College Placement Analysis Report

Madhav Khanal

March 4, 2025

1 Introduction

The dataset is an open-source dataset from Kaggle that contains students' academic training and college placement status [1]. There are 9 features for the dataset: CGPA, Internships, Projects, Workshops/Certifications, Aptitude Test Scores, Training with Soft Skills, Extra Curricular Activities, Placement Training and Senior Secondary and Higher Secondary School Marks. The final column is a target column with a boolean whether the student was placed or not. Each of the columns can be summarized as follows:

- CGPA - It is the overall grades achieved by the student.
- Internships - It tells the number of internships a student has done.
- Projects - Number of projects a student has done.
- Workshops/Certifications - As there are multiple courses available online students opt for them to upskill themselves.
- ApptitudeTestScore - Aptitude tests are generally a part of the recruitment process to understand the quantitative and logical thinking of the student. This is on a scale of 1-100.
- SoftSkillrating - Communication is a key role that plays in the placement or in any aspect of life. This is on a scale of 1-5
- Extracurricular Activities - This helps provide insight into the personality of an individual regarding how much he is active other than academics.
- Placement Training - This training is provided to students in college so they can ace the placement process.
- SSC and HSC - Senior Secondary and Higher Secondary Marks. This is on a scale of 1-100.
- PlacementStatus - This is our target column with two outputs: placed and not placed.

The dataset can be categorized as:

- Numeric: CGPA, Internships, Projects, Workshops, Aptitude Test Score, Soft Skill Rating, SSC and HSC
- Boolean/Binary: ExtraCurriculars Placement Training, Placement Status

Overall, the dataset is a binary classification problem that depends on 9 features given above. Each feature might have varying levels of impact on the college placement.

1.1 Relevance

This dataset has a lot of room for basic data analytics and visualization that ties back to the data analytics class, for instance plotting relations between different columns using ggplot, dplyr for manipulating rows and columns, cleaning the data before processing and so on.

1.2 Challenges

- The dataset is huge with 10000 rows and 10 columns. Planning what's important and how to handle might be challenging
- There might be missing columns and we need to either remove them or replace the values appropriately
- The boolean values are the strings. We might need to convert them into a binary 0/1 for further analysis.
- We need to clean the data for missing rows, inconsistent values and so on before processing

2 Individual Report

2.1 Hypothesis

- H01: The CGPA Correlates with aptitude test score and can enhance college placement
- HO2: Higher Aptitude Score implies better college placement
- HO4: GPA indicates college placement success
- HO5: Placement training influences placement outcome
- How do numeric variables relate to each other?.
- Which are the top 5 most important features for college admissions process?

The above questions are significant because often when students apply to colleges, they and their families want to know what factors are the most important and what factor to focus on the most.

2.2 Analysis Methodology

Since there are so many columns, my goal was to find the most important. I first used the heatmap to find the highest correlations and analyzed them. I backed it up by common sense as well–for instance, a correlation is probably high between placement status and high gpa and so on. For each of the above steps, I did the following:

- I created a scatter plot of CGPA with aptitude test scores coloured by placement status. The placed people in the upper right half (majority) indicate a good correlation (visual).
- To test whether a higher aptitude score means better college placement, i created boxplots with two boxes for each placed and not placed and saw where in the boxes these values lie.
- To test whether GPA is actually important, i made box plots for both placed and non-placed in the same plot and saw where the GPA percentiles of these two categories fall into.
- I made a bar plot of the proportion of students against whether they received training or not
- Calculated the mean decrease gini(how much removing each variable causes the loss in accuracy) for each column feature and plotted the bar chart with descending order.

2.3 Results & Insights

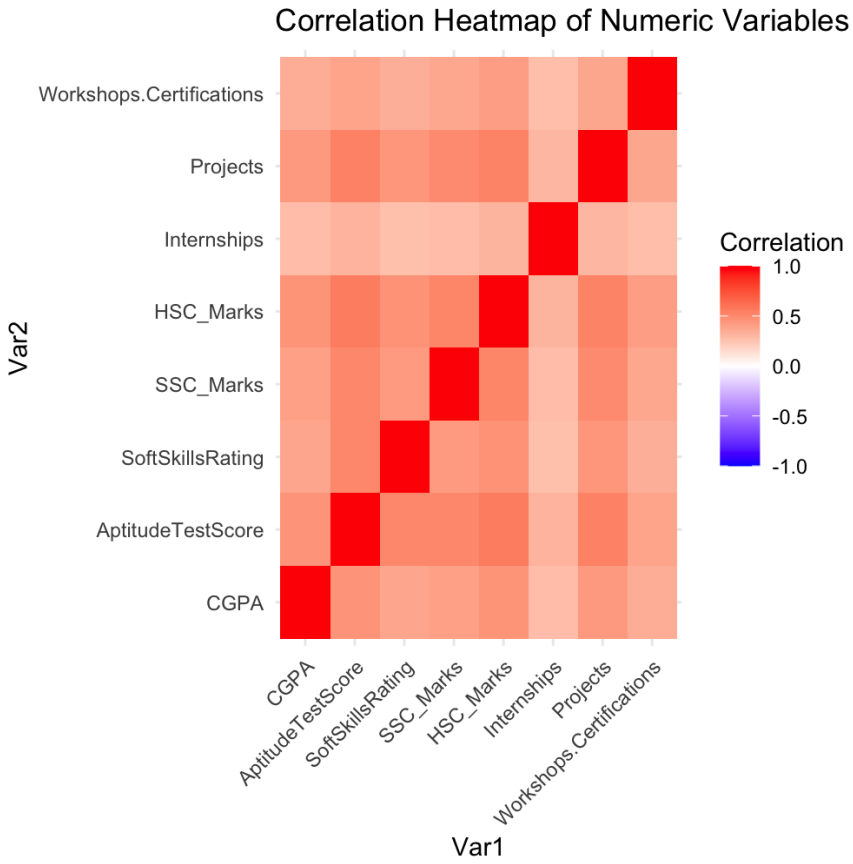

Figure 1: The HeatMap

The Heatmap patterns show several key insights:

- Internships are the least related to any other columns (transparent row and column for internship field). This makes sense as none of the internships mightn't be that much impacted by CGPA, Aptitude Score, and so on.
- None of these fields have negative or zero correlation as all of the students are high school students applying to college and many of these components together.
- There's a good correlation between CGPA and Aptitude Test Scores which implies that intelligent students have good GPAs.
- There's a good correlation between aptitude test scores and the number of projects done, marking that intelligent students are explorative and do projects.

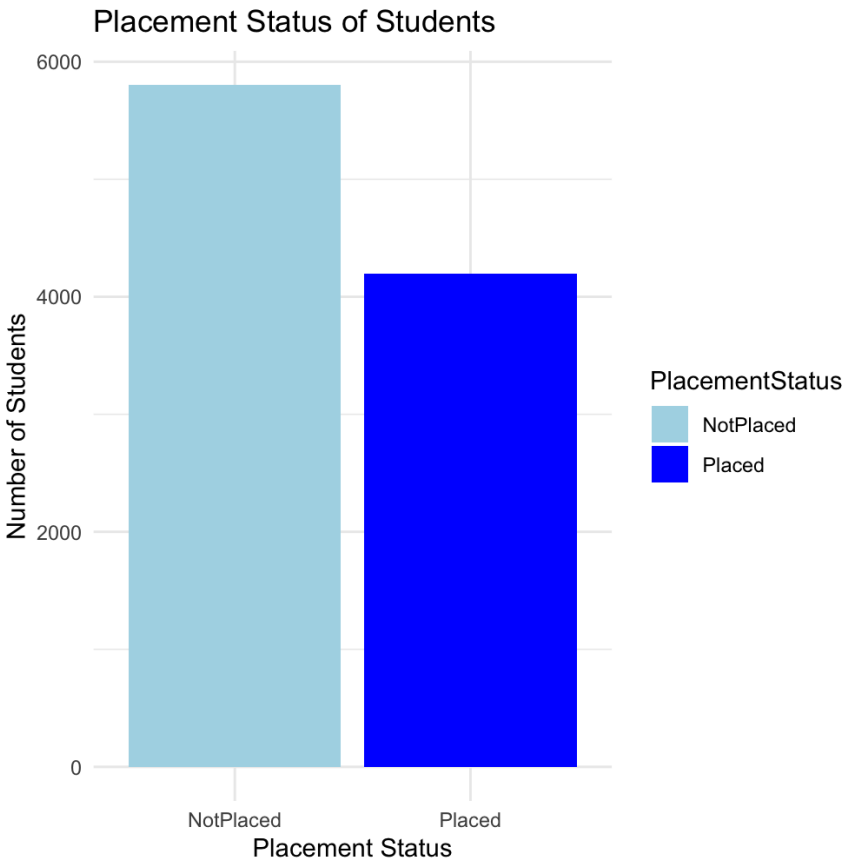

Figure 2: Bar Plot of Placement Frequency

The basic bar plot shows that the majority of students didn't receive college placement, highlighting the competitive nature of college admissions processes.

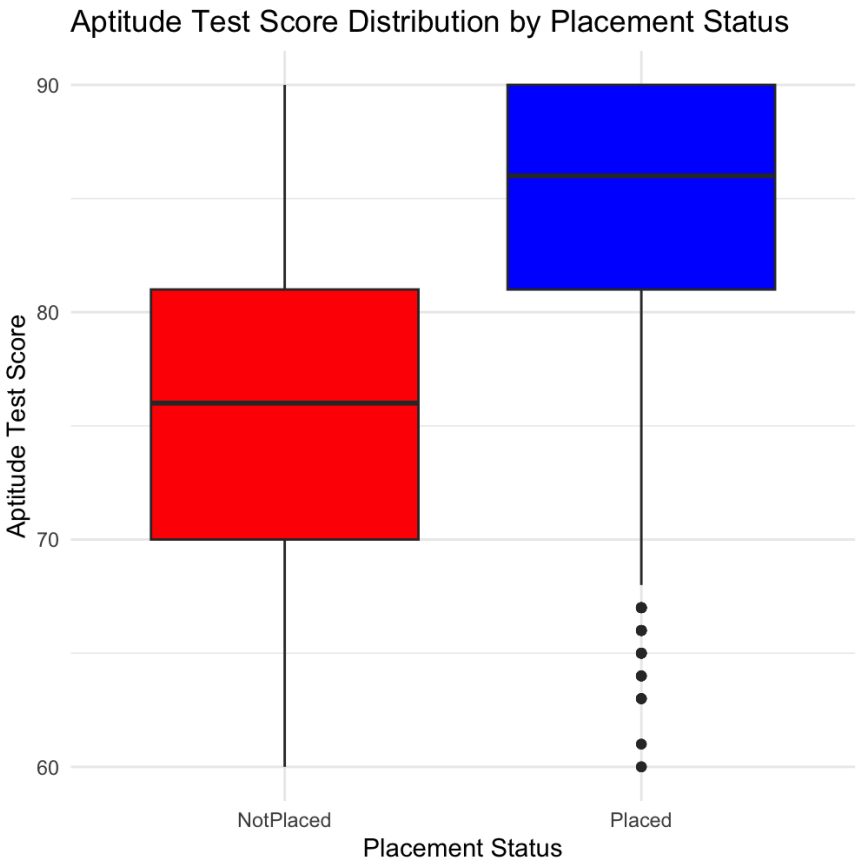

Figure 3: Box plot of Aptitude Test by Placement

Key observations from the aptitude score boxplot:

- The median and all percentiles are higher in terms of aptitude scores for students who received a placement.
- This concludes that intelligent students were more likely to receive college placements.
- There's a section with outlier values where students with less aptitude scores were placed into colleges.
- This supports that aptitude scores aren't the sole indicators, accounting for the fact that these test scores might be flawed, incorrectly measured, or that aptitude isn't the sole deciding factor in the holistic college admissions process.

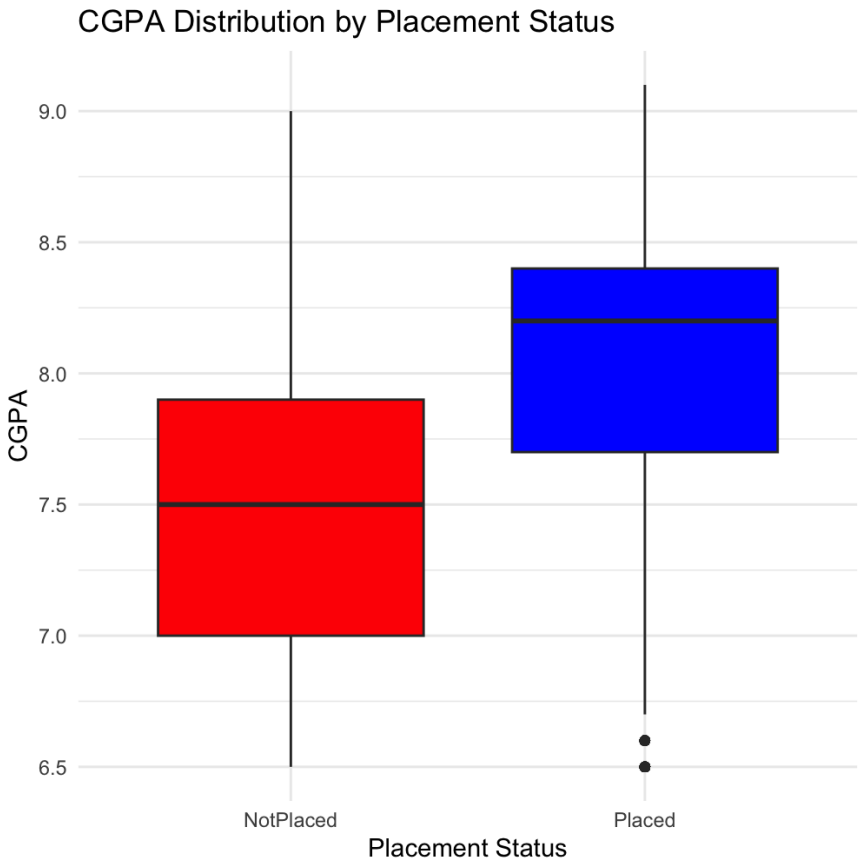

Figure 4: Box plot of Cumulative GPA by Placement

The GPA boxplot reveals that:

- The GPA median and all percentiles are higher for students who were placed compared to the ones who didn't.
- This conclusively shows that GPA is an important factor in the college admissions process.

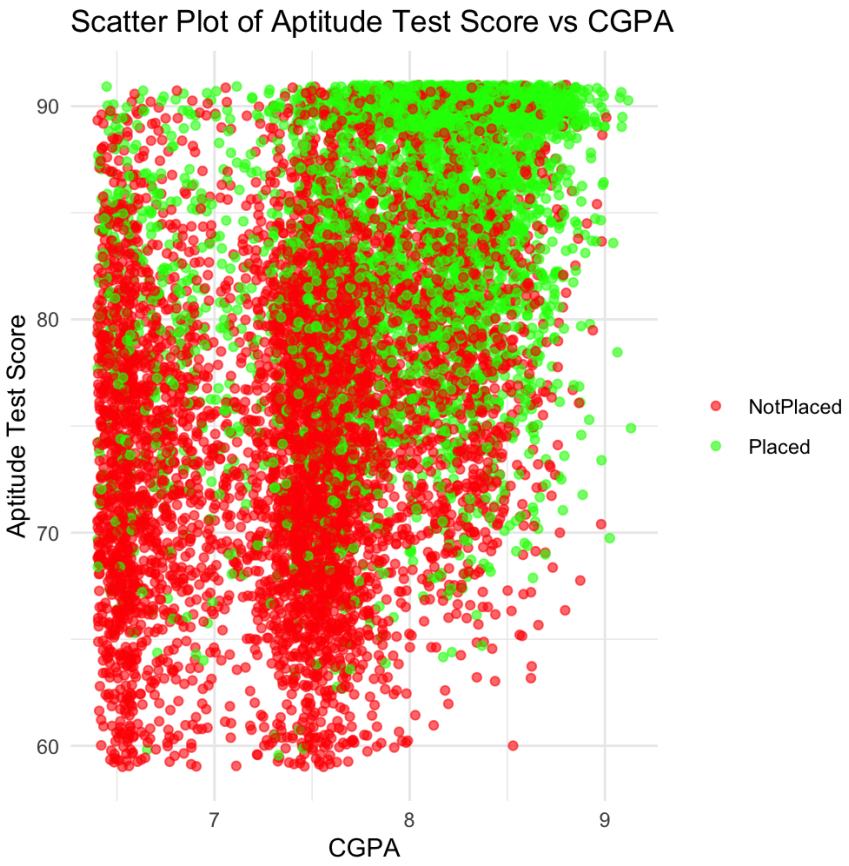

Figure 5: Scatter plot of Aptitude Test Score vs CGPA by Placement

Insights from the scatter plot:

- There is a trend that higher aptitude test scores align with higher CGPA.
- The upper right section of the graph is mostly green, concluding that higher GPA and higher aptitude scores are correlated and can enhance college placement.

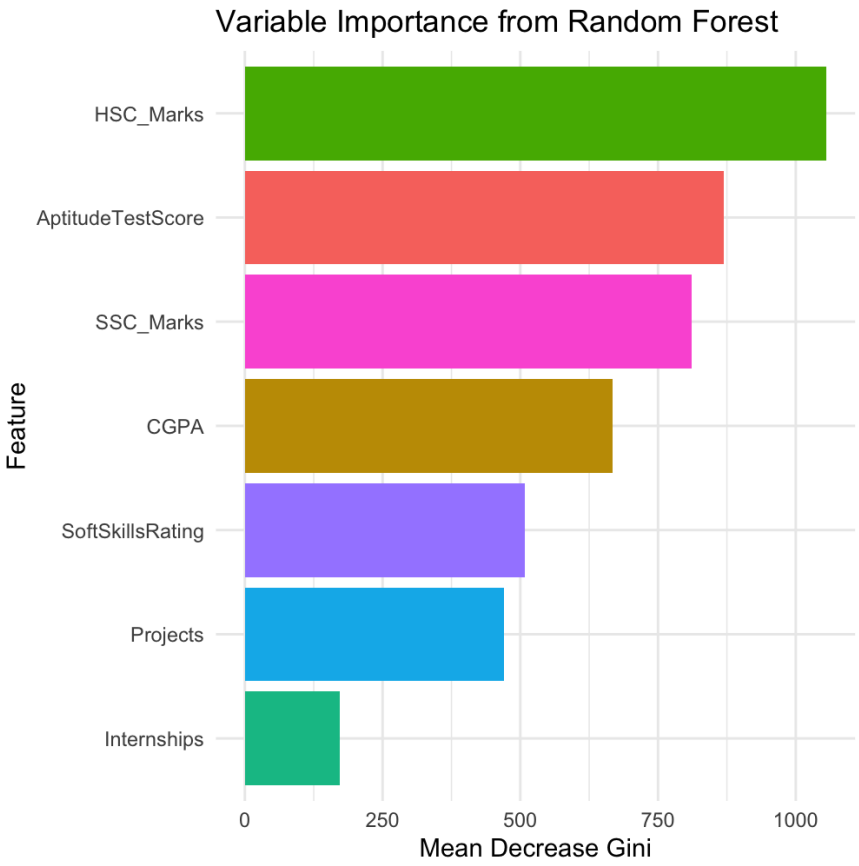

Figure 6: Bar Plot of Mean Decrease Gini

The Mean Decrease Gini plot highlights that:

- There is a high mean decrease gini with the variables HSC marks, aptitude test score, and SSC marks.
- This concludes that academics is the most important feature contributing to the college placement of students.

2.4 Analysis with DeepLearning (Optional)

In this approach, the binary columns are converted to 0 and 1 and the final target of placement is also converted into a binary number. The student ID column is dropped. This is basically a binary classification problem where 9 features determine the college admissions process.

We build a feedforward neural network with 5 layers including the dense and dropout using TensorFlow, splitting the data into 80/20 train tests. Since there are no empty columns (from the previous analysis) we don't need to worry about that. The final layer is a binary output (placed/not placed) from the sigmoid function. The learning rate is adjusted to 0.001 for a slower gradient descent based on the experimentation with different rates (the higher rates led to the model randomly jumping here and there in the learning of the weighting

process). The model is trained with 2800 parameters, and to prevent overfitting, we have used the dropout layer as well as decreased the epoch cycles. For instance, when we trained over 500 epochs, while the model's accuracy increased over the training data, the accuracy in validation or test data started decreasing. Hence, we adjusted the epochs to an equilibrium from where the accuracy of validation started declining (or loss function started increasing).

After training, learning the weights, and predicting placement on the test data, we found an average accuracy score of around 80% which suggests the model is effective. As we can see from the confusion matrix and F1 scores, we correctly predicted the majority of people who were not placed and placed. There's a room for improvement in successful placements (the model is only 74% accurate) on predicting successful placements while it's more accurate on people who don't get placed.

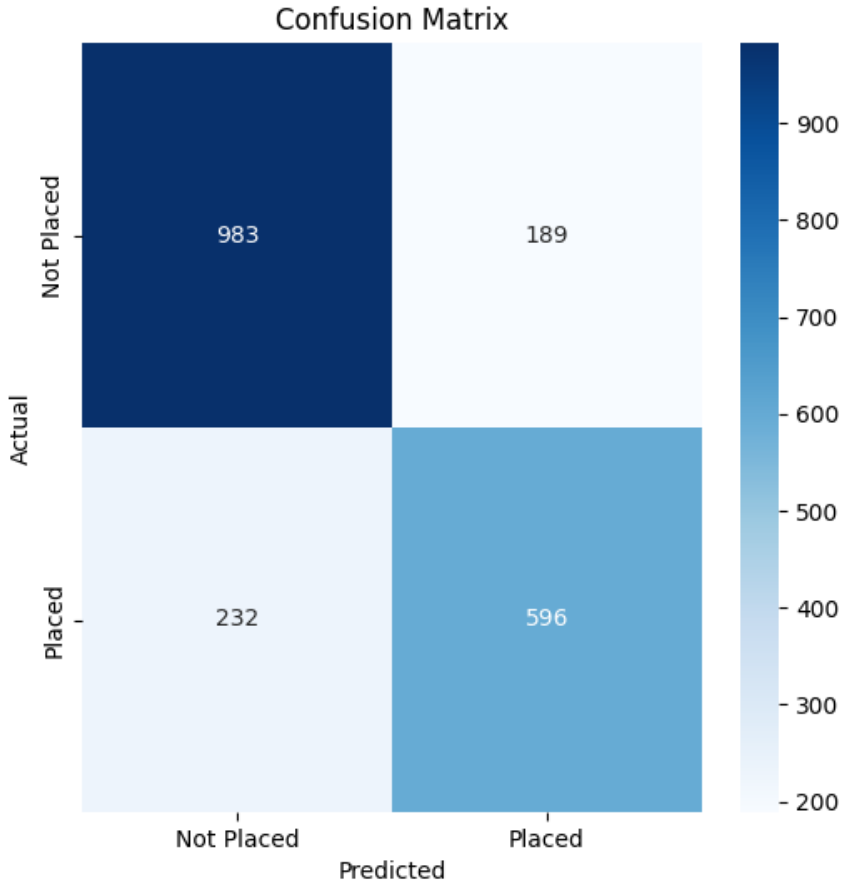

Figure 7: Deep Learning Analysis Images

2.5 Analysis with RandomForest

Using the random forest model on the same test and train data, and a max depth of 5, we obtain an accuracy score of 0.79. As with the previous model, when the model says someone

is "placed", it's only accurate 72% of the time while when it says "not placed" it's fairly accurate i.e. 85% of the time. The accuracy is slightly lower than the deep learning model, nevertheless, the models perform similarly. However, we obtained an important insight from the random forest model's feature importance score: the top five features are: the HSC Score, Aptitude Test Score, Extra extracurricular activities, Projects and SSC Marks.

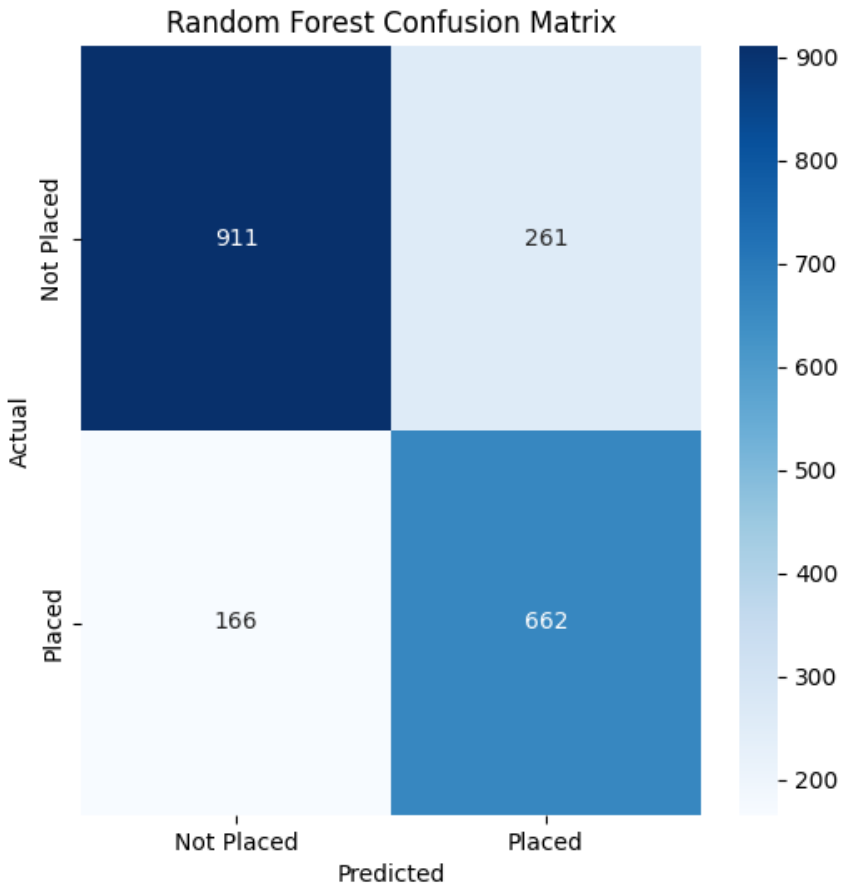

Figure 8: Random Forest Analysis

Further random forest model is run and analysis is done after identifying the most important features. For this analysis, we only took the top 5 features to see if the model improved, but the model accuracy was decreased to 0.78 and the performance remained similar. This might suggest that only these features are the most important as the model accuracy didn't change that much.

The same process is done again with the deep learning model: the model is rerun with only the top 5 features identified above. The model accuracy remains the same as before, but it does better in balancing the precision: when it says "placed" it's right 79% of the time and when it says "not placed" it's right 76% of the time, contrary to the previous deep learning and random forest models where the model was way better at predicting the "not

placed" than "placed".

2.6 Analysis with Logistic Regression and Important Insight

A simple logistic regression model with y as the binary outcome (placed/not placed) and 5 important features is defined as:

$$y = \sigma(w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + b)$$

where σ is the sigmoid function, and w_1, w_2, w_3, w_4, w_5 are the weights corresponding to the features x_1, x_2, x_3, x_4, x_5, with b as the bias. This model yields an accuracy of 78% on the test data, which is very similar to the performance of both the deep learning and random forest models.

The key insight here is that the simple logistic model only has 6 parameters while the neural network above had 2800 parameters. Our analysis concludes that a high number of parameters doesn't necessarily mean a better model. There's always a chance of overfitting with the high number of parameters.

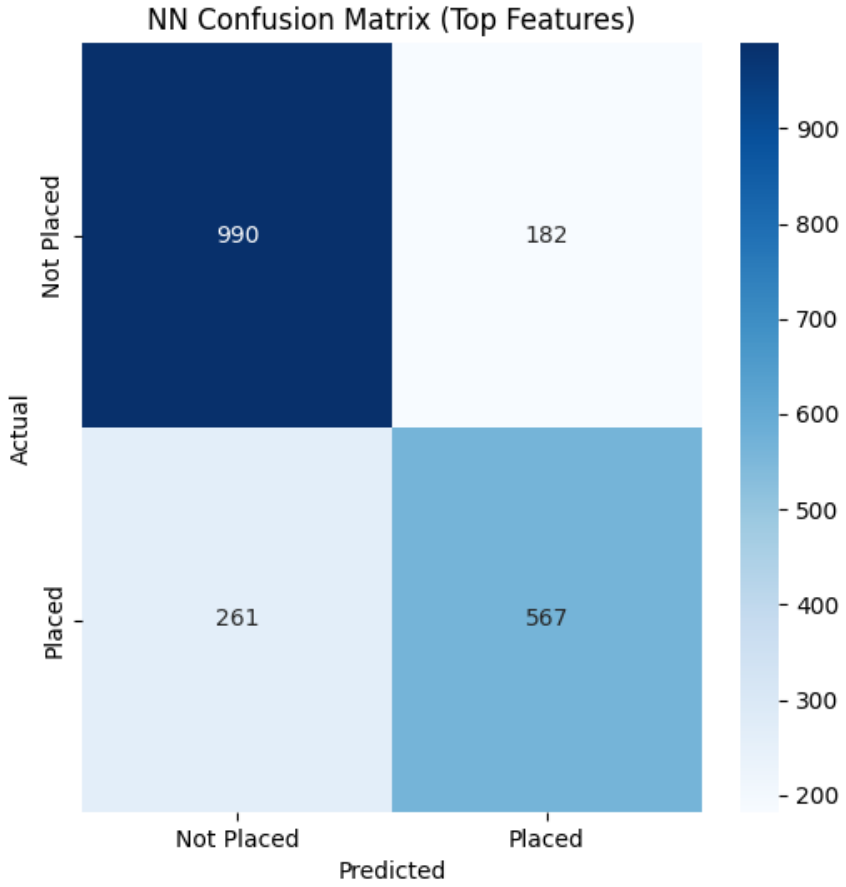

Figure 9: Regression Analysis

3 Conclusion

In the above dataset, we drew important insights into the college admission stats data of 10000 students through various data analytics techniques. We first cleared the data to see if the data had missing entries. Then we started visualizing the dataset to see the relations between different columns and how they impact the college placement. We came across an important conclusion that marks in exams and aptitude scores are very high indicators of college placements. Furthermore, we dived into various models like neural networks and random forests to see how they perform in predicting the college placements of unseen data. In doing so, we found the five most important factors for college admissions: HSC Score, Aptitude Test Score, Extra extracurricular activities, Projects and SSC Marks, concluding that students should focus mainly on these for successfully getting into college. Finally, we ran the prediction task on a simple logistic regression with 6 parameters highlighting it as being as powerful as the thousands of parameter-trained neural networks, with an important insight that a model isn't necessarily better just because it has a huge number of parameters.

References

[1] Ruchika Kumbhar. Placement prediction dataset. Kaggle, 2023. Accessed: 2025-03-04.