**Table of Contents**

[1.0 Introduction 6](#_Toc204554050)

[2.0 Domain Knowledge 7](#_Toc204554051)

[2.1 Problem Statement 7](#_Toc204554052)

[2.2 Aim, Objectives & Scopes 8](#_Toc204554053)

[2.2.1 Aim 8](#_Toc204554054)

[2.2.2 Objectives 8](#_Toc204554055)

[2.2.3 Scopes 8](#_Toc204554056)

[2.3 Data Dictionary 9](#_Toc204554057)

[2.4 Methodology Selection – SEMMA 11](#_Toc204554058)

[2.4.1 Phase 1: Sample 12](#_Toc204554059)

[2.4.2 Phase 2: Explore 12](#_Toc204554060)

[2.4.3 Phase 3: Modify 12](#_Toc204554061)

[2.4.4 Phase 4: Model 13](#_Toc204554062)

[2.4.5 Phase 5: Assess 13](#_Toc204554063)

[3.0 Exploratory Data Analysis 14](#_Toc204554064)

[3.1 Statistical Analysis 15](#_Toc204554065)

[3.2 Graphical Analysis 18](#_Toc204554066)

[3.3 Data Quality Issues 23](#_Toc204554067)

[4.0 Data Preprocessing 24](#_Toc204554068)

[4.1 Impute 24](#_Toc204554069)

[4.2 Drop 24](#_Toc204554070)

[4.3 Sample 25](#_Toc204554071)

[4.4 Data Partition 26](#_Toc204554072)

[4.5 Transform Variables Node 26](#_Toc204554073)

[4.6 Replacement Node 27](#_Toc204554074)

[5.0 Decision Tree, Gradient Boosting, Random Forest, HP Tree (Law Zhuo Wei) 28](#_Toc204554075)

[5.1 Model Construction and Optimization 28](#_Toc204554076)

[5.1.1 Model Construction and optimization for Decision Tree 28](#_Toc204554077)

[5.1.2 Model Construction and optimization for Gradient Boosting 32](#_Toc204554078)

[5.1.3 Model Construction and optimization for Random Forest (HP Forest) 36](#_Toc204554079)

[5.1.3 Model Construction and optimization for High Performance Decision Tree (HP Tree) 38](#_Toc204554080)

[5.2 Model Validation 42](#_Toc204554081)

[5.3 Critical Interpretation of Outcomes 43](#_Toc204554082)

[5.3.1 Fit Statistic of Ensemble Node (Gradient Boosting 3 Branch) 43](#_Toc204554083)

[5.3.2 Event Classification Table: Analysis and Calculations 46](#_Toc204554084)

[6.0 Regression Modelling (Eugene Tan Ting Siang) 49](#_Toc204554085)

[6.1 Model Construction and Optimization 49](#_Toc204554086)

[6.1.1 Regression without Variable Selection 50](#_Toc204554087)

[6.1.2 Regression with Variable Selection 56](#_Toc204554088)

[6.1.3 HP Regression 60](#_Toc204554089)

[6.2 Model Validation 64](#_Toc204554090)

[6.3 Critical Interpretation of Outcomes 66](#_Toc204554091)

[6.3.1 Fit Statistics 66](#_Toc204554092)

[6.3.2 Iteration Plot 67](#_Toc204554093)

[6.3.3 Summary of Backward Elimination 68](#_Toc204554094)

[6.3.4 Analysis of Maximum Likelihood Estimates 70](#_Toc204554095)

[6.3.5 Odds Ratio Estimates 72](#_Toc204554096)

[6.3.6 Event Classification Table 73](#_Toc204554097)

[7.0 Clustering (Beh Swee Shen) 76](#_Toc204554098)

[7.1 Model Construction and Optimization 76](#_Toc204554099)

[7.2 Model Validation 90](#_Toc204554100)

[7.3 Critical Interpretation of Outcomes 92](#_Toc204554101)

[8.0 Discussion and Conclusion 95](#_Toc204554102)

[8.1 Model Comparison 95](#_Toc204554103)

[8.1.1 ROC Curve 96](#_Toc204554104)

[8.1.2 Cumulative Lift 97](#_Toc204554105)

[8.1.3 Fit Statistics 97](#_Toc204554106)

[8.2 Conclusion 99](#_Toc204554107)

[9.0 References 101](#_Toc204554108)

**List of Figures**

[Figure 1: SEMMA Methodology Diagram 12](#_Toc199768490)

[Figure 2: Editing Variables 15](#_Toc199768491)

[Figure 3: Summary statistics of class variables 16](#_Toc199768492)

[Figure 4: Summary statistics of interval variables 17](#_Toc199768493)

[Figure 5: Variable Worth chart 18](#_Toc199768494)

[Figure 6: Histogram of Aptitude Test Scores 19](#_Toc199768495)

[Figure 7: Histogram of CGPA 19](#_Toc199768496)

[Figure 8: Bar Chart of Extracurricular Activities 19](#_Toc199768497)

[Figure 9: Histogram of HSC Marks 20](#_Toc199768498)

[Figure 10: Pie Chart of Internship 20](#_Toc199768499)

[Figure 11: Bar Chart of Placement Status 20](#_Toc199768500)

[Figure 12: Bar Chart of Placement Training 21](#_Toc199768501)

[Figure 13: Pie Chart of Projects 21](#_Toc199768502)

[Figure 14: Boxplot of SSC Marks 21](#_Toc199768503)

[Figure 15: Histogram of Soft Skills Rating 22](#_Toc199768504)

[Figure 16: Pie Chart of Workshops / Certifications 22](#_Toc199768505)

[Figure 17: Boxplot of CGPA vs Placement Status 22](#_Toc199768506)

[Figure 18: Bar Chart of Placement Status for each Placement Training 23](#_Toc199768507)

[Figure 19: 3D Bar Chart of Internships vs Extracurricular Activities vs CGPA 23](#_Toc199768508)

[Figure 20: Data quality issue: imbalance classes 24](#_Toc199768509)

[Figure 21: Impute node 25](#_Toc199768510)

[Figure 22: Drop node 25](#_Toc199768511)

[Figure 23 Sample Node 26](#_Toc199768512)

[Figure 24: Sample properties 26](#_Toc199768513)

[Figure 25 Data Partition Node 27](#_Toc199768514)

[Figure 26 Data Set Allocations 27](#_Toc199768515)

[Figure 27 Transform Variables Node 27](#_Toc199768516)

[Figure 28 Replacement Node 28](#_Toc199768517)

[Figure 29 Decision Tree Node 29](#_Toc199768518)

[Figure 30 Decision Tree Properties 29](#_Toc199768519)

[Figure 31 Decision Tree Diagram 30](#_Toc199768520)

[Figure 32 Model Comparison Table for decision tree 32](#_Toc199768521)

[Figure 33 Subtree Assessment Plot for chosen Decision Tree 32](#_Toc199768522)

[Figure 34 Gradient Boosting Node 33](#_Toc199768523)

[Figure 35 Gradient Boosting Properties 34](#_Toc199768524)

[Figure 36 Gradient Boosting Diagram 34](#_Toc199768525)

[Figure 37 Model Comparison table for gradient boosting 36](#_Toc199768526)

[Figure 38 HP Forest node 37](#_Toc199768527)

[Figure 39 HP Forest Properties 37](#_Toc199768528)

[Figure 40 HP Forest diagram 38](#_Toc199768529)

[Figure 41 Model Comparison for HP Forest 39](#_Toc199768530)

[Figure 42 HP Tree Node 39](#_Toc199768531)

[Figure 43 HP Tree Properties 40](#_Toc199768532)

[Figure 44 HP Tree diagram 41](#_Toc199768533)

[Figure 45 Model Comparison for HP Tree 42](#_Toc199768534)

[Figure 46 Best Model Diagram 43](#_Toc199768535)

[Figure 47 Model Comparison for Finding the Best Model 43](#_Toc199768536)

[Figure 48 Ensemble Node 44](#_Toc199768537)

[Figure 49 Fit Statistic of ensemble (Gradient Boosting 3 branch) 45](#_Toc199768538)

[Figure 50 Fit Statistic 2 46](#_Toc199768539)

[Figure 51 Event Classification Table of Best Model (Gradient Boosting) 47](#_Toc199768540)

[Figure 52: Regression node 50](#_Toc199768541)

[Figure 53: Regression equation properties 50](#_Toc199768542)

[Figure 54: Regression class targets properties 51](#_Toc199768543)

[Figure 55: Regression model selection properties 51](#_Toc199768544)

[Figure 56: Regression diagram (1) 52](#_Toc199768545)

[Figure 57: Regression comparison table (1) 56](#_Toc199768546)

[Figure 58: Variable Selection Node 57](#_Toc199768547)

[Figure 59: Regression diagram (2) 58](#_Toc199768548)

[Figure 60: Variables rejected when default criteria were used 58](#_Toc199768549)

[Figure 61: HP Regression node 61](#_Toc199768550)

[Figure 62: HP Regression diagram 61](#_Toc199768551)

[Figure 63: HP Regression comparison table 64](#_Toc199768552)

[Figure 64: Regression comparison diagram 65](#_Toc199768553)

[Figure 65: Regression comparison table (2) 66](#_Toc199768554)

[Figure 66: Fit Statistics of 2nd Degree Polynomial Backward Regression 67](#_Toc199768555)

[Figure 67: Iteration Plot of 2nd Degree Polynomial Backward Regression 68](#_Toc199768556)

[Figure 68: Summary of backward elimination 69](#_Toc199768557)

[Figure 69: Analysis of maximum likelihood estimates 71](#_Toc199768558)

[Figure 70: Odds ratio estimates 73](#_Toc199768559)

[Figure 71: Event classification table 74](#_Toc199768560)

[Figure 72: Cluster Node 77](#_Toc199768561)

[Figure 73: Cluster’s properties 77](#_Toc199768562)

[Figure 74: Model 78](#_Toc199768563)

[Figure 75: Data Partition Node 78](#_Toc199768564)

[Figure 76: First Cluster segment size 78](#_Toc199768565)

[Figure 77: First Cluster segment plot 79](#_Toc199768566)

[Figure 78: First Cluster mean statistics 80](#_Toc199768567)

[Figure 79: First Cluster variable importance 81](#_Toc199768568)

[Figure 80: 5 Cluster 82](#_Toc199768569)

[Figure 81: 5 Cluster segment plot 82](#_Toc199768570)

[Figure 82: 5 Cluster mean statistics 84](#_Toc199768571)

[Figure 83: 7 Cluster 85](#_Toc199768572)

[Figure 84: 7 Cluster segment plot 86](#_Toc199768573)

[Figure 85: 7 Cluster mean statistics 87](#_Toc199768574)

[Figure 86: 3 Cluster 88](#_Toc199768575)

[Figure 87: 3 Cluster segment plot 88](#_Toc199768576)

[Figure 88: 3 Cluster mean statistics 90](#_Toc199768577)

[Figure 89: Decision Tree 90](#_Toc199768578)

[Figure 90: Model Comparision table 91](#_Toc199768579)

[Figure 91: Cumulative lift 92](#_Toc199768580)

[Figure 92: Segment Profile 93](#_Toc199768581)

[Figure 93: Segment profile variable worth 95](#_Toc199768582)

[Figure 94: Model comparison with Ensemble 96](#_Toc199768583)

[Figure 95: ROC Curve 97](#_Toc199768584)

[Figure 96: Cumulative Lift 98](#_Toc199768585)

[Figure 97: Fit Statistics 98](#_Toc199768586)

# **1.0 Introduction**

Predictive modeling has emerged as a crucial instrument in education in recent years, assisting institutions in making data-driven choices to improve student outcomes (Ruparel & Swaminarayan, 2024). As the amount of student data increases, machine learning methods have become increasingly popular for providing insightful information on a range of academic and professional factors, including placement success. Predictive models can also identify patterns and trends in previous data that are not immediately apparent, enabling better resource allocation and advice (Haque et al., 2024).

This project focuses on developing multiple prediction models using a student placement dataset to evaluate and compare their effectiveness. By using a variety of machine learning algorithms, we can retrieve a more comprehensive understanding of the variables affecting placement results and the potential predictive capability of various strategies. The goal is to create reliable models that will help schools identify students who are at risk, improve training, and eventually increase placement rates.

# **2.0 Domain Knowledge**

## **2.1 Problem Statement**

Even with the abundance of historical student data available, many educational institutions find it difficult to identify students who might need extra help finding work after graduation (Swami et al., 2025). Traditional approaches usually rely on generalized metrics or limited criteria, which may overlook students who are silently struggling or at risk of being left behind in the placement process (Ruparel & Swaminarayan, 2025). Since student datasets nowadays are growing in size and complexity, there is a need for more advanced methods to analyze historical trends and predict future outcomes accurately.

Using historical data, machine learning models can help identify students who are less likely to be placed based on past patterns of academic performance, training attendance, and other relevant factors. This enables institutions to pinpoint individuals who may require additional resources or interventions. The ability to predict placement outcomes not only supports strategic planning but also provides a foundation for personalized assistance, ultimately helping more students secure employment after graduation.

## **2.2 Aim, Objectives & Scopes**

### **2.2.1 Aim**

The aim of this research is to build and compare multiple prediction models for forecasting student placement outcomes based on student’s performance in school by exploring and pre-processing a student placement dataset to identify the best and most accurate model for prediction. Clustering will also be used to identify underlying patterns and structures within the dataset.

### **2.2.2 Objectives**

1. To explore a dataset on student’s placement data to determine if they get placed in a job after graduation based on student’s performance in school using data mining techniques.
2. To pre-process the student’s placement dataset to ensure a clean dataset that will yield the best results when creating a predictive model, by using pre-processing techniques in SAS Enterprise Miner.
3. To predict student placement rates and patterns by building various prediction models and comparing the performance of each model to find the best performing one.
4. To group data points that contain similarities using clustering and reveal underlying patterns and structures.

### **2.2.3 Scopes**

|  |  |
| --- | --- |
| Type of dataset | Classification Dataset |
| Geography | India |
| Time | Q1 of 2025 |
| License | CC0: Public Domain |
| Link to Dataset | <https://www.kaggle.com/datasets/ruchikakumbhar/placement-prediction-dataset> |

## **2.3 Data Dictionary**

|  |  |
| --- | --- |
| **Dataset Title** | Placement Prediction Dataset |
| **Dataset Description** | This dataset contains information about the students’ academic and training experiences along with their respective placement status. The data in this dataset can allow researchers to understand which factors affect student’s placement outcomes the most and build optimized prediction models based on those findings. |
| **Method of Data Collection** | Manual data collection |
| **Sources** | Education Institutions in India |
| **Date Published** | January 2025 |
| **Number of Columns** | 12 |
| **Size** | 10,000 entries |

|  |  |  |
| --- | --- | --- |
| Attribute | Data Type | Description |
| StudentID | Integer | Unique identifier for each student record. |
| CGPA | Double | Student’s cumulative grade point average (CGPA) on a 10-point scale (lowest is 0, highest is 10.0) |
| Internships | Integer | Number of internships done by a student. |
| Projects | Integer | Number of projects done by a student. |
| Workshops/Certifications | Integer | Number of certification courses that a student has participated in. |
| AptitudeTestScore | Integer | Student’s score in an aptitude test (a test that measures the student’s skill and potential conducted during the recruitment process). |
| SoftSkillsRating | Double | A rating of the student’s proficiency in soft skills (e.g. communication skills). |
| ExtracurricularActivities | Binary | Whether the student participated in extracurricular activities or not. |
| PlacementTraining | Binary | Whether the student has gone through placement training provided in college. |
| SSC\_Marks | Integer | Student’s senior secondary marks. |
| HSC\_Marks | Integer | Student’s higher secondary marks. |
| PlacementStatus | String | Whether the student was placed in a job placement after graduation. This is the target variable. |

## **2.4 Methodology Selection – SEMMA**

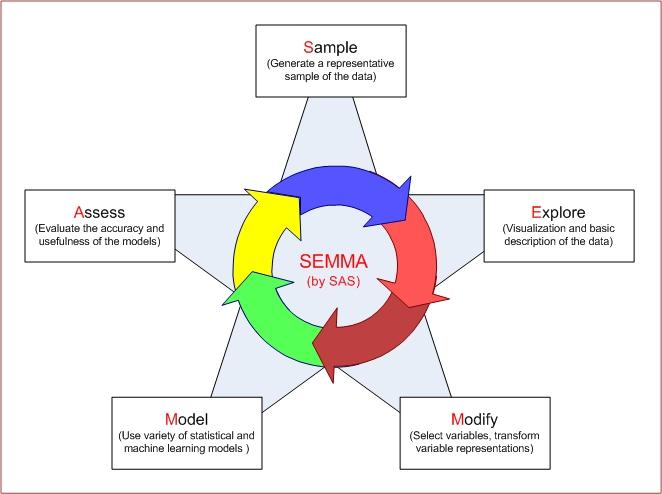


Figure 1: SEMMA Methodology Diagram

For this project, the SEMMA methodology – Sample, Explore, Model, Modify, Assess – was chosen to conduct the analysis. This methodology was developed by the SAS institute and is closely associated with SAS Enterprise Miner, the software we are using for this project. SEMMA is also more focused on the technical aspects of data mining, unlike CRISP-DM which relates to the whole project lifecycle (business understanding to deployment). Furthermore, SEMMA puts more emphasis on iteratively exploring the data to ensure that we retrieve data of the best quality that is free from errors such as noise, missing values, and irrelevant attributes, thus making it more suited for modelling (Gautam et al., 2017). These are the reasons why we went with this methodology instead of others.

### **2.4.1 Phase 1: Sample**

The SEMMA methodology starts with the Sample phase, where a representative subset of the data is obtained. Some key tasks involved in this phase include extracting large datasets from sources and partitioning data into training and validation sets to ensure unbiased model evaluation. The primary goal of this stage is to identify independent and dependent variables that influence the target outcome. For example, in our case, the data set for student placement should include factors such as historical academic records, training attendance, and placement status. By sampling effectively, the models are trained on data that mirrors real-world scenarios while ensuring that the evaluation remains objective and generalizable.

### **2.4.2 Phase 2: Explore**

Then comes the Explore phase which involves analyzing the sampled data using techniques such as statistical analysis and graphical analysis. This allows us to identify any data quality issues that arise like missing values, noisy data, inconsistencies, outliers, and more. It also allows us to better understand the structure of the data and identify any patterns or relationships between variables through univariate or multivariate analysis. For this case study, exploration includes tasks like examining the distribution of test scores in a histogram and finding correlations between a student’s CGPA and their placement status.

### **2.4.3 Phase 3: Modify**

The Modify phase focuses on data transformation and preparing the data for modeling. In this phase, missing values identified from the previous phase can be addressed by imputing it with the attribute’s mean value, mode value, or even values determined from a decision tree which considers its relationships with other attributes. Distributions that are skewed can be transformed as well through various methods like logarithmic transformation or normalization to ensure that the outliers do not greatly affect the prediction results. Besides that, if the target variable is imbalanced (e.g. more placement status = yes than placement status = no), we can apply sampling to under sample the majority class while keeping all of the data in the minority class to help the model learn more effectively and reduce bias towards the dominant outcome.

### **2.4.4 Phase 4: Model**

During the Model phase, we will build various predictive models and apply their algorithms to the modified dataset. In this study, models such as decision trees, regression models, and neural networks are trained to predict whether the student will be placed in a job or not after graduation. We can also tune each model’s hyperparameters to improve their performance, such as the learning rate, tree depth, polynomial degree, and more depending on the algorithm used. In addition to that, we can experiment with different feature combinations, data partition ratios, and other factors since they also affect the performance of the model.

### **2.4.5 Phase 5: Assess**

Finally, the Assess phase evaluates the models’ predictive performance using metrics such as accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (ROC-AUC). This evaluation helps determine which model best identifies students at risk of not being placed. For the case study, this step is very important as it informs decision-makers about which model can reliably support early interventions, helping institutions allocate resources to those who need additional training or career guidance.

It is important to note that the SEMMA methodology supports iterative movement in any of the phases, ensuring flexible adjustments and delivering a well-refined model. This iterative nature allows us to attain the best and most accurate results possible which align best with the objectives of the analysis. For example, if the Modeling phase produces poor performing models, we can freely go back to the Explore phase to investigate any data distributions or outliers that were previously missed.

# **3.0 Exploratory Data Analysis**

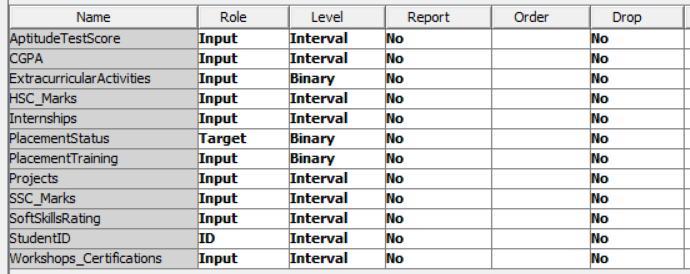


Figure 2: Editing Variables

After importing the dataset to SAS Enterprise Miner, the variables in the dataset were explored. First, StudentID’s role was changed from “Input” to “ID”. In addition, the role of PlacementStatus was also changed from “Input” to “Target” as it is the target variable. The levels for the following columns (ExtracurricularActivities, PlacementStatus, PlacementTraining) were changed to binary as well since they only contain two values.

## **3.1 Statistical Analysis**

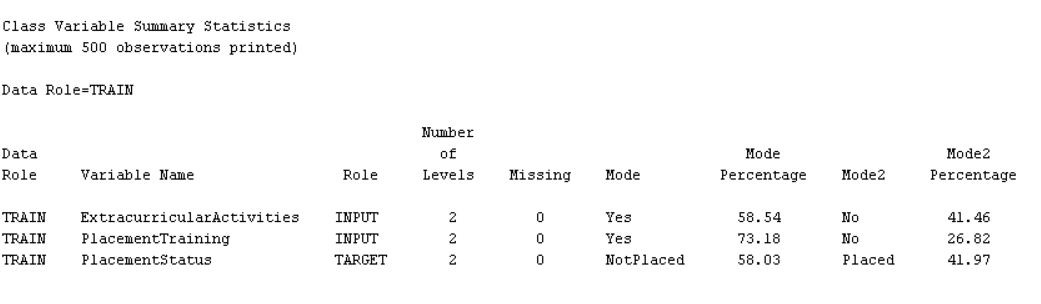


Figure 3: Summary statistics of class variables

|  |  |
| --- | --- |
| Class Variable | Findings |
| ExtracurricularActivities | First, we will focus on the class variables. This variable contains 2 values: Yes (58.54%) and No (41.46%). The two categories are quite evenly split, with Yes slightly greater than No. There are no missing values in this variable. |
| PlacementTraining | This variable also contains 2 values: Yes (73.18%) and No (26.82%). As we can see, the Yes category is the majority by a huge margin. There are also no missing values in this variable. |
| PlacementStatus | This is the target variable for our predictive models. It contains 2 values: NotPlaced (58.03%) and Placed (41.97%).    These statistics show that 5803 records were classified as NotPlaced while 4197 records were classified as Placed. There is a slight class imbalance observed in this variable, so sampling will have to be conducted. |

A number of numbers on a white background

AI-generated content may be incorrect.

Figure 4: Summary statistics of interval variables

|  |  |
| --- | --- |
| Class Variable | Findings |
| AptitudeTestScore | First, we will focus on the interval variables. This variable contains continuous values ranging from 60 to 90. The distribution has a mean of 79.44, with a standard deviation of 8.16. The skewness is -0.36, indicating a slight left-skew in the data. There are no missing values in this variable. |
| CGPA | This variable contains continuous values ranging from 6.5 to 9.1. The mean CGPA is 7.69, with a standard deviation of 1.64. The skewness is -0.42, indicating a slight left-skew. There are no missing values in this variable. |
| HSC\_Marks | This variable contains continuous values ranging from 57 to 88. The mean HSC marks are 74.50, with a standard deviation of 8.19. The skewness is -0.00099, indicating a nearly normal distribution. There are no missing values in this variable. |
| Internships | This variable contains continuous values ranging from 0 to 5. The mean number of internships is 1.05, with a standard deviation of 0.87. The skewness is 0.10, indicating a nearly symmetrical distribution. There are no missing values in this variable. |
| Projects | This variable contains continuous values ranging from 0 to 2. The mean number of projects is 2.03, with a standard deviation of 0.87. The skewness is -1.60, indicating a strong left-skew in the data. There are no missing values in this variable. |
| SSC\_Marks | This variable contains continuous values ranging from 35 to 90. The mean SSC marks are 69.15, with a standard deviation of 10.43. The skewness is 0.03, indicating a near-normal distribution. There are no missing values in this variable. |
| SoftSkillsRating | This variable contains continuous values ranging from 3 to 4.8. The mean soft skills rating is 4.32, with a standard deviation of 0.41. The skewness is -0.67, indicating a slight left-skew. There are no missing values in this variable. |
| StudentID | This variable contains unique identifiers for each student. It is an integer type variable ranging from 1 to 10,000. As an ID variable, it is used to uniquely identify each student record and does not contribute to the analysis for prediction purposes. There are no missing values in this variable. |
| Workshops\_Certifications | This variable contains continuous values ranging from 0 to 3. The mean number of workshops or certifications is 1.01, with a standard deviation of 0.90. The skewness is 0.20, indicating a slightly positive skew. There are no missing values in this variable. |

A screenshot of a graph

AI-generated content may be incorrect.

Figure 5: Variable Worth chart

The Variable Worth chart highlights the importance of each predictor variable in predicting PlacementStatus. The chart shows that AptitudeTestScore, HSC\_Marks, and Projects are the most significant variables, with higher weights, indicating a strong correlation with the likelihood of placement. These variables are crucial in determining whether a student gets placed or not. On the other hand, StudentID and Internships have lower weights, suggesting that they are less important predictors in this context. The chart reinforces that focusing on the more influential variables, such as test scores and academic performance, will improve the accuracy of placement predictions.

## **3.2 Graphical Analysis**

|  |  |
| --- | --- |
| Graphical View of Variables | Interpretation |
| Student ID | This is a unique identifier for each student record. |
| Figure 6: Histogram of Aptitude Test Scores | The histogram shows the distribution for Aptitude Test Scores, which ranges from 60 to 90. The results are divided into 10 bins. From the histogram, it can be observed that a huge majority of students (2618) scored between 87 to 90 in their aptitude tests. There are no missing values in this column. |
| Figure 7: Histogram of CGPA | The histogram of CGPA ranges from 6.5 to 9.1. The results are divided into 10 bins. The bin between 8.06 and 8.32 has the highest frequency, suggesting this is the most common CGPA range among students. A small number of students also fall above 9.0, which could be outliers. There are no missing values in this column. |
| Figure 8: Bar Chart of Extracurricular Activities | This bar chart shows that within this dataset a higher number of individuals participate in extracurricular activities compared to those who do not. 5854 individuals engage in extracurricular activities while 4146 individuals do not. There are no missing values in this column. |

|  |  |
| --- | --- |
| Figure 9: Histogram of HSC Marks | This histogram reveals a bimodal distribution, indicating two concentrations of marks: one group around the low 60 and the larger group scoring in the mid-80. The histogram shows that HSC marks range from 57 to 68 and are separated into 10 bins as well. There are no missing values in this column. |
| Figure 10: Pie Chart of Internship | This pie chart indicates that the majority of students (55.42%) completed one internship. The remaining students are divided between 0 (19.83%) and 2 (24.75%) internships completed, indicating that around 80% of students completed an internship program throughout their university days. There are no missing values in this column. |
| Figure 11: Bar Chart of Placement Status | This bar chart shows that 5803 individuals were not placed in jobs while 4197 individuals were. This indicates that non-placement is more prevalent than placement within this observed group. Since this is the target variable, sampling will have to be implemented to ensure balanced classes. There are no missing values in this column. |
| Figure 12: Bar Chart of Placement Training | The bar chart clearly shows that a significantly higher number of individuals in the dataset participated in placement training (7318) compared to those who did not (2682). This indicates that participation in placement training is substantially more common within this observed group. There are no missing values in this column. |
| Figure 13: Pie Chart of Projects | This pie chart shows the number of projects by students. Students with 0 project completion only occupy 0.3% in this dataset, while students who completed 1 project account for 35.47%, followed by students who completed 2 projects (25.50%), and majority of students completed 3 projects (38.73%). This distribution suggests that almost every student completed 1 project and more than half of the students completed two or more projects. There are no missing values in this column. |
| Figure 14: Boxplot of SSC Marks | The boxplot of SSC marks shows a distribution with a median of 70. The majority of scores fall between 59 (Q1) and 78 (Q3), with the full range being from 55 to 90. The distribution appears to be slightly left-skewed, meaning there are relatively more scores concentrated at the higher end, with a tail extending towards the lower scores. There are no missing values in this column. |
| Figure 15: Histogram of Soft Skills Rating | This histogram shows a left-skewed distribution. However, this is only the case because the range is small (3 to 4.8). The histogram shows that the majority of individuals have high soft skills ratings (2863), while a smaller number of individuals have lower ratings, creating a tail towards the left. There are no missing values in this column. |
| Figure 16: Pie Chart of Workshops / Certifications | The pie chart indicates that the largest proportion of students (37.35%) falls into a category that has zero completion of workshops/certification. This is followed by 32.7% of students who have completed 2 workshops/certification and 27.02% have completed 1 certification. A very small minority (2.75%) completed the highest number of workshops/certifications, which is 3. There are no missing values in this column. |
| Figure 17: Boxplot of CGPA vs Placement Status | The boxplot clearly indicates a strong positive correlation between higher CGPA and placement status. Students who are placed generally have significantly higher median CGPAs than those who are not placed. While the spread of the middle 50% of scores is similar, the entire distribution of CGPA for placed student is shifted upward. Besides that, there are also 2 outliers detected in the Placed category. |
| Figure 18: Bar Chart of Placement Status for each Placement Training | The bar chart strongly suggests that placement training significantly increases the likelihood of a student being placed. Of those students who did not participate in placement training, a huge majority of them did not get placed, whereas among those who did receive placement training, the majority did get placed, although it is split quite evenly. While some students get a placed without training, and some remain unplaced even with training, the proportion of placed students is substantially higher among those who undergo placement training. |
| Figure 19: 3D Bar Chart of Internships vs Extracurricular Activities vs CGPA | This multivariate bar graph is created to investigate whether non-academic activities such as extracurricular activities and internships affect a student’s CGPA. From the graph, it can be seen that students who participated in extracurricular activities and went through 2 internships have a higher mean CGPA (~8.14) than other students. This suggests a positive correlation between involvement and performance since participation in such activities usually helps the student develop time management and discipline skills. |

## **3.3 Data Quality Issues**

|  |  |
| --- | --- |
| Missing Values | None of the variables contain missing values |
| Skewed Distributions | The distributions were skewed but still within the acceptable range (-2 < x < 2) |
| Inconsistencies / Noisy Data | No inconsistencies or noisy data were spotted |
| Imbalanced Classes | The target variable contains imbalance in placed vs not placed |

A group of words on a white background

AI-generated content may be incorrect.

Figure 20: Data quality issue: imbalance classes

One significant data quality issue in this dataset is the imbalanced classes in the PlacementStatus target variable. The distribution shows that 58.03% of students are categorized as NotPlaced, while 41.97% are Placed. This slight imbalance may affect model performance, as machine learning algorithms tend to develop a bias towards predicting the majority class (in this case, NotPlaced), potentially leading to inaccurate predictions for the minority class (Placed).

To address this imbalance, sampling techniques such as oversampling the minority class or undersampling the majority class can be applied. SAS Enterprise Miner provides tools for performing these sampling methods, ensuring that both classes are adequately represented during model training. Additionally, it is important to use evaluation metrics like Precision, Recall, and F1-Score, which are more suitable for imbalanced datasets compared to accuracy, to ensure more reliable model performance.

# **4.0 Data Preprocessing**

## **4.1 Impute**

A close-up of a blue rectangle

AI-generated content may be incorrect.

Figure 21: Impute node

There were no missing values in our dataset, so impute will not be necessary. However, if there were missing values, we could use the impute node to replace missing values with the variable mode value, mean value, maximum value, and minimum value. We could also use trees to determine the missing values, which could be more accurate since it considers its relationship with other variables.

## **4.2 Drop**

A close-up of a sign

AI-generated content may be incorrect.

Figure 22: Drop node

The Drop Node is used to eliminate unnecessary or irrelevant variables from the dataset, ensuring that the predictive models only use relevant features. In this case, StudentID is identified as an irrelevant variable since it only serves as a unique identifier and does not contribute to predicting placement status. However, since we marked its role as ID, removing it will not be necessary.

## **4.3 Sample**

A close-up of a logo

AI-generated content may be incorrect.

Figure 23 Sample Node

A screenshot of a computer

AI-generated content may be incorrect.

Figure 24: Sample properties

The Sample Node in SAS Enterprise Miner adjusts the class distribution to handle imbalance in the dataset. For this project, equal sampling will be used to maintain the proportion of Placed and NotPlaced students in the sample. 100 is entered into the percentage so that the model will select 100% of the minimum class size and match that number across all classes. This method ensures that both classes are adequately represented, preventing the model from favoring the majority class and allowing for more accurate predictions on the minority class.

## **4.4 Data Partition**

A close-up of a blue rectangle

AI-generated content may be incorrect.

Figure 25 Data Partition Node

A screenshot of a data sheet

AI-generated content may be incorrect.

Figure 26 Data Set Allocations

The Data Partition Node divides the dataset into training, validation, and test sets. For this project, the data will be split as follows: 80% for training, 20% for validation, and 0% for testing. The training set will be used to build the model, while the validation set will help tune hyperparameters and prevent overfitting. The test set is not needed for this project, as the focus is on training and validating the model's performance.

## **4.5 Transform Variables Node**

A close-up of a sign

AI-generated content may be incorrect.

Figure 27 Transform Variables Node

The Transform Variables Node allows for data normalization and transformation to ensure that variables are on a comparable scale and properly prepared for modeling. Common transformations include logarithmic scaling for highly skewed data or standardization for variables with different units. These transformations help prevent variables with larger scales from dominating the model. In this project, transformations like normalization may not be necessary since the variables are relatively consistent in scale, but this method can be applied if required.

## **4.6 Replacement Node**

A close-up of a sign

AI-generated content may be incorrect.

Figure 28 Replacement Node

The Replacement Node in SAS Enterprise Miner is used to replace specific values in the dataset based on conditions or rules. For instance, incorrect values can be replaced with predefined values such as the mean, median, or mode, or even using more advanced methods like predictive imputation based on other features. While this method is often useful for datasets with missing or inconsistent data, in this project, no noisy values were identified. As such, the Replacement Node is not required. However, if inconsistencies were found, this would be the appropriate node to use for cleaning the data.

# **5.0 Decision Tree, Gradient Boosting, Random Forest, HP Tree (Law Zhuo Wei)**

## **5.1 Model Construction and Optimization**

A blue sign with black text

AI-generated content may be incorrect.

Figure 29 Decision Tree Node

A decision tree in SAS Enterprise Miner is a machine learning algorithm used for both classification and regression tasks. It divides the dataset into subsets based on feature values, creating a tree structure where each internal node represents a decision or split, and each leaf node shows the outcome. The model is easy to interpret, helping identify important variables and patterns in the data, making it useful for predictive analysis. (Vasiliki Matzavela, 2021)

### **5.1.1 Model Construction and optimization for Decision Tree**

A screenshot of a computer

AI-generated content may be incorrect.

Figure 30 Decision Tree Properties

Before I constructed the decision tree model, I changed the maximum branch to 3 to limit the complexity of the decision tree and prevent overfitting. Setting the subtree method to N and the number of leaves to 11 helps in making the model more interpretable and manageable.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 31 Decision Tree Diagram

|  |  |  |
| --- | --- | --- |
| Model Properties | Subtree Assessments and Misclassification Rate | MISC Results |
| Max Branch = 2  Subtree Method: N = 10 |  | Train = 0.2043  Validation = 0.2125 |
| Max Branch = 3  Subtree Method: N = 11 |  | Train = 0.2126  Validation = 0.2077 |
| Max Branch = 4  Subtree Method: N = 10 |  | Train = 0.2043  Validation = 0.2125 |

A screenshot of a computer

AI-generated content may be incorrect.

Figure 32 Model Comparison Table for decision tree

Justification:

The reason I changed and choose the maximum branch to 3 because to limit the complexity of the decision tree and prevent overfitting. Setting the subtree method to N and the number of leaves to 11 helps in making the model more interpretable and manageable. This configuration provides the lowest misclassification rate on validation compared to branches 2 and 4.

A screen shot of a graph

AI-generated content may be incorrect.

Figure 33 Subtree Assessment Plot for chosen Decision Tree

The decision to set the number of leaves to 11 was based on the subtree assessment plot, which shows the misclassification rates for 80% training data and 20% validation data. A total of 41 leaves were generated by the decision tree. However, the misclassification rate remained stable after the 11th leaf. As a result, the misclassification rate of the 11th leaf was selected for model construction. The difference between the training and validation misclassification rates is minimal, indicating a balanced fit without overfitting or underfitting. The minimal difference between the training and validation misclassification rates indicates a good fit, without overfitting or underfitting. The training misclassification rate is 0.2126, while the validation misclassification rate is 0.2077, confirming that the model is appropriately fitted to the data.

### **5.1.2 Model Construction and optimization for Gradient Boosting**

A close-up of a sign

AI-generated content may be incorrect.

Figure 34 Gradient Boosting Node

Gradient Boosting in SAS Enterprise Miner is a powerful ensemble learning method used for both classification and regression tasks. It builds models by combining multiple weak learners, typically decision trees, to create a stronger predictive model. The technique sequentially corrects the errors of previous models, improving accuracy. It is highly effective for handling complex datasets with non-linear relationships.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 35 Gradient Boosting Properties

Before I constructed the gradient boosting model, I changed the maximum branch to 3 and the maximum depth to 5 in the gradient boosting model to prevent overfitting and control the complexity. Keeping the interval bins at the default of 100 ensures sufficient granularity without excessive data partitioning. (Maria D. Guillen, 2023)

A screenshot of a computer

AI-generated content may be incorrect.

Figure 36 Gradient Boosting Diagram

|  |  |  |
| --- | --- | --- |
| Model Properties | Misclassification Rate and Average Square Error | MISC Results |
| Maximum Branch = 2 |  | Train = 0.1876  Validation = 0.2065 |
| Maximum Branch = 3 |  | Train = 0.1367  Validation = 0.2047 |
| Maximum Branch = 4 |  | Train = 0.1006  Validation = 0.2059 |

A screenshot of a computer

AI-generated content may be incorrect.

Figure 37 Model Comparison table for gradient boosting

The reason I chose and changed the maximum branch to 3 and the maximum depth to 5 in the gradient boosting model is because I need to control the model complexity and avoid overfitting. The default of 100 interval bins was retained to maintain a balanced granularity for data partitioning. The decision to set the maximum branch to 3 was based on the lowest misclassification rate observed on the validation set compared to branches 2 and 4. This adjustment ensures that the model remains well-structured without overfitting or underfitting. The misclassification rate for the training data is higher than that of the validation data, indicating that the model is a good fit. (Constantin Aliferis, 2024) The minimal difference between the training and validation misclassification rates confirms the model's ability to generalize well to unseen data, ensuring effective prediction.

### **5.1.3 Model Construction and optimization for Random Forest (HP Forest)**

A close-up of a sign

AI-generated content may be incorrect.

Figure 38 HP Forest node

HP Forest in SAS Enterprise Miner is a high-performance implementation of random forests, an ensemble learning method for classification and regression tasks. It builds multiple decision trees using random subsets of data and features. The model aggregates results from all trees, improving accuracy and robustness. (Bette Loef, 2022) HP Forest handles large datasets efficiently, making it suitable for complex predictive tasks.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 39 HP Forest Properties

Before I built the HP Forest model, I changed the maximum depth to 8 in the HP Forest model to balance model complexity and accuracy. This adjustment prevents overfitting, allowing the model to capture more detailed patterns without becoming too complex.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 40 HP Forest diagram

|  |  |  |
| --- | --- | --- |
| Model Properties | Iteration Plot | MISC Results |
| Maximum Depth = 8 |  | Train = 0.1496  Validation = 0.2089 |
| Maximum Depth = 50 (System default) is the same as Maximum Depth = 20 |  | Train = 0.0055  Validation = 0.2148 |

A screenshot of a computer

AI-generated content may be incorrect.

Figure 41 Model Comparison for HP Forest

The reason I changed the maximum depth to 8 in the HP Forest model is to control complexity and improve accuracy. This adjustment was made because the depth of 8 resulted in the lowest misclassification rate on validation, compared to depths of 20 and 50, the system's default. This setting prevents overfitting, as deeper trees would capture unnecessary noise in the training data. It also avoids underfitting by ensuring the model is complex enough to capture meaningful patterns. The misclassification rate for the training data is higher than the validation data, confirming a good fit without overfitting or underfitting, allowing for accurate predictions.

### **5.1.3 Model Construction and optimization for High Performance Decision Tree (HP Tree)**

A close up of a sign

AI-generated content may be incorrect.

Figure 42 HP Tree Node

HP Tree in SAS Enterprise Miner is an optimized decision tree algorithm designed for high-performance computing. It builds decision trees efficiently using parallel processing, which allows it to handle large datasets quickly. HP Tree improves model accuracy by reducing errors in classification and regression tasks, making it suitable for complex, large-scale predictive modeling. (Mohammad Arifuzzaman, 2023)

A screenshot of a computer

AI-generated content may be incorrect.

Figure 43 HP Tree Properties

Before I built the HP Tree model, I changed the maximum branch to 3 and set the maximum depth to 10, the system defaulted, to control model complexity and prevent overfitting. I chose the Cost-Complexity subtree method to improve pruning and reduce model complexity. Additionally, I selected the automatic selection method to optimize variable selection, enhancing model performance.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 44 HP Tree diagram

|  |  |  |
| --- | --- | --- |
| Model Properties | Iteration Plot | MISC Results |
| Maximum Branch = 2  Subtree Method = Cost-Complexity |  | Train = 0.2053  Validation = 0.2119 |
| Maximum Branch = 3  Subtree Method = Cost-Complexity |  | Train = 0.2007  Validation = 0.2095 |
| Maximum Branch = 4  Subtree Method = Cost-Complexity |  | Train = 0.2192  Validation = 0.2250 |

A screenshot of a computer

AI-generated content may be incorrect.

Figure 45 Model Comparison for HP Tree

The reason I changed the maximum branch to 3 and set the maximum depth to 10, the system defaulted, is to control the model's complexity and prevent overfitting. The Cost-Complexity subtree method was chosen to help prune the tree effectively and avoid overly complex models. I also set the selection method to automatic instead of N, to allow the model to optimize variable selection more effectively. The decision to set the maximum branch to 3 was based on the lowest misclassification rate observed on the validation set compared to branches 2 and 4. This choice ensures the model is neither overfitting nor underfitting. The misclassification rate for the training data is higher than that of the validation data, indicating that the model is a good fit. This confirms that the model is appropriately capturing the underlying data patterns without being too complex or too simple.

## **5.2 Model Validation**

A diagram of a software tree

AI-generated content may be incorrect.

Figure 46 Best Model Diagram

A screenshot of a computer

AI-generated content may be incorrect.

Figure 47 Model Comparison for Finding the Best Model

In this section, I will compare the best models from each scenario such as Decision Tree, Gradient Boosting, HP Forest, and HP Tree, then select the model with the highest accuracy. After testing the different configurations, the Gradient Boosting model emerged as the best performer, with the highest accuracy of 79.53%. It achieved a misclassification validation rate (MISC) of 0.2047, making it the top model among the four tested. This result indicates that Gradient Boosting effectively captured the data's underlying patterns and performed well on unseen data.

The second-best model is the Decision Tree, with a maximum branch of 3, which resulted in a validation misclassification rate of 0.2077, leading to an accuracy of 79.23%. Although it performed slightly lower than Gradient Boosting, it still delivered reliable results, with a minimal difference in accuracy.

In third place is the HP Forest model, which used a maximum depth of 8. It achieved a misclassification validation rate of 0.2089, corresponding to an accuracy of 79.11%. While this model performed well, it was slightly outpaced by the Decision Tree and Gradient Boosting models.

Finally, the HP Tree model, with a maximum branch of 3, had the lowest accuracy of 79.05% and a misclassification validation rate of 0.2095. Despite being the least accurate, it still performed reasonably well but did not match the results of the other models.

Based on this comparison, the Gradient Boosting model is selected as the optimal model for this project.

## **5.3 Critical Interpretation of Outcomes**

### **5.3.1 Fit Statistic of Ensemble Node (Gradient Boosting 3 Branch)**

A blue rectangle with black text

AI-generated content may be incorrect.

Figure 48 Ensemble Node

The Ensemble node in SAS Enterprise Miner combines multiple predictive models into one ensemble model. This approach enhances accuracy by aggregating predictions from different models, allowing them to complement each other. By selecting the highest accuracy model, Gradient Boosting, the ensemble model improves predictive performance and minimizes errors, creating a more reliable outcome. (Yen-Liang Chen, 2022)

A screenshot of a computer

AI-generated content may be incorrect.

Figure 49 Fit Statistic of ensemble (Gradient Boosting 3 branch)

The **Misclassification Rate (MISC)** for the validation set is **0.2047**, which means that approximately **20.47%** of the predictions made by the model during the validation phase were incorrect. This metric represents the proportion of misclassified instances out of the total number of instances in the validation set. A misclassification rate of 0.2047 implies that around 20% of the time, the model failed to predict the correct placement status for students.

To calculate **accuracy**, we subtract the misclassification rate from 1 and then multiply by 100 to get the percentage of correct predictions:

**Accuracy = (1 − 0.2047) × 100 = 79.53%**

This means that the model successfully predicted the placement status of students about **79.53%** of the time during the validation phase. While this accuracy is quite good, it indicates that there is still room for improvement. The model correctly identified the placement status for almost 80% of the cases but made errors in the remaining 20%.

The misclassification rate of 0.2047 highlights that the model is relatively effective but may need further tuning or adjustments in feature selection, model parameters, or data quality to improve its performance. Lowering the misclassification rate and increasing the accuracy could further enhance the model's predictive capabilities.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 50 Fit Statistic 2

In the **Ensemble Node**, the **Misclassification Rate (MISC)** for the **train** set is **0.14**, while for the **validation** set, it is **0.20**.

### **5.3.2 Event Classification Table: Analysis and Calculations**

In this section, we will analyse the Event Classification Table for both the training and validation datasets. We will calculate key performance metrics such as Accuracy, Precision, Recall, Specificity, and F1 Score based on the provided data. These metrics help evaluate the performance of the model in predicting the Placement Status of students, which is the target variable.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 51 Event Classification Table of Best Model (Gradient Boosting)

**Training Data:**

The **training data** provided gives the following classification counts:

|  |  |  |
| --- | --- | --- |
| Actual \ Predicted | Not Placed (Predicted) | Placed (Predicted) |
| Not Placed (Actual) | 2914 (True Negative - TN) | 443 (False Positive - FP) |
| Placed (Actual) | 475 (False Negative - FN) | 2882 (True Positive - TP) |

* **True Negative (TN)**: 2914 instances where the model correctly predicted "Not Placed".
* **False Positive (FP)**: 443 instances where the model incorrectly predicted "Placed" for students who were actually "Not Placed".
* **False Negative (FN)**: 475 instances where the model incorrectly predicted "Not Placed" for students who were “Placed”.
* **True Positive (TP)**: 2882 instances where the model correctly predicted "Placed".

**Validation Data:**

The **validation data** provided gives the following classification counts:

|  |  |  |
| --- | --- | --- |
| Actual \ Predicted | Not Placed (Predicted) | Placed (Predicted) |
| Not Placed (Actual) | 675 (True Negative - TN) | 165 (False Positive - FP) |
| Placed (Actual) | 179 (False Negative - FN) | 661 (True Positive - TP) |

Justification:

* **True Negative (TN)**: 675 instances where the model correctly predicted "Not Placed".
* **False Positive (FP)**: 165 instances where the model incorrectly predicted "Placed" for students who were actually "Not Placed".
* **False Negative (FN)**: 179 instances where the model incorrectly predicted "Not Placed" for students who were “Placed”.
* **True Positive (TP)**: 661 instances where the model correctly predicted "Placed".

**Calculation:**

* **Accuracy:** (675 + 661) / (675 + 165 + 179 + 661) = 1336 / 1680 = **0.7952 (79.52%)**
* **Precision:** 661/ (661 + 165) = **0.8002 (80.02%)**
* **Recall:** 661/ (661 + 179) = **0.7869 (78.69%)**
* **Specificity:** 675 / (675 + 165) = **0.8035 (80.35%)**
* **F1 Score**: 2\* [ (0.8002 \* 0.7869) / (0.8002 + 0.7869)] = **0.7934 (79.34%)**

**Justification:**

After analysing the classification results for the **training data**, we can calculate several key performance metrics to assess the model's effectiveness in predicting student placement status.

For **accuracy**, the model achieved **79.52%**, which means that it correctly classified the placement status of 79.52% of the students in the validation dataset. **Precision** is **80.02%**, indicating that the model correctly predicted "Placed" for 80.02% of the students it classified as placed. This shows that the model has a relatively low rate of false positives, meaning it seldom misclassifies "Not Placed" students as "Placed."

The **recall**, at **78.69%**, demonstrates the model’s ability to correctly identify most of the students who were Placed. However, it still missed around 21.31% of those students. **Specificity**, calculated at **80.35%**, tells us that the model correctly classified 80.35% of students who were **Not Placed**. This suggests that the model is equally reliable in identifying students who were not placed in jobs.

Finally, the **F1 Score** is **79.34%**, which is a balanced measure between precision and recall. It shows that the model maintains a good equilibrium between accurately identifying placed students and minimizing false positives.

All in all, the model performs well with balanced results across these metrics, indicating that it is reliable for predicting student placement status and identifying those who need additional support. However, there is room for improvement, particularly in boosting recall for placed students.

**Findings Conclusion:**

The comparison of Decision Tree, Gradient Boosting, Random Forest (HP Forest), and High-Performance Decision Tree (HP Tree) revealed significant differences in performance. Gradient Boosting emerged as the most accurate model with an impressive 79.53% accuracy, followed closely by Decision Tree at 79.23%. While HP Forest and HP Tree delivered decent results, their performance lagged slightly behind. These findings suggest that Gradient Boosting, with its ensemble nature, offers superior prediction accuracy for student placement outcomes. Decision Trees, however, provide a more interpretable alternative, making them a strong contender for applications requiring explainability alongside accuracy.

# **6.0 Regression Modelling (Eugene Tan Ting Siang)**

A blue rectangular sign with black text

AI-generated content may be incorrect.

Figure 52: Regression node

Regression modelling is a fundamental statistical technique used to examine the relationships between variables (Dodge, 2008). Not only can it be used to explore data associations, but it can also be used for predictive purposes. It allows researchers to quantify the influence of predictors, identify trends, and forecast outcomes based on observed data. There are many types of regression like simple regression, forward selection regression, backward elimination regression, stepwise regression, and more. Simple regression uses all independent variables in the dataset; forward selection regression starts with no variables and slowly adds them one by one; backwards elimination regression starts with all variables and removes each variable one by one; and stepwise regression is a combination of the former two models. These techniques help improve model accuracy and interpretability, especially when dealing with large datasets containing multiple predictors (James et al., 2021).

## **6.1 Model Construction and Optimization**

Before modelling the regression model, some settings will have to be adjusted.

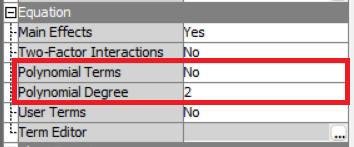


Figure 53: Regression equation properties

The first settings are the polynomial terms and polynomial degree in the Equation section. Polynomial terms refer to the inclusion of higher-order terms of input variables in the regression model. Selecting “Yes” for this option would transform the input features by raising them to powers (x, x2, x3) to allow the model to capture nonlinear relationships. As for polynomial degree, it refers to the highest power of the independent variable included in the model. By increasing the degree, we can fit more complex patterns in the data, but this increases the chance of overfitting, where the model understands the training data too well and is unable to generalize unseen data properly.

A close up of a login box

AI-generated content may be incorrect.

Figure 54: Regression class targets properties

Next is the regression type and link function in the Class Targets section. Since our target is a binary data type, we will be using logistic regression instead of linear regression and a logit link function for all cases. This allows the model to predict probabilities (between 0 and 1) instead of unbounded values.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 55: Regression model selection properties

Besides that, the selection model and selection criterion can be modified as well in the Model Selection section. Changing the selection model will allow us to create forward selection regression models, backward elimination regression models, and stepwise regression models. As for the selection criterion, we went with validation misclassification to evaluate the misclassification rate of the models since we are trying to predict binary values for unseen data.

### **6.1.1 Regression without Variable Selection**

To start the analysis, a series of regression models without the involvement of the variable selection node will be created. This will include all variables inside the models, regardless of whether it is highly associated with the target or not.

A diagram of a data flow

AI-generated content may be incorrect.

Figure 56: Regression diagram (1)

|  |  |  |
| --- | --- | --- |
| **Model Name & Properties** | **Iteration Plot** | **Results** |
| **Simple Regression**  Polynomial Terms: No  Polynomial Degree: 2  Selection Model: None  Selection Criterion: Default | None | Train Misc:  0.200923  Valid Misc:  0.208929  Accuracy:  79.1% |
| **Forward Regression**  Polynomial Terms: No  Polynomial Degree: 2  Selection Model: Forward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.205392  Valid Misc:  0.207143  Accuracy:  79.3% |
| **Backward Regression**  Polynomial Terms: No  Polynomial Degree: 2  Selection Model: Backward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.200923  Valid Misc:  0.208929  Accuracy:  79.1% |
| **Stepwise Regression**  Polynomial Terms: No  Polynomial Degree: 2  Selection Model: Stepwise  Selection Criterion: Validation Misclassification |  | Train Misc:  0.205392  Valid Misc:  0.207143  Accuracy:  79.3% |
| **2nd Degree Polynomial Forward Regression**  Polynomial Terms: Yes  Polynomial Degree: 2  Selection Model: Forward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.202413  Valid Misc:  0.202381  Accuracy:  79.8% |
| **2nd Degree Polynomial Backward Regression**  Polynomial Terms: Yes  Polynomial Degree: 2  Selection Model: Backward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.199434  Valid Misc:  0.20119  Accuracy:  79.9% |
| **2nd Degree Polynomial Stepwise Regression**  Polynomial Terms: Yes  Polynomial Degree: 2  Selection Model: Stepwise  Selection Criterion: Validation Misclassification |  | Train Misc:  0.202413  Valid Misc:  0.202381  Accuracy:  79.8% |
| **3rd Degree Polynomial Forward Regression**  Polynomial Terms: Yes  Polynomial Degree: 3  Selection Model: Forward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.206881  Valid Misc:  0.201786  Accuracy:  79.8% |
| **3rd Degree Polynomial Backward Regression**  Polynomial Terms: Yes  Polynomial Degree: 3  Selection Model: Backward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.197945  Valid Misc:  0.213095  Accuracy:  78.7% |
| **3rd Degree Polynomial Stepwise Regression**  Polynomial Terms: Yes  Polynomial Degree: 3  Selection Model: Forward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.206881  Valid Misc:  0.201786  Accuracy:  79.8% |

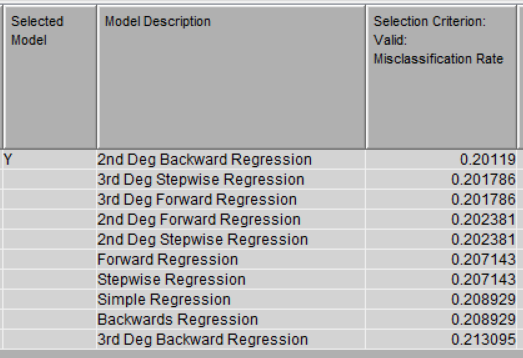


Figure 57: Regression comparison table (1)

This table is a comparison of all models with their misclassification rates. It can be seen that 2nd Degree Polynomial Backward Regression was chosen as the best model out of the 10 models due to it having the lowest misclassification rate (accuracy of 79.9%). On the contrary, 3rd Degree Polynomial Backward Regression is considered the worst model because of its lower accuracy (78.7%).

### **6.1.2 Regression with Variable Selection**

Apart from the polynomial degree and the regression selection model, the variables involved also play an important role in determining the model’s performance. In this section, we experiment on how different combinations of variables will impact the accuracy of the models by using the variable selection node to perform feature selection.

A close-up of a button

AI-generated content may be incorrect.

Figure 58: Variable Selection Node

In the SAS diagram, the variable selection node was placed before the data partition node. This node works by selecting variables based on the minimum chi-square and R-square criteria. If any variable fails to meet the criteria, the variable will be rejected, and it will not be used in building the predictive model. This ensures that only the most relevant predictors are retained, thus potentially increasing the model’s accuracy.

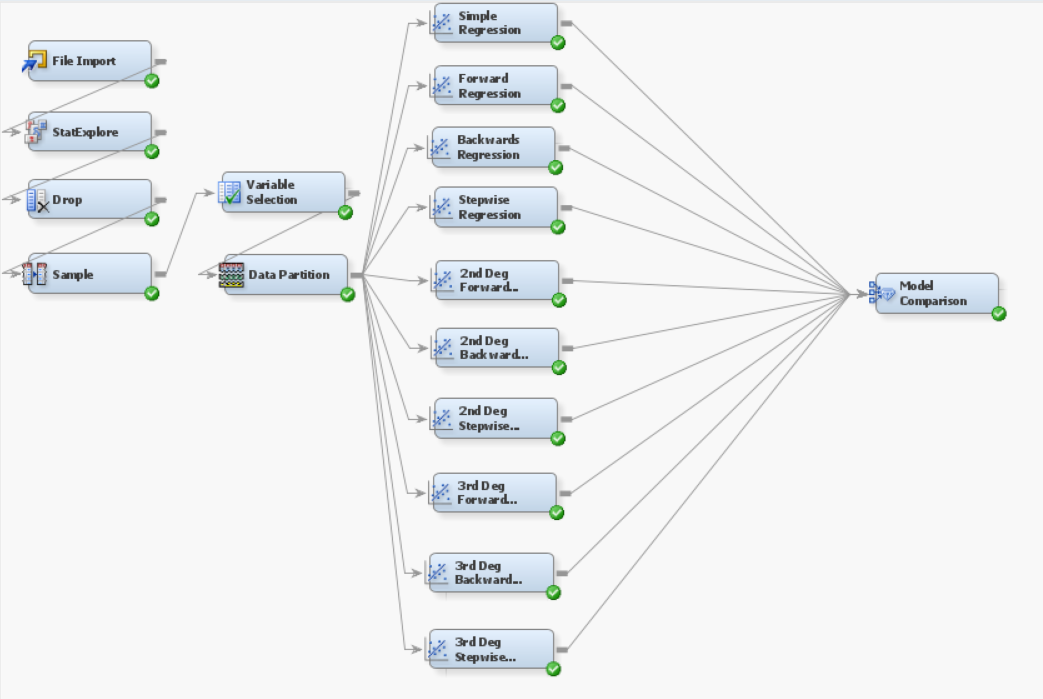


Figure 59: Regression diagram (2)

After using the default criteria of minimum chi-square = 3.84 and minimum r-square = 0.005, it was observed that only a few variables were excluded from the model due to small r-square values. Therefore, to analyze the impact when more variables were excluded, the r-square criteria was slowly increased, and the changes were observed and recorded.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 60: Variables rejected when default criteria were used

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Minimum R-square** | **No. of Variables Selected** | **Variables** | **Model** | **Misc Rate** |
| 0.005  (Default) | 9 | **Accepted:**  AptitudeTestScore  CGPA  ExtracurricularActivities  HSC\_Marks  PlacementTraining  Projects  SSC\_Marks  SoftSkillsRating  Workshops\_Certifications  **Rejected:**  StudentID  Internships | Simple Regression | 0.209524 |
| Forward Regression | 0.207143 |
| Backward Regression | 0.209524 |
| Stepwise Regression | 0.207143 |
| 2nd Degree Polynomial Forward Regression | 0.202381 |
| 2nd Degree Polynomial Backward Regression | 0.207143 |
| 2nd Degree Polynomial Stepwise Regression | 0.202381 |
| 3rd Degree Polynomial Forward Regression | 0.201786 |
| 3rd Degree Polynomial Backward Regression | 0.208929 |
| 3rd Degree Polynomial Stepwise Regression | 0.201786 |
| 0.12 | 8 | **Accepted:**  AptitudeTestScore  CGPA  ExtracurricularActivities  HSC\_Marks  Projects  SSC\_Marks  SoftSkillsRating  Workshops\_Certifications  **Rejected:**  StudentID  Internships  PlacementTraining | Simple Regression | 0.207738 |
| Forward Regression | 0.207738 |
| Backward Regression | 0.207738 |
| Stepwise Regression | 0.207738 |
| 2nd Degree Polynomial Forward Regression | 0.205952 |
| 2nd Degree Polynomial Backward Regression | 0.205357 |
| 2nd Degree Polynomial Stepwise Regression | 0.205952 |
| 3rd Degree Polynomial Forward Regression | 0.208333 |
| 3rd Degree Polynomial Backward Regression | 0.213095 |
| 3rd Degree Polynomial Stepwise Regression | 0.208333 |
| 0.15 | 7 | **Accepted:**  AptitudeTestScore  CGPA  ExtracurricularActivities  HSC\_Marks  Projects  SSC\_Marks  SoftSkillsRating  **Rejected:**  StudentID  Internships  PlacementTraining  Workshops\_Certifications | Simple Regression | 0.207738 |
| Forward Regression | 0.207738 |
| Backward Regression | 0.207738 |
| Stepwise Regression | 0.207738 |
| 2nd Degree Polynomial Forward Regression | 0.202381 |
| 2nd Degree Polynomial Backward Regression | 0.204762 |
| 2nd Degree Polynomial Stepwise Regression | 0.205952 |
| 3rd Degree Polynomial Forward Regression | 0.208929 |
| 3rd Degree Polynomial Backward Regression | 0.204762 |
| 3rd Degree Polynomial Stepwise Regression | 0.208929 |
| 0.19 | 6 | **Accepted:**  AptitudeTestScore  ExtracurricularActivities  HSC\_Marks  Projects  SSC\_Marks  SoftSkillsRating  **Rejected:**  StudentID  Internships  PlacementTraining  Workshops\_Certifications  CGPA | Simple Regression | 0.214286 |
| Forward Regression | 0.208929 |
| Backward Regression | 0.214286 |
| Stepwise Regression | 0.208929 |
| 2nd Degree Polynomial Forward Regression | 0.214286 |
| 2nd Degree Polynomial Backward Regression | 0.211310 |
| 2nd Degree Polynomial Stepwise Regression | 0.214286 |
| 3rd Degree Polynomial Forward Regression | 0.213095 |
| 3rd Degree Polynomial Backward Regression | 0.208333 |
| 3rd Degree Polynomial Stepwise Regression | 0.213095 |

From the results gathered, it was observed that the model with the highest accuracy was found in the category when 9 variables were selected, which is the 3rd Degree Polynomial Forward Regression with an accuracy of 79.8%. This model will be selected to compare with other best performing regression models in later sections of the report.

### **6.1.3 HP Regression**

A close-up of a sign

AI-generated content may be incorrect.

Figure 61: HP Regression node

HP Regression, also known as high performance regression, is a different type of regression that we can use to build our model. It utilizes parallel processing to create models at a faster rate, which is why it is useful for very large datasets like ours that have 10000 records. In this section, we will experiment with using this node to see if it can predict placement status better than the other regression model.

A diagram of data processing

AI-generated content may be incorrect.

Figure 62: HP Regression diagram

For the HP regression models, the parameters used, such as the regression type, polynomial degree, and selection criterion are the same as those used in regular regression previously. While HP regression provides additional selection methods in the model selection section, such as LARS and LASSO which tackle the variable selection problems, these methods only work on interval targets. Since our target is binary, these methods are not applicable and will not be used in our models.

|  |  |  |
| --- | --- | --- |
| **Model Name & Properties** | **Fit Statistics** | **Results** |
| **Forward Regression**  Polynomial Terms: No  Polynomial Degree: 2  Selection Model: Forward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.200477  Valid Misc:  0.209524  Accuracy:  79.0% |
| **Backward Regression**  Polynomial Terms: No  Polynomial Degree: 2  Selection Model: Backward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.200626  Valid Misc:  0.210119  Accuracy:  78.9% |
| **Stepwise Regression**  Polynomial Terms: No  Polynomial Degree: 2  Selection Model: Stepwise  Selection Criterion: Validation Misclassification |  | Train Misc:  0.200477  Valid Misc:  0.209524  Accuracy:  79.0% |
| **2nd Degree Polynomial Forward Regression**  Polynomial Terms: Yes  Polynomial Degree: 2  Selection Model: Forward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.201966  Valid Misc:  0.210714  Accuracy:  78.9% |
| **2nd Degree Polynomial Backward Regression**  Polynomial Terms: Yes  Polynomial Degree: 2  Selection Model: Backward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.198987  Valid Misc:  0.207143  Accuracy:  79.3% |
| **2nd Degree Polynomial Stepwise Regression**  Polynomial Terms: Yes  Polynomial Degree: 2  Selection Model: Stepwise  Selection Criterion: Validation Misclassification |  | Train Misc:  0.201966  Valid Misc:  0.210714  Accuracy:  78.9% |
| **3rd Degree Polynomial Forward Regression**  Polynomial Terms: Yes  Polynomial Degree: 3  Selection Model: Forward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.203307  Valid Misc:  0.214881  Accuracy:  78.5% |
| **3rd Degree Polynomial Stepwise Regression**  Polynomial Terms: Yes  Polynomial Degree: 3  Selection Model: Forward  Selection Criterion: Validation Misclassification |  | Train Misc:  0.203307  Valid Misc:  0.214881  Accuracy:  78.5% |

A screenshot of a computer

AI-generated content may be incorrect.

Figure 63: HP Regression comparison table

Out of all the models, the 2nd Degree Polynomial Backward HP Regression was selected as the best performing model due to its lowest misclassification rate and highest accuracy (79.3%).

## **6.2 Model Validation**

This section will focus on comparing the best models from each scenario and selecting the one with the highest accuracy. When conducting regression without variable selection, 2nd Degree Polynomial Backward Regression was chosen as the best model with an accuracy of 79.9%. When conducting regression with variable selection, 3rd Degree Polynomial Forward Regression was chosen as the best model when 9 variables were selected with an accuracy of 79.8%. As for HP regression, 2nd Degree Polynomial Backward HP Regression was selected as the best performing model with an accuracy of 79.3%.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 64: Regression comparison diagram

A screenshot of a computer program

AI-generated content may be incorrect.

Figure : Regression comparison table (2)

From the comparison, it can be seen that 2nd Degree Polynomial Backward Regression was selected as the best performing model because it has the highest accuracy. We will be evaluating the outcomes of this model in the next section.

## **6.3 Critical Interpretation of Outcomes**

### **6.3.1 Fit Statistics**

A screenshot of a computer

AI-generated content may be incorrect.

Figure 66: Fit Statistics of 2nd Degree Polynomial Backward Regression

This table shows many statistics for the regression model, but we will only be focusing on the misclassification rate (MISC) to evaluate the model’s performance. From the table, we can see that the model’s validation misclassification rate is 0.20119. By subtracting this number from 1 and multiplying it by 100%, we can calculate the accuracy of the model which is 79.881%. This means that the model is able to correctly predict a student’s placement status about 79.881% of the time.

### **6.3.2 Iteration Plot**

A graph showing a number of different colored lines

AI-generated content may be incorrect.

Figure 67: Iteration Plot of 2nd Degree Polynomial Backward Regression

The figure above shows the iteration plot of the 2nd Degree Polynomial Backward Regression model’s misclassification rate. This graph shows how the misclassification rates change as multiple iterations of the model are trained. The red line indicates the validation data misclassification rates, whereas the blue line represents the testing data misclassification rates.

As seen in the graph, the model selected at step 4 has the best balance between training and validation error and it has the lowest misclassification rate of 0.20119. After step 4, the validation misclassification rate tends to increase more and more, which suggests that underfitting begins to occur beyond that point. As for the training error, it seems to remain low throughout, reinforcing that further steps may overfit the training data without improving generalization. It also continues to improve slightly after step 4, but at the cost of validation performance which is not desirable. Therefore, the most optimal model was chosen at step 4 to prevent overfitting while ensuring solid predictive performance on new data.

### **6.3.3 Summary of Backward Elimination**

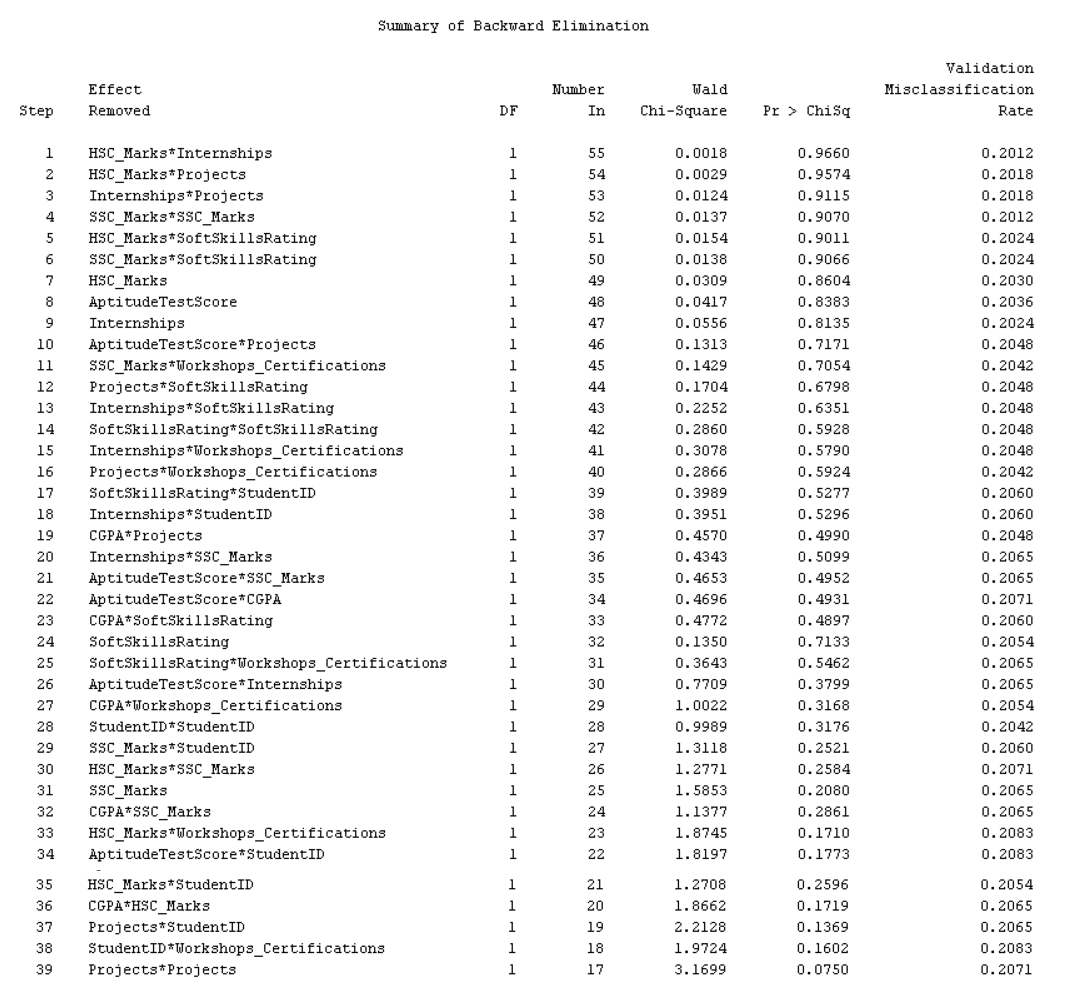


Figure 68: Summary of backward elimination

The figure above is the summary of backward elimination for the regression model in a table. It shows the steps or iterations of the model and how at each step, different variables were removed starting from the least significant one and the resulting validation misclassification rate was shown. In addition, the table also shows the degree of freedom (DF), the number of variables still in the model after a variable was removed (Number In), the significance of the removed variable (Wald Chi-Square), and the p-value of the test, where higher p-value means less significant (Pr > ChiSq).

The model started off by removing variables with little statistical importance (variables with high p-values > 0.9). The best misclassification rate (0.2012) was found in step 4, after removing the variable SSC\_Marks\*SSC\_Marks and leaving the model with only 52 variables. After step 4, there is a trend where the validation misclassification rate slowly increases until the final step of step 39, which is a typical sign of underfitting. At this point, only 17 variables remain, and the model is very simplified, but error has increased to 0.2071.

Therefore, it can be deduced that the optimal model is at step 4 with 52 variables and going beyond this point sacrifices predictive accuracy.

### **6.3.4 Analysis of Maximum Likelihood Estimates**

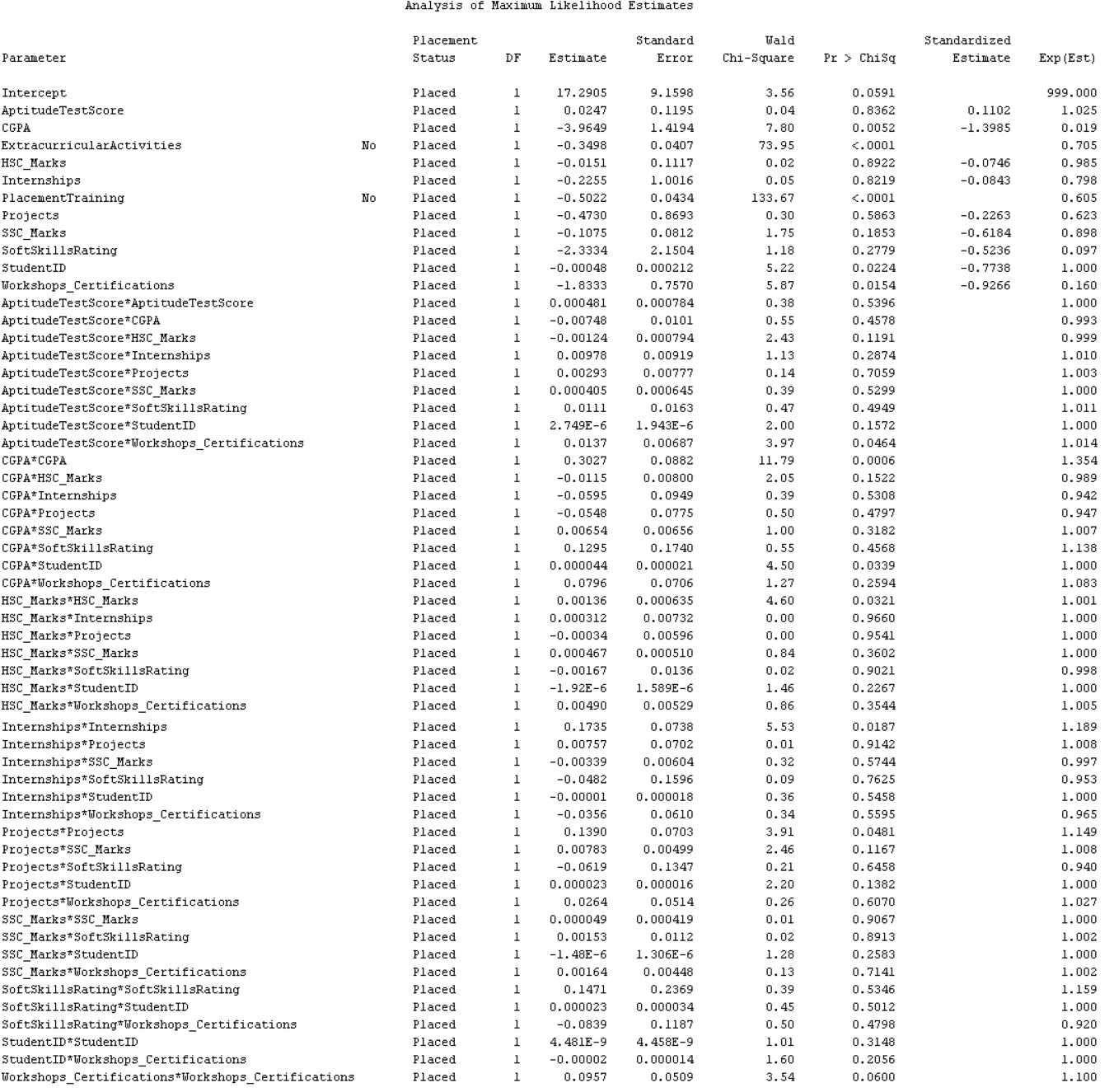


Figure 69: Analysis of maximum likelihood estimates

This table shows how each parameter contributes to the likelihood of the outcome variable (in this case, the placement status Placed vs Not Placed). The table has 9 columns, starting with the parameter, which is the variable used in the model, followed by the outcome of the predicted placement status, degree of freedom, estimate, standard error, Wald Chi-Square, p-value, standardized estimate, and the odds ratio (Exp(Est)).

From the table, it can be seen that the most statistically significant predictors are extracurricular activities = no with high Wald Chi-Square value of 73.95 and a very low p-value in the Pr > ChiSq column (<0.0001), as well as placement training = no with an even higher Wald Chi-Square value of 133.67 and the same p-value. On the other hand, some non-significant variables can also be spotted due to their high p-values, such as aptitude test scores, internships, and HSC marks. These variables are not useless, but they just do not explain variation in placement status at a statistically confident level.

Besides that, the table also shows that the Exp(Est) (odds ratio) of extracurricular activities = no is 0.705. This odds ratio shows how the odds of the outcome change with a one-unit change in the predictor variable, compared to a reference group, while holding other variables constant. An odd ratio greater than 1 signifies positive association, less than 1 signifies negative association, and equal to 1 means that there is no effect. In this case, 0.705 is less than 1, which suggests that students who do not participate in extracurricular activities have 29.5% lower odds (1 - 0.705 = 0.295) of getting placed compared to students who do participate in extracurricular activities. Similarly, the odds ratio for placement training = no is 0.605, which means that students who do not go through placement have 39.5% lower odds (1 - 0.605 = 0.395) of being placed, compared to those who received training. This indicates a stronger negative association with placement than the absence of extracurricular activities.

### **6.3.5 Odds Ratio Estimates**

A white background with black text

AI-generated content may be incorrect.

Figure 70: Odds ratio estimates

The odds ratio estimates table shows how the two factors (extracurricular activities and placement training) affect the placement status, similar to what was explored previously in the analysis of maximum likelihood estimates. From the table, we can see that extracurricular activities (No vs Yes) have a point estimate of 0.497. This indicates that students who did not participate in extracurricular activities (No) had 0.497 times the odds of being placed compared to those who did participate (Yes). As for placement training (No vs Yes), it has a point estimate of 0.366, which indicates that students without placement training (No) had 0.366 times the odds of being placed compared to those with training (Yes).

These results reiterate that both extracurricular activities and placement training significantly improve placement opportunities for students, even though placement training has a larger impact than extracurricular activities. Therefore, students are encouraged to participate in more events or clubs and enter more placement training programs if they want to have a better chance at finding a job after graduation.

### **6.3.6 Event Classification Table**

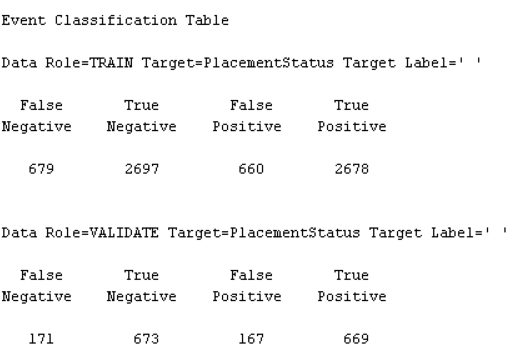


Figure 71: Event classification table

The figure above provides an event classification table for both the training and validation datasets. This table enables us to evaluate the performance of the regression model through false negatives (model predicted negative when it should be positive), true negatives (model correctly predicted negative), false positives (model predicted positive when it should be negative), and true positives (model correctly predicted positive).

**Training Data:**

* **False Negative (FN):** The training dataset incorrectly classified 679 students’ placement status as not placed.
* **True Negative (TN):** 2697 students’ placement status were correctly classified as not placed in the training dataset.
* **False Positive (FP):** The training dataset incorrectly classified 660 students’ placement status as placed.
* **True Positive (TP):** 2678 students’ placement status were correctly classified as placed in the training dataset.

|  |  |  |
| --- | --- | --- |
|  | **Predicted: Not Placed** | **Predicted: Placed** |
| **Actual: Not Placed** | 2697 (**TN**) | 660 (**FP**) |
| **Actual: Placed** | 679 (**FN**) | 2678 (**TP**) |

**Validation Data:**

* **False Negative (FN):** The validation dataset incorrectly classified 171 students’ placement status as not placed.
* **True Negative (TN):** 673 students’ placement status were correctly classified as not placed in the validation dataset.
* **False Positive (FP):** The validation dataset incorrectly classified 167 students’ placement status as placed.
* **True Positive (TP):** 669 students’ placement status were correctly classified as placed in the validation dataset.

|  |  |  |
| --- | --- | --- |
|  | **Predicted: Not Placed** | **Predicted: Placed** |
| **Actual: Not Placed** | 673 (**TN**) | 167 (**FP**) |
| **Actual: Placed** | 171 (**FN**) | 669 (**TP**) |

**Accuracy:** (673 + 669) / (673 + 167 + 171 + 669) = 1342 / 1680 = **0.7988 (79.88%)**

**Precision:** 669 / (669 + 167) = **0.8005 (80.05%)**

**Recall:** 669 / (669 + 171) = **0.7965 (79.65%)**

**Specificity:** 673 / (673 + 167) = **0.8012 (80.12%)**

**F1 Score:** 2 \* [ (0.8005 \* 0.7965) / (0.8005 + 0.7965) ] = **0.7985 (79.85%)**

After calculating the key metrics for the validation data, we can see that the regression model created has 79.88% accuracy. The precision, which measures how reliable the positive predictions are, is 80.05%, which means the model makes relatively few false placement predictions. Besides that, a recall of 79.65% indicates that the model successfully identifies most of the students who were actually placed. On the other hand, the calculated specificity shows that of all students who were not placed, the model correctly labeled 80.12% of them as "Not Placed". Lastly, the model had an F1 score of 79.85%, which means that the model maintains a good balance between catching placed students and being accurate about it.

Overall, this model showed balanced and consistent performance across all metrics, so it is quite a well performing and reliable model for predicting student placement status.

# **7.0 Clustering (Beh Swee Shen)**

## **7.1 Model Construction and Optimization**

A close-up of a sign

AI-generated content may be incorrect.

Figure 72: Cluster Node

Clustering is a technique used in various fields of computer science, including machine learning, artificial intelligence, data analysis, and data mining, to group comparable data points based on information contained within them, such as traits or attributes. This method is very effective when working with huge datasets, that contain unstructured unlabelled or incomplete data (Nancholas, 2023).

A screenshot of a computer

AI-generated content may be incorrect.

Figure 73: Cluster’s properties

Before building the model, the specification method is selected as automatic, which means that the system typically employs one or more statistical criteria to evaluate different number of clusters; while the maximum number of cluster set the highest number of groups the system will consider.



Figure 74: Model

A close-up of a sign

AI-generated content may be incorrect.

Figure 75: Data Partition Node

Data partitioning is the process of dividing big datasets into smaller, more manageable segments known as partitions. This process is designed to improve data organization, management, and retrieval. It is critical in database management and data warehousing since it allows for faster data access and fewer I/O operations (Dremio, 2024). In my case, I split all the data into 60% of training 20% of validate and 20% to test to preventing overfitting and increase model evaluation.

A pie chart with numbers with Crust in the background

AI-generated content may be incorrect.

Figure 76: First Cluster segment size

A screenshot of a computer

AI-generated content may be incorrect.

Figure 77: First Cluster segment plot

Based on this proportion plots, ExtracurricularActivities shows that majority of the bar for segment 1 is orange (Yes) and minority is blue. Come to segment 2, it almost entirely “Yes” for extracurricular (very high orange proportion) and segment 3 is a mix with a notable proportion of “No” but also substantial of “Yes”. Lastly, segment 4 showing predominantly “No”, indicate that segments 3 and 4 are characterized by a lack of extracurricular activities while segment 1 and 2 is strongly characterized by the presence of extracurricular.

Beside that, the variable 2 internship plots, from bottom to top: Light green (0), Darker green (0.25), Light brown (0.75-1), Purple (1.25) which implies that the number doesn’t perfectly align with typical internship counts like 0, 1, 2, 3. They might represent categories of weighted values and lets assuming it as a ranges of internship counts For segment 1, it shows a high percentage in the lower ranges (light green, light brown and a little of purple) , indicating that fewer or no internships. While segment 2 shows a significant proportion in the purple and light brown, although there is a substantial lower range of light green but it still likely indicating more internships. Segments 3 and 4 are similarly to segment 2, suggesting there is a fewer or no internships.

Come to Placement Training, the segments 1,2 and 4 are showing a largely “Yes” for placement training while segment 3 shows an entire “No” for placement training. Segment 1 and 2 stands out with nearly all individuals undergone placement training, while segment 3 and 4 primarily consists of individuals with/without placement training.

Variable projects are categorized into ranges which similarly to internships as (light blue = 0.75-1.125, light pink = 1.875,2.25, light brown = 2.625-3). Segment 1 shows a great mix of light blue, light pink and light brown indicate that one ore more project. While segment 2 shows a much higher proportion of light brown which tells that individuals with the most projects. Segments 3 and 4 are similar with mixed profile, with a fair amount in the mid-to-higher ranges, but also significant lower ranges.

Variable soft skills rating is categorized into mange fine-grained range from 0.0375 up to 4.125, this suggest rating scale that has been binned. Segment 1 appears to have a distribution that from lower to higher soft skills rating, while segment 2 have even more higher distribution across various soft skills rating, possibly indicating a broader range or higher average. Segments 3 and 4 are more concentrated in the lower soft skills rating ranges.

Variable 6 workshops certifications are a numerical ranges as internships and projects from ( 0-0.375,0.75-1.125,1.875-2.25,2.625-3). Segments 1 are primarily in the lower ranges, suggesting fewer or no workshops certifications; while segment 2 shows a higher proportion of 1.875-2.25 which indicating individuals in this group are like to have one or more workshops certification. Segments 3 and 4 are similar with a mixed distribution but mostly in the lower ranges with few or no workshops certifications.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 78: First Cluster mean statistics

In the mean statistics, cluster 2 have the largest data points at the group which is 2993, and has the lowest Root-Mean-Square standard deviation (0.53334), indicating it is the most cohesive or “tight” cluster; and also perform a 3.534939 in maximum distance from cluster seed suggesting that a less potential outlier. Come to the variable, internships, projects, soft skills rating and workshops/certification, cluster 2 are having a mean of 1.657534 internships, which is higher than cluster 1 and cluster 3, which tells that cluster 2 tend to have more advantages. To conclude this results, segment 2 stands out as the most “accomplished” characterized by high engagement across all development activities and excellent soft skills, while segment 1 shows a nearest neighbour of cluster 2 it is also highly engaged, particularly in extracurriculars and placement training, but slightly less extensive in internships and project than segment 2. Lastly, segment 3 and 4 represent individuals with significantly lower engagement in most of the aspects but slightly better than segment 3 with better soft skills but still moderate.

A white screen with black text

AI-generated content may be incorrect.

Figure 79: First Cluster variable importance

This tables shows that the input variables I choose for cluster. The internships variable has the highest importance score of 1 indicating a strong influence while soft skills rating and project followed closely, but they still perform well in major differentiators in student profiles. Come to placement training and workshops certification, they also play meaningful roles. Lastly, although extracurricular activities have the least impact on cluster formation, but it still shows with a 0.703 shows non- academic activities have some effect on segmentation.

A screenshot of a pie chart with Crust in the background

AI-generated content may be incorrect.

Figure 80: 5 Cluster

A screenshot of a computer

AI-generated content may be incorrect.

Figure 81: 5 Cluster segment plot

The first top left variable extracurricular activities, for segment 1 it is almost entirely “No” with a slightly of ‘Yes” indicating low participation in this segment which is on the contrary of segment 3 shows a high proportion of participation. While come to segment 2, it has a great mix of yes and no nearly half for each indicating individuals in this segment might participate in extracurricular or might not. Comes to segment 3 and segment 4, segment3 almost entirely “Yes” with a slightly of “No” indicating most individuals participate, and segment 4 shows an entire proportion of “Yes” which is a contrary with segment 5.

Variable 2 internships (dark green =0-0.25, dark brown = 0.75-1,dark purple = 1.75-2), segment 1 are having a high proportion of lower ranges (dark green and dark brown) indicating fewer and less internship similar to segment 4 and segment 5, while segment 2 also shows the similar results but slightly different on dark purple, indicating more internship. And segment 3 is heavily concentrated in the higher ranges (dark purple) which imply more internships in this segment.

In variable 3 placement training, segment 1 and segment 2 is entirely “No” showing a zero partition in placement training while segments 4 and 5 have an entire “Yes” showing a high partition in placement training. Lastly segment 2 also mostly “Yes” but a slightly “No” indicting individual in this group likely to have more participate in placement training.

In variable 4 projects (0-0.25 ,0.75-1.25, 1.875-2.25, 2.625-3, in segment 1 moderate distribution across lower to mid-range, indicate that individuals are likely to have one or more projects in this segment.

Comes to variable 5 soft skills rating, segment 1: shows a spread across mid-to-high ratings around (3-4.35) which are indicate a high soft skill rating. While segment 2, the light purple (4.575-4.8) is occupied a higher proportion which shows a higher soft skills rating compared to segment 1, like segment 4. And segment 3 are having the highest proportion of high soft skills rating within this segmentation, shows that a highest consistently soft skill. Lastly segment 5 have a great mix of each categorized, is consider stronger than segment 1 but weaker than segment 2 and segment 4.

Lastly variable 6 which is workshops certification. Segment 1 and segment 5 are with notable a proportion in lower and mid-ranges, indicating that low participation in workshops/certification. While segment 2 also a large amount of lower range, but a strong concentration in higher range than segment 1 this indicating one or more workshop/certifications like segment 4 and 5. Lastly segment 3 highly concentrated at the highest range, suggesting that a high workshops/certification.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 82: 5 Cluster mean statistics

Segment 3 are the highly accomplished with 28.72% of data with demonstrates high mean values for Internships (1.92), projects (2.65). soft skills rating (4.63) and workshop/certification (1.89) also shows a strong participation in placement training and extracurricular, indicate most cohesive cluster. This segment represents the top-tier individuals who are highly engage across all the activities and training and possess a excellent soft skills and practical experience, followed by segment 4 and segment 5 also not excellent like segment 3 but also moderately engaged with good individuals’ profile.

A screenshot of a pie chart

AI-generated content may be incorrect.

Figure 83: 7 Cluster

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 84: 7 Cluster segment plot

Variable 1 extracurricular activities, segment 1, 3 , 6 and 7 shows a very strong participation in activities while segment 2, 4 and 5 shows a less participation but segment 5 shows a slightly proportion of “Yes” indicate that individuals in this segment is moderate low in participation.

Variable 2 internship has been categorized into 7 segment while segment 1,2,4,5,6,7 are high proportion in the lower ranges indicating fewer internship. However, segment 3 heavily concentrated in high ranges, indicating individuals in this segment having one or two internship.

Variable 3 placement training also similar to variable 2, segment 1,2 ,3,6 and 7 high proportion of “Yes” indicating a high participation in placement training and segment 4 and 5 almost entirely “No” indicate low participation in placement training.

Variable 4 projects segment 1 and 3 heavily concentrated at the higher range, suggesting very high or more projects while segments 4, 5 and 7 are more proportion on the lower range indicating less project. And segment 2 a have a great mix indicating zero or one project.

Variable 5 soft skills rating segments 1, 3 and 6 have a strong concentration in the higher soft skill rating ranges, indicating consistently high soft skills while the other segment have a more proportion on lower range which shows that a lower soft skills rating.

Variable workshops/certification shows that segment 1 and segment 3 individuals are more likely to have one or more certification but segment 2,4,5,7 are suggesting very few or no workshops or certifications.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 85: 7 Cluster mean statistics

In this mean statistics segment 3 are shows a highly accomplished, it consistently demonstrates high mean values for internship (1.92), projects (2.65), soft skills rating (4.63), and Workshops/certification (1.89) and a strong participation in placement training and extracurricular activities, it is most cohesive cluster followed by segment 4 and 5, and other are low engagement.

A screenshot of a pie chart

AI-generated content may be incorrect.

Figure 86: 3 Cluster

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 87: 3 Cluster segment plot

Segment 1 in extracurricular activities’s proportion are mostly “Yes” indicating very active and a slightly on “No”. Comes to segment 2, it only a little of “No” and mostly entire “Yes” means that these two segment’s individuals are active in participate extracurricular. While the segment 3 is mostly “No” with a slightly of “yes” individuals within this segment are more likely not to participate in extracurricular.

Variable internships with segment 1 shows that a mixed, but mostly mid-range values meaning that individuals in this group are having one or more internships. Different from segment 2 and segment 3, both of them are heavily clustered in mid-ranges and low range, might indicating that individuals in this group have one or less internship/

Come to variable 3, segment 1 and 2 in placement training are having a high participation means that individuals within this group are more likely to have a placement training, and for segment 3 it shows a evenly split, individuals in this group might accept for placement training and might not.

Segment in projects also shows a mostly mid to high project count, while segment two shows a little of lower range (light blue) not as high as segment 1. And segment 3 are dominated by lower project counts.

Soft skills rating in segment 1 are very high rating dominate and segment 2 are spread across mid to high levels indicating that individual are tend to have a high soft skills rating in segment 1 and 2 but segment 2 is not high as segment 1 and segment 3 includes some low range rating.

Lastly, segment 1 in workshops/certification are majority in high certification while segment 2 possess a medium-high range in participate of workshops/certification. And segment 3 are mostly mid-range meaning they are less participation.

To conclude, segment 1 individuals like includes the most prepared students with high on soft skills, training and activities. They’re likely top candidates for placement. While segment 2 are still actively working on skills, placement training and workshops/certification, but they might lack project experience compared to segment 1 but strongly focused on employability. Lastly segment 3 individuals are more focus on academics and less holistic development such as extracurricular activities, placement training, and soft skills, and might need some support for placement preparation.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 88: 3 Cluster mean statistics

In these mean statistics, we also can see that segment 1 performs better with a highly engaged, well-rounded, and accomplished group with an extracurricular activities (Yes:93.50) and Placement Training (Yes:89.77) also shows a high mean values for internships (2.14) projects (1.64) soft skills rating (4.62) and workshops/certification (1.84), while segment 2 is characterized by a focus on formal placement training, with moderate practical experience but less extracurricular. Lastly, segment 3 individuals possess a minimal engagement across all placement training activities and all development.

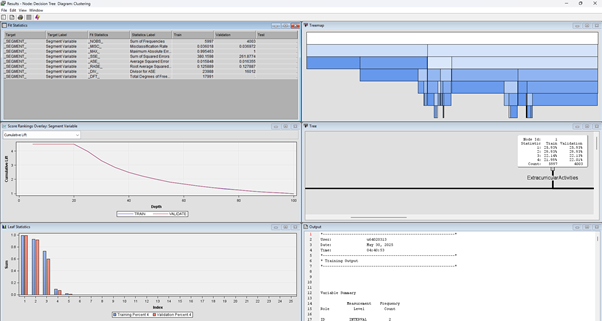


Figure 89: Decision Tree

This decision tree is used to provide a set of simple, rules-based conditions that can be easily understood acted upon by non-technical stakeholders. For example, a rule might be “If Internships < X AND soft skills training < Y, then segment = 3”. This directly translates the complex clustering results into actionable business logic. It also helps to understand how to identify future member of those clusters based on a clear decision path. It provides a “classification engine” for segments.

## **7.2 Model Validation**

A screenshot of a computer

AI-generated content may be incorrect.

Figure 90: Model Comparision table

This table is to evaluate the performance to determine their error rates and how the cluster will perform.

|  |  |
| --- | --- |
| Column Name | Explanation |
| Misclassification Rate | This is to select the best model, the lower percentage is typically chosen because it indicates the best performance on unseen data. |
| Train: Misclassification | This is the percentage of observations in the training dataset that the model classified incorrectly, the lower indicating the better performance |

Tree 4 are being selected as the best performing model based on its performance on validation set with a Train Misclassification Rate of 0.0310 and Valid Misclassification Rate of 0.0295. This is considered as very low and suggest that tree 4 is highly accurate at predicting segment membership on unseen data; also means that the model is not overfitting the training data as it performs as well on data it has never seen before as it does on the data it learned from. While compared to other tree models, can see that tree, tree 2 and tree 3 all have higher valid misclassification rates (0.0299, 0.0385, 0.0695), this furthers shows that tree 4 is indeed has the lowest error on the critical validation set and the very small differences between training and validation misclassification rates.

Since tree 4 is the lowest accuracy model, then later will proceed with segment profile to determine how good it is.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 91: Cumulative lift

In this cumulative lift, we can see that tree 4 shows the highest cumulative lift across most of the “Depth” range, especially at lower depths (e.g. up to around 20-30%). In this case meaning that it is more effective at identifying the target segment within that top percentage of the population. While the other trees might cross path or have similar lift at very deep levels, Tree 4 consistently demonstrates a stronger ability to concentrate the positive responses at the top of its predictions. It would yield a substantially greater concentration of the desired outcome than the other decision trees. This superior performance at the initial, high-impact segments of the population makes Tree 4 the most effective and desirable model for this particular task, as it maximizes the identification of the target segment with fewer resources or efforts.” Tree 4 also has the lowest misclassification rate of 0.02995 which is lower than tree 3, tree 2 and tree, means fewer incorrect predictions and is typically performs well across squared errors and root average squared error supporting its overall predictive accuracy.

## **7.3 Critical Interpretation of Outcomes**

A screenshot of a graph

AI-generated content may be incorrect.

Figure 92: Segment Profile

This is the segment profile of 3 clusters, In first row which is segment 3, it clearly shows most individuals in this segment are less participate in extracurricular activities which is also weaker than full dataset. Come to the project histogram, it is nearly left right-skewed most have fewer project and few have many, indicating they are weaker in project exposure too. Comparing the soft skills rating with full dataset it is slightly right-skewed also as mostly moderate soft skills, but fewer high rating compared to dataset. While workshops/certification also perform the same way with right-skewed and individuals took part in certification but with a slight lean toward fewer. The distribution of placement training is fairly balance with a mixed participation in placement training; not dominant and still weaker than full dataset. Lastly, internship distribution are most in middle bar, is slightly right-skewed indicating that most individuals with 0 or 1 internship very few with more in this segment. To conclude this group under performs on most attributes vs. the overall dataset.

Followed by segment 1, its workshops/certifications are mostly above average and is left-skewed many students completed at least one certification in this segment some have 2. For the internship the tall bar is on right side, so it is left skewness most student have 2 internships some have 1. Comes to the project’s histogram it is left-skewed and flat then tall bar right, majority completed many projects, segment excels here. Same goes to soft skills rating, its peak is on the right, so it is left-skewed indicating individuals possess strong soft skills most rating is high better than dataset average. In extracurricular activities and placement training pie chart, it is mostly “Yes” segment has high placement training exposure also tend to participate in activities and is average above than the dataset. To conclude this is the best-performing cluster, above average in nearly all areas.

In Segment 2, extracurricular is performing very well, almost entire “Yes” shows a very high participation in activities and average above when compared to dataset. While internships are peak in middle and slight right skewed, many of them have 1 internship some have 0. Distributions mirrors segment 1. And for workshop/certificate, high bars left and is right-skewed majority of individuals completed few certifications segments is weaker here. While soft skills rating is right peak and left-skewed indicating strong soft skills and ratings concentrated at high end. To conclude, segment 2 focuses more on activities and soft skills, with average of workshops/certificates and less internship.

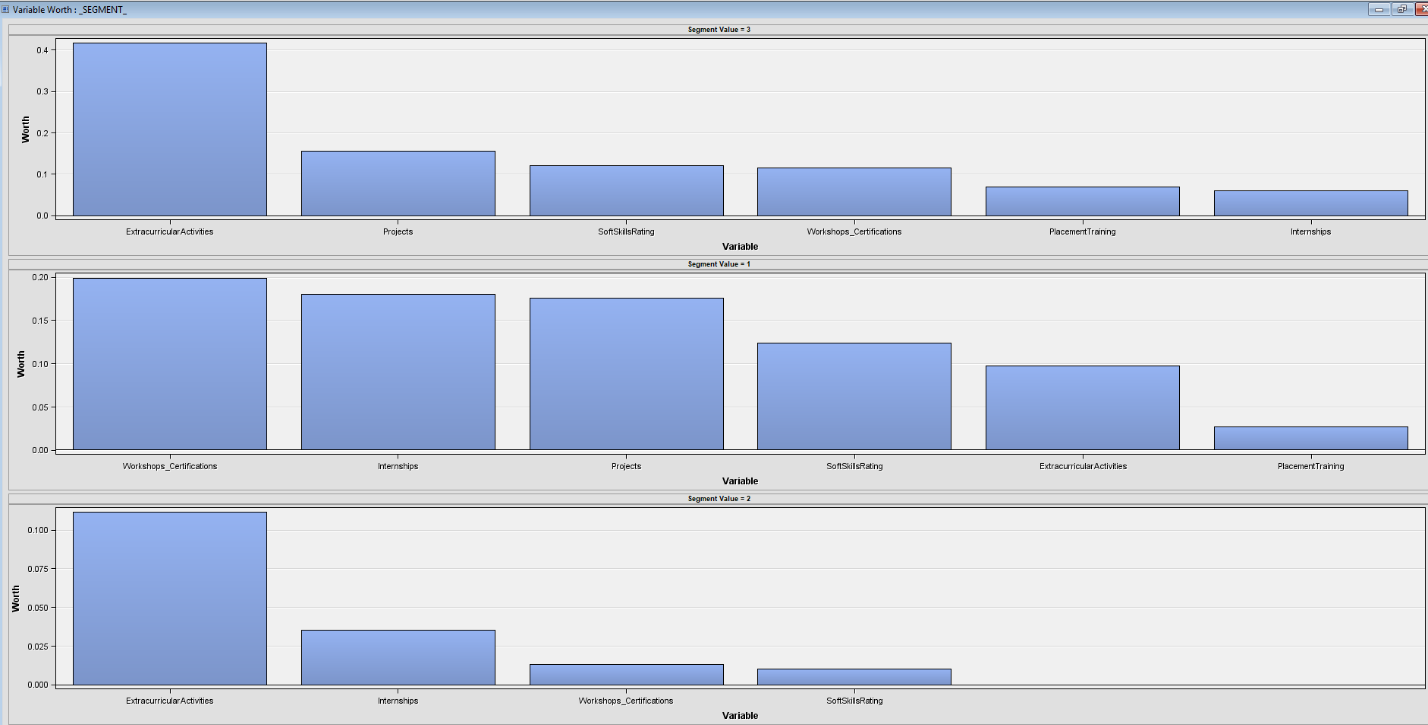


Figure 93: Segment profile variable worth

This bar charts each representing the “Variable Worth” for a different “Segment Value” (3, 1, and 2). These

Segment value = 3 (top chart) the highest worth variable is extracurricular activities suggesting it’s the most defining characteristic of this segment and project, soft skills rating and workshop certification have moderate and roughly similar worth. The lowest worth variable placement training and internships also indicating that they are less influential in characterizing segment 3.

Segment value = 1 (Middle Chart) shows the workshops/certification, internship and project are the highest and relatively similar levels of worth, while soft skills rating and extracurricular have moderate worth, lastly placement training is suggesting the least important variable as it has the lowest worth.

Come to the last segment value = 2 (bottom chart), the extracurricular activities shows again the highest worth, which is similar to segment 3, and variable such as internships, workshops certifications and soft skills rating have considerably lower and somewhat similar worth compared to extracurricular activities. Notably, projects and placement training are not visible or have negligible worth in this segment’s chart.

# **8.0 Discussion and Conclusion**

## **8.1 Model Comparison**

A screenshot of a diagram

AI-generated content may be incorrect.

Figure 94: Model comparison with Ensemble

In the diagram, an ensemble model was created using the best regression model and the best gradient boosting model. Then, all 3 models were compared to see which performs better.

### **8.1.1 ROC Curve**

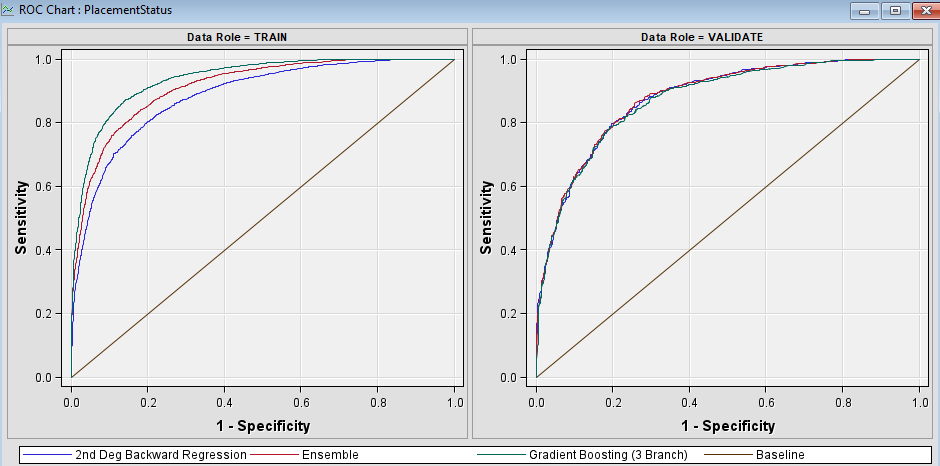


Figure 95: ROC Curve

The Receiver Operating Characteristic (ROC) chart is a graphical representation used to evaluate the performance of a classification model across all possible threshold values. It plots sensitivity (True Positive Rate) against 1-specificity (False Positive Rate), providing insight into the trade-off between correctly identifying positive cases and avoiding false alarms.

In the training set, the gradient boosting (green line) shows the steepest and highest curve, indicating the best sensitivity and specificity, while ensemble model (red line) is close and slightly below boosting. Lastly, regression (blue line) performs the lowest with a less sharp curve.

In the validation set. All models show a very close performance. But again, gradient boosting is slightly superior, and the regression model remains slightly under the others, suggesting it’s less robust.

### **8.1.2 Cumulative Lift**

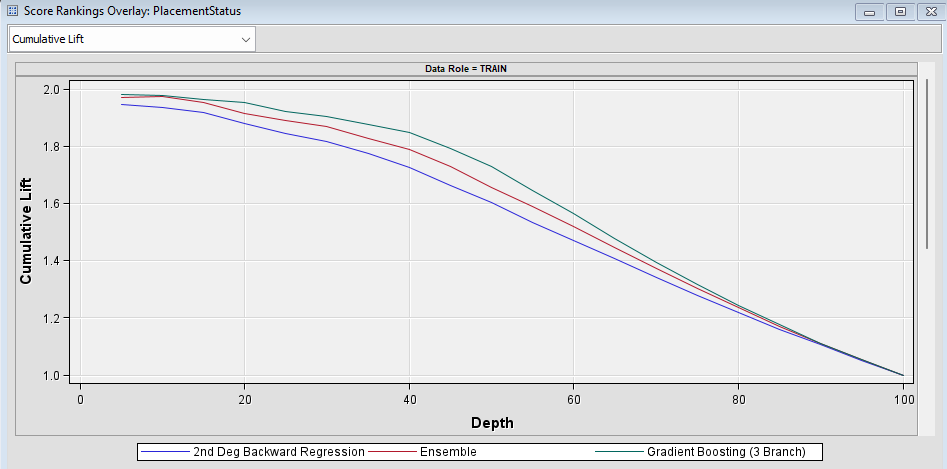


Figure 96: Cumulative Lift

Cumulative Lift is a metric used to evaluate the effectiveness of a classification or predictive model, particularly in marketing and risk modeling. It measures how much better the model performs at identifying positive outcomes compared to random selection. From the graph, it can be seen that Gradient Boosting starts with the highest (at depth 0-30) and is followed by the ensemble model which is slightly lower. Regression still consistently underperforms others.

### **8.1.3 Fit Statistics**

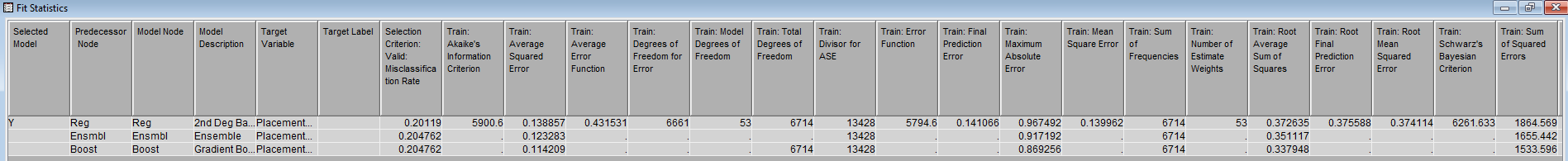


Figure 97: Fit Statistics

In these fit statistics, regressions show the lowest misclassification, which is 0.20119, meaning that it is best for prediction. To sum up, 2nd degree backward regression is simple, interpretable with low misclassification rate, however it underperforms in ranking (ROC, lift) indicating a weaker generalization. As for gradient boosting, it shows a high accuracy with the best lift and ROC performance, but it has a slightly higher misclassification due to it being more complex. Lastly, ensemble combines both model benefits, thus reducing overfitting risk and robustness, although it is slightly worse than boosting alone but averages strengths.

## **8.2 Conclusion**

To conclude, a comprehensive analysis was conducted on the student placement dataset, which provided valuable insights into the key factors that affect student placement outcomes after graduation. By employing the SEMMA methodology, the study followed a structured approach that enhanced the clarity and effectiveness of the data analysis process. Besides that, predictive modelling efforts were also included with the aim of developing robust models that can predict student placement outcomes accurately. The findings from these models can help serve as a foundation for educational institutions to identify students at risk of not being placed and develop interventions at an early stage.

The dataset we chose was quite clean, as there were no missing values, outliers, nor inconsistencies in any of the columns. Although the distribution for each variable was slightly skewed, they were still within the acceptable range, so there was no need to transform the variables. However, there was a slight imbalance in the target class between “Placed” and “Not Placed” students, so a sampling method was used to address that by ensuring an equal representation of both classes. This will help improve the reliability of the predictive modelling process.

The following models were developed in this study: decision tree, gradient boosting, random forest (HP forest), high performance decision tree (HP Tree), regression (with/without variable selection), HP regression, and an ensemble model that combined the best tree and best regression. Out of all the models, the 2nd degree polynomial backward elimination regression without variable selection was the model that produced the highest accuracy of 79.9% (although it has lower ROC and lift), followed by the ensemble model and the gradient boosting model with a shared accuracy of 79.5%. These findings suggest that the selected regression model offers superior prediction accuracy for student placement outcomes. Besides that, the regression model also showed that Placement Training and Extracurricular Activities were among the most influential factors towards student’s placement status, suggesting that students should actively participate in more training and after school activities to increase their chances of employment after graduation.

In addition, clustering analysis was also conducted to identify underlying patterns within the dataset. The clustering analysis offered an unsupervised approach to group students based on similarities in their academic and behavioural profiles without using the target variable placement status. Based on the 3 clustering results, it was observed that high-performing students who participated in placement training sessions, internships and workshops/certification formed distinct clusters with high placement rates. In contrast, students with average academic records and minimal participation in extracurricular or receive a low soft skills rating were clustered together with moderate to low placement outcome. This analysis highlighted that natural segment within the student population and served as a strategic tool to identify target groups for intervention, clustering is valuable for understanding patterns and designing supports programs tailored to different student profiles.

# **9.0 References**

Bette Loef, A. W.-C. (2022, June 20). *Using random forest to identify longitudinal predictors of health in a 30-year cohort study*. Retrieved from Scientific Reports: https://www.nature.com/articles/s41598-022-14632-w

Constantin Aliferis, G. S. (2024, March 5). *Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI*. Retrieved from Springer Nature Link: https://link.springer.com/chapter/10.1007/978-3-031-39355-6\_10#citeas

Maria D. Guillen, J. A. (2023, March 13). *Gradient tree boosting and the estimation of production frontiers*. Retrieved from Science Direct: https://www.sciencedirect.com/science/article/pii/S0957417422021522

Mohammad Arifuzzaman, M. R. (2023, February). *An Advanced Decision Tree-Based Deep Neural Network in Nonlinear Data Classification*. Retrieved from Research Gate: https://www.researchgate.net/publication/367979886\_An\_Advanced\_Decision\_Tree-Based\_Deep\_Neural\_Network\_in\_Nonlinear\_Data\_Classification

Vasiliki Matzavela, E. A. (2021, October 13). *Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments*. Retrieved from Scinece Direct: https://www.sciencedirect.com/science/article/pii/S2666920X21000291

Yen-Liang Chen, C. H. (2022, June). *An ensemble model for link prediction based on graph embedding*. Retrieved from Scinece Direct: https://www.sciencedirect.com/science/article/abs/pii/S0167923622000240

Dodge, Y. (2008). Regression Analysis. *The Concise Encyclopedia of Statistics*, 450–452. <https://doi.org/10.1007/978-0-387-32833-1_348>

Gautam, P., Singh, Y. P., Gautam, P., & Shaikh, P. (2017). ADVANCED DATA MINING: A SCIENTIFIC APPROACH. In *International Journal For Technological Research In Engineering ISSN (Online* (pp. 2347–4718). ETEBMS. <https://ijtre.com/images/scripts/16169.pdf>

Haque, R., Quek, A., Ting, C.-Y., Goh, H.-N., & Hasan, M. R. (2024). Classification Techniques Using Machine Learning for Graduate Student Employability Predictions. *International Journal on Advanced Science Engineering and Information Technology*, *14*(1), 45–56. <https://doi.org/10.18517/ijaseit.14.1.19549>

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer.

Ruparel, M., & Swaminarayan, P. (2024). Student Placement Prediction Using Various Machine Learning Techniques. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3), 2107–2113. <https://www.researchgate.net/publication/386448173_Student_Placement_Prediction_Using_Various_Machine_Learning_Techniques>

Ruparel, M., & Swaminarayan, P. (2025). Enhanced Student Placement Prediction Using Machine Learning: A Comparative Evaluation of Algorithms. *International Journal of Engineering Trends and Technology*, *73*(1), 225–236. <https://doi.org/10.14445/22315381/ijett-v73i1p119>

Swami, S., Gupta, S., Kumar, R., Raj, S., Shrivastava, V., & Bandil, D. K. (2025). ML-Based: Placement Prediction Application. *International Journal of Emerging Science and Engineering (IJESE)* , *13*(6). <https://doi.org/10.2139/ssrn.5246128>