine Learning Models and their Efficiency Comparison", Irene Treesa Jose, Daibin Raju, Jeebu Abraham Aniyankunju, Joel James, Mereen Thomas Vadakke International Journal of Innovative Science and Research Technology, Volume 5, Issue 5, May – 2020.

Kamishima, T., Akaho, S., & Asoh, H. Fairness-aware learning through regularization approach. In Proceedings of the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 643-654, 2012.

Ribeiro, M. T., Singh, S., & Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144, 2016.

D. G. Kleinbaum and M. Klein, Logistic Regression: A Self-Learning Text, 3rd ed. New York: Springer, 2010.

Wickramasinghe, I., Kalutarage, H. Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Comput 25, 2277–2293, 2021.

Quinlan, J.R. Induction of decision trees. Mach Learn 1, 81–106, 1986.

Couronné, R., Probst, P. & Boulesteix, AL. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 19, 270, 2018.

T. Cover and P. Hart, "Nearest neighbor pattern classification," in IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, Jan. 1967.