# IBM HACK CHALLENGE 2023

APPLIED DATA SCIENCE

***IDENTIFYING PATTERNS AND TRENDS IN CAMPUS PLACEMENT DATA USING MACHINE LEARNING***

**TEAM QUADRUPLE**

ABIRAMI GURUSHANKER
DEEPTHI B
HARSHITA V
AKILAA M

(VELLORE INSTITUTE OF TECHNOLOGY CHENNAI CAMPUS)

# ***IDENTIFYING PATTERNS AND TRENDS IN CAMPUS PLACEMENT DATA USING MACHINE LEARNING***

## 1. INTRODUCTION

### 1.1. Overview

The objective of this analysis is to extract valuable insights from campus placement data using machine learning techniques, to understand the factors influencing placement success, and to develop strategies for improving the placement process.

The campus placement data comprises comprehensive information about students, including academic performance, skills, internships, and their ultimate placement outcomes. Leveraging machine learning,this solution focuses on identifying patterns and trends within this data to provide meaningful recommendations for optimizing the placement process and enhancing student employability.

### 1.2 Purpose

The purpose of this project is to leverage machine learning techniques to analyze campus placement data and extract valuable insights. These insights aim to understand the factors influencing placement success and develop strategies for improving the placement process in educational institutions. The project encompasses the following key objectives:

- *Data-Driven Decision Making*: Enable colleges and universities to make informed, data-driven decisions by providing actionable insights from the analysis of placement data.
- *Optimizing Placement Process*: Identify patterns and trends in the data to optimize the placement process, tailor strategies to match employer preferences, and enhance student placement success.
- *Enhancing Employability*: Analyze factors impacting placement outcomes to design specialized training programs and workshops, addressing specific skill gaps and thereby enhancing students' employability.
- *Personalized Student Support*: Understand the individual factors contributing to students' success and provide personalized support to increase their chances of securing placements.
- *Industry Collaboration*: Strengthen collaboration between educational institutions and industries by using insights to offer more relevant internships and job opportunities to students.

- ➤ *Enhancing Reputation*: Achieving successful placements and higher employability rates can enhance an institution's reputation, making it more attractive to both prospective students and employers.

**Where It Can Be Used:**

This project can be used in various educational institutions, including colleges and universities, as well as by academic administrators, placement coordinators, and career services teams. It can also be valuable for employers looking to understand the factors influencing the placement pool and tailor their recruitment strategies accordingly.

**What Can Be Achieved Using It:**

By implementing this project, the following outcomes and achievements can be realized:

- ○ *Informed Decision-Making*: Educational institutions can make informed decisions based on data-driven insights, ensuring that resources are allocated effectively to improve placement outcomes.
- ○ *Enhanced Student Employability*: Institutions can focus on addressing skill gaps and providing targeted support to students, leading to improved employability and increased chances of successful placements.
- ○ *Optimized Placement Strategies*: Institutions can optimize their placement strategies by tailoring them to industry preferences and the specific needs of their students, resulting in higher placement success rates.
- ○ *Stronger Industry Collaborations*: Collaboration between educational institutions and industries can be strengthened, leading to more meaningful internships and job opportunities for students.
- ○ *Improved Reputation*: Successful placements and higher employability rates can enhance the institution's reputation, attracting more prospective students and employers.

*Continuous Improvement*: The project's insights allow for continuous monitoring and refinement of placement strategies to adapt to changing requirements and industry trends.

# 2. LITERATURE SURVEY

## 2.1 Existing Problem:

The existing problem in campus placement revolves around several challenges, including understanding the factors influencing placement success, optimizing the placement process, and enhancing students' employability. Traditional methods often lack the data-driven insights needed for effective decision-making. *Existing Approaches or Methods to Solve This Problem:*

In the context of campus placement, existing approaches include:

- *Historical Data Analysis*: Institutions rely on historical placement records, but this approach often lacks predictive capabilities.
- *Skill Development Programs*: Offering skill development programs to students to improve their readiness for job placements.
- *Recruitment Drives*: Organizing job fairs and recruitment drives to connect students with potential employers.
- *Career Counseling*: Providing career counseling services to guide students in making informed choices.
- *Alumni Networks*: Leveraging alumni networks for industry connections and placement opportunities.

## 2.2 Proposed Solution

For your project, the proposed solution is a data-driven approach leveraging machine learning techniques:

*Data Collection and Preprocessing:*

Gather comprehensive campus placement data, including academic performance, skills, internships, and placement outcomes.
Pre process the data to handle missing values, outliers, and ensure data consistency.
Exploratory Data Analysis (EDA):

- Conduct EDA to gain insights into the dataset, visualize trends, and identify potential relationships between different variables.

- Perform statistical analyses, such as correlation analysis, to understand the impact of academic performance, skills, and internships on placement outcomes.

*Feature Engineering:*

Create relevant features from the available data that may significantly influence placement outcomes. For example, aggregate academic scores into a single metric or calculate a weighted score based on skill relevance to job roles.

*Machine Learning Model Selection:*
Based on the nature of the problem, select appropriate machine learning models. Since our project involves predicting placement success, models such as Logistic Regression, KNN,Model, Naïve Bayes,Decision tree classifier, random forest classifier are used and the accuracy of each model is selected.

*Model Training and Evaluation:*

- Split the dataset into training and testing sets.

- Train selected models on the training data and evaluate their performance using metrics like accuracy, precision, recall, and F1-score.

- Implement cross-validation techniques to ensure robust model evaluation.

- Identifying Factors Influencing Placement Success: Analyze feature importance scores from machine learning models to identify the key factors that influence placement success. This could include factors like academic performance, specific skills, or internships.

*Strategy Development:*

Develop data-driven strategies based on the identified influential factors. For instance, if skills are found to be critical, consider offering targeted skill development programs or curricular enhancements.

Continuous Monitoring and Refinement:

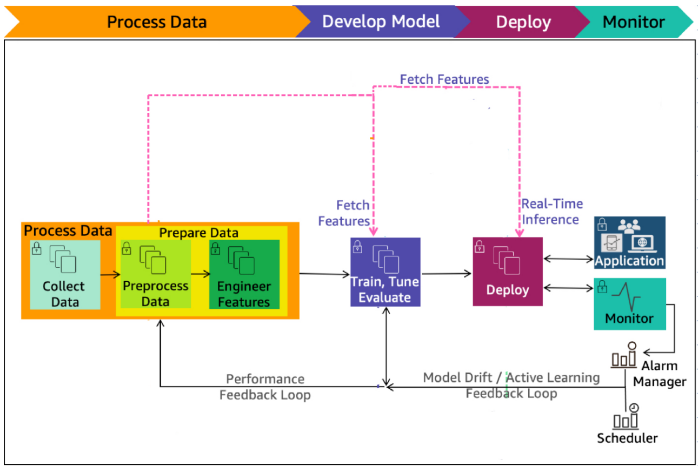
- Continually monitor placement outcomes and collect new data.

- Update and refine machine learning models and strategies to adapt to changing requirements and trends.

This specific approach aims to provide actionable insights into factors affecting campus placement success, enabling institutions to make informed decisions and optimize their placement processes. It ensures a data-driven, adaptable, and continuous improvement approach to campus placements.

## 3. THEORETICAL ANALYSIS

### 3.1 BLOCK DIAGRAM



## 3.2 HARDWARE AND SOFTWARE REQUIREMENTS

***Hardware Requirements***

- *Computer or Server*: You'll need a computer or server to run the project. The specific hardware requirements depend on the size of the dataset and the complexity of the machine learning models. For small to medium-sized datasets and models, a standard laptop or desktop computer with a modern CPU and sufficient RAM (8 GB or more) may suffice. For larger datasets and more complex models, a more powerful workstation or cloud-based resources may be necessary.
- *Storage*: Sufficient storage space to store the dataset, intermediate results, and trained machine learning models. Depending on the dataset size, you may need several gigabytes or more of free storage.
- *GPU (Optional)*: If you plan to work with deep learning models, especially large neural networks, having access to a compatible GPU (Graphics Processing Unit) can significantly speed up model training. GPUs from NVIDIA (e.g., GeForce or Tesla series) are commonly used for deep learning tasks.
- *Internet Connection*: An internet connection is required to download datasets, libraries, and packages, as well as to access online resources for research and development.

***Software Requirements***

- *Operating System*: Most machine learning libraries and tools are compatible with Windows, macOS, and Linux. Choose the operating system you are most comfortable with.
- *Python*: Python is the most commonly used programming language for machine learning and data analysis. You'll need Python 3.x installed on your system.
- *Integrated Development Environment (IDE)*: While optional, using a Python IDE can enhance your productivity. Popular choices include Jupyter Notebook, PyCharm, Visual Studio Code, and Spyder.
- *Machine Learning Libraries*: You'll need various Python libraries for machine learning, such as:
    - ◎ NumPy: For numerical computations.
    - ◎ pandas: For data manipulation and analysis.
    - ◎ scikit-learn: For machine learning algorithms and tools.
    - ◎ TensorFlow or PyTorch: For deep learning models (if applicable).
    - ◎ XGBoost or LightGBM: For gradient boosting algorithms (if applicable).
    - ◎ Data Visualization Libraries: Matplotlib and Seaborn for data visualization and plotting.

## 4. EXPERIMENTAL INVESTIGATIONS

- It is observed that the average SSC marks secured by students in Science stream is higher than those who pursued Commerce or Arts in HSC.

- It is observed that students pursuing degree in Commerce and Management possess higher grades in HSC than those in Science and other courses.

- The Male to Female ratio between students pursuing studies in Science and Technology is higher compared to other fields of study.

- It can be observed that marketing and finance is dominant in commerce and management field, whereas, all other degree types have students specializing in marketing and HR.

- It is observed that students studying in commerce and management field have higher chances of getting placed, whereas students belonging to the 'other' category have almost equal chances of getting placed or not.

- It can be observed that median salary cannot be the determining factor to decide the salary package of students.

- We can see that most students pursuing commerce and management have higher chances to get placement offers.

- Commerce and Management majors have higher salary package.

- It is observed that the average salary earned by Science and Technology majors are higher than the other department students.

- It can be seen that students with prior experience have better chances of getting placed. Moreover, many students in the marketing and finance field tend to get placed irrespective of level of experience.

- It is observed that candidates with more experience tend to earned higher salary packages.

- Over two-third of candidates are placed.

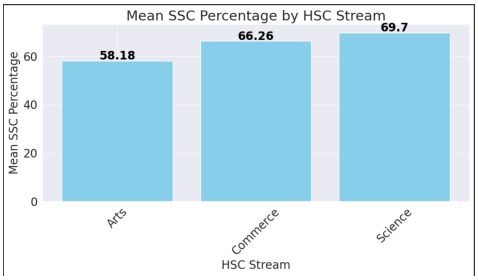
- It is observed that male candidates are higher in number. Proportion of male candidates getting placed is three-fifth of the entire male candidate applications whereas, about half the female candidates are placed.

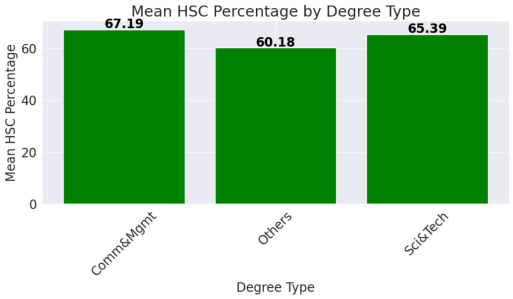
- Out of all the students participating in campus placement during a given year, half of them are placed.

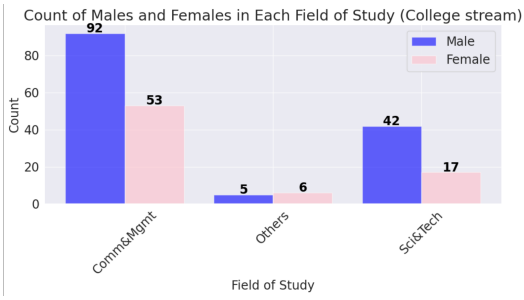
- Salary acquired by the candidate is strongly correlated with their MBA percentage.

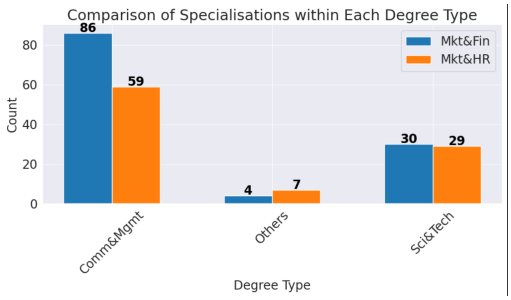
- It is observed that the logistic regression model has a higher accuracy to predict whether the student can get placed.

# 5. RESULTS

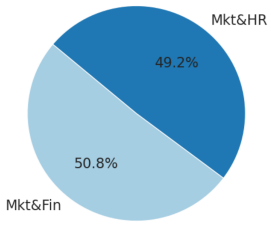


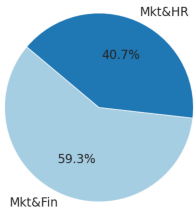



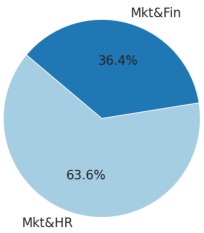

Distribution of Specialisations within Sci&Tech Degree Type

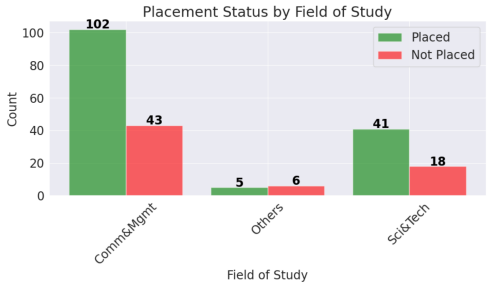


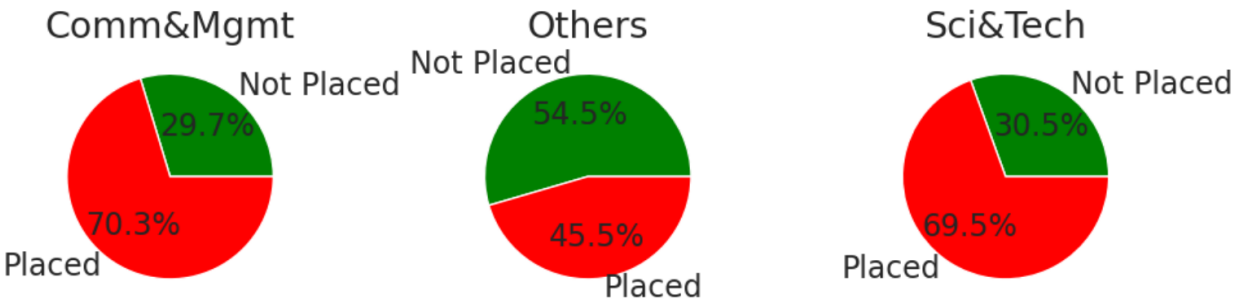

Distribution of Specialisations within Comm&Mgmt Degree Type

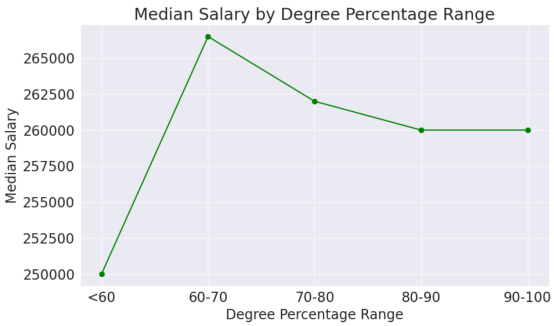


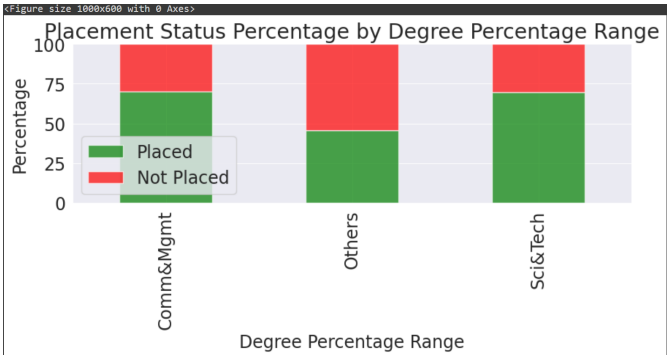

Distribution of Specialisations within Others Degree Type

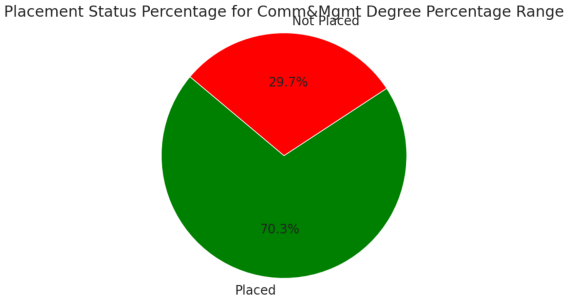


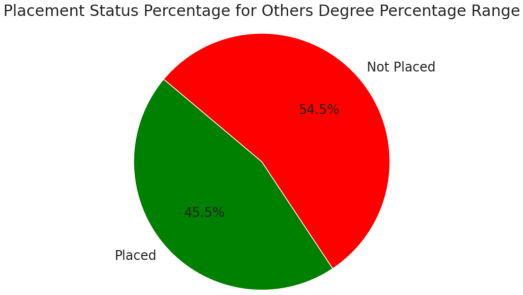



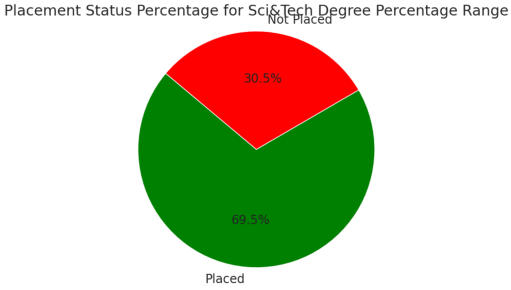

Comm&Mgmt
Not Placed
29.7%
70.3%
Placed
Others
Not Placed
54.5%
45.5%
Placed
Sci&Tech
Not Placed
30.5%
69.5%
Placed

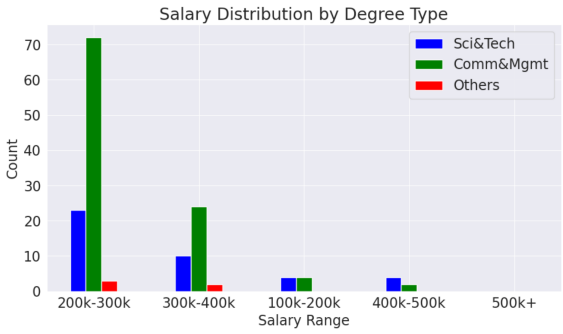

Median Salary by Degree Percentage Range
Median Salary
265000
262500
260000
257500
255000
252500
250000
<60
60-70
70-80
80-90
90-100
Degree Percentage Range

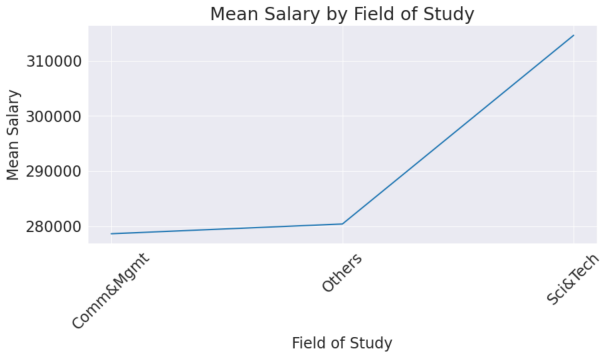

<Figure size 1000x600 with 0 Axes>
Placement Status Percentage by Degree Percentage Range
Percentage
100
75
50
25
0
Placed
Not Placed
Comm&Mgmt
Others
Sci&Tech
Degree Percentage Range

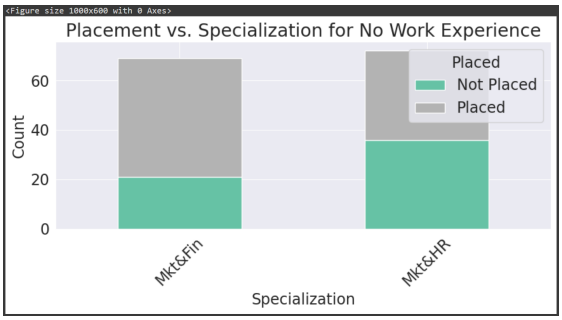

Placement Status Percentage for Comm&Mgmt Degree Percentage Range
Not Placed
29.7%
70.3%
Placed

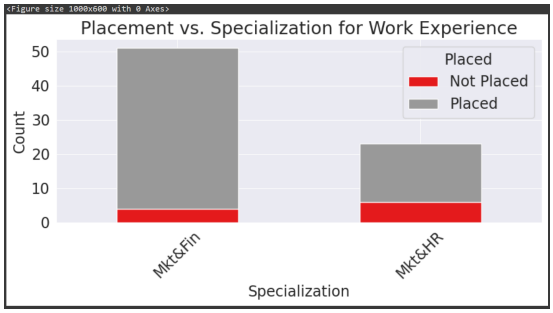

Placement Status Percentage for Others Degree Percentage Range
Not Placed
54.5%
45.5%
Placed

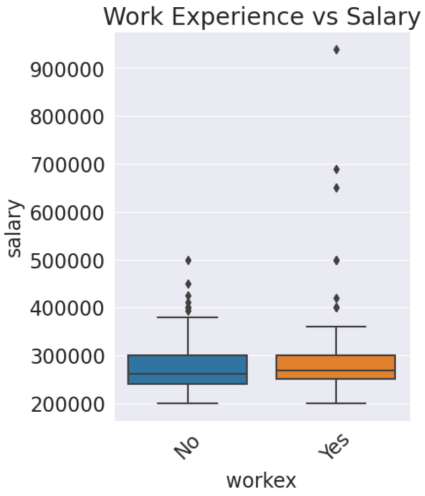

Placement Status Percentage for Sci&Tech Degree Percentage Range
Not Placed
30.5%
69.5%
Placed

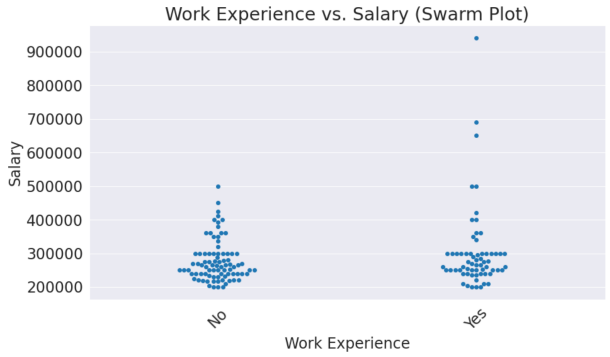

Salary Distribution by Degree Type
Sci&Tech
Comm&Mgmt
Others
Count
70
60
50
40
30
20
10
0
200k-300k
300k-400k
100k-200k
400k-500k
500k+
Salary Range

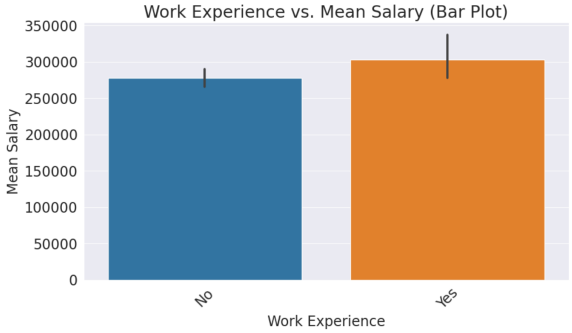

Mean Salary by Field of Study
Mean Salary
310000
300000
290000
280000
Comm&Mgmt
Others
Sci&Tech
Field of Study

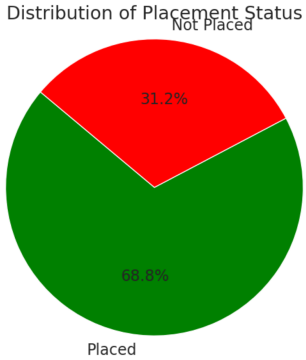

<Figure size 1000x600 with 0 Axes>
Placement vs. Specialization for No Work Experience
Placed
Not Placed
Placed
Count
60
40
20
0
Mkt&Fin
Mkt&HR
Specialization

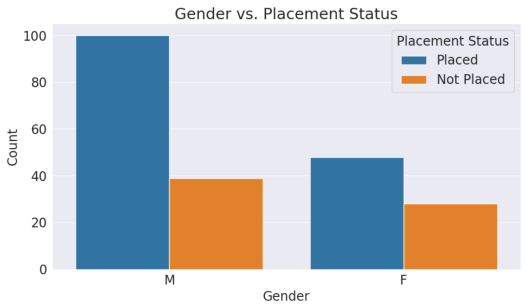

<Figure size 1000x600 with 0 Axes>
Placement vs. Specialization for Work Experience
Placed
Not Placed
Placed
Count
50
40
30
20
10
0
Mkt&Fin
Mkt&HR
Specialization

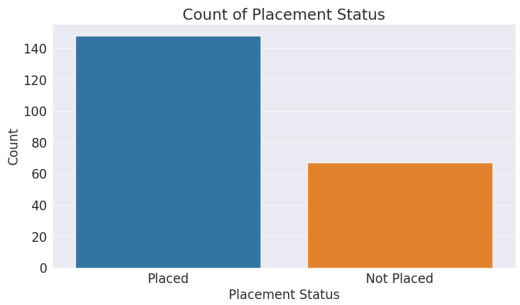

Work Experience vs Salary
salary
900000
800000
700000
600000
500000
400000
300000
200000
No
Yes
workex

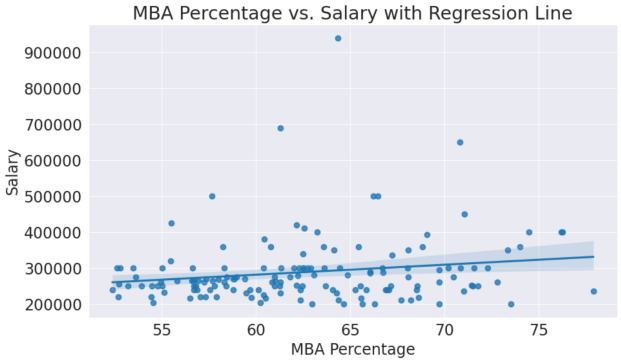

Work Experience vs. Salary (Swarm Plot)
Salary
900000
800000
700000
600000
500000
400000
300000
200000
No
Yes
Work Experience

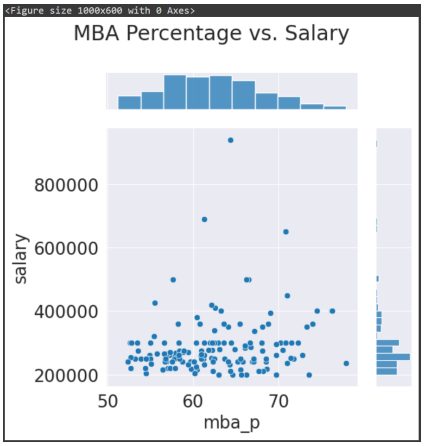

Work Experience vs. Mean Salary (Bar Plot)
Mean Salary
350000
300000
250000
200000
150000
100000
50000
0
No
Yes
Work Experience


Distribution of Placement Status
Not Placed
31.2%
68.8%
Placed



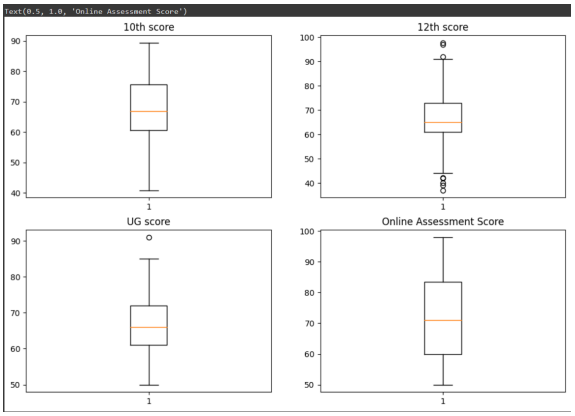



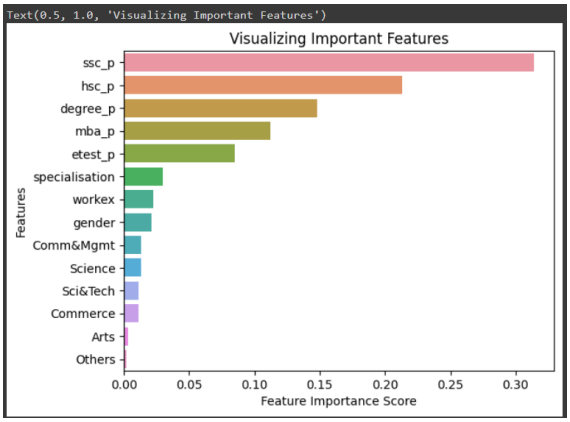




## Data Preprocessing: Data Inspection

| | sl_no | gender | ssc_p | ssc_b | hsc_p | hsc_b | hsc_s | degree_p | degree_t | workex | etest_p | specialisation | mba_p | status | salary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | M | 67.00 | Others | 91.00 | Others | Commerce | 58.00 | Sci&Tech | No | 55.0 | Mkt&HR | 58.80 | Placed | 270000.0 |
| **1** | 2 | M | 79.33 | Central | 78.33 | Others | Science | 77.48 | Sci&Tech | Yes | 86.5 | Mkt&Fin | 66.28 | Placed | 200000.0 |
| **2** | 3 | M | 65.00 | Central | 68.00 | Central | Arts | 64.00 | Comm&Mgmt | No | 75.0 | Mkt&Fin | 57.80 | Placed | 250000.0 |
| **3** | 4 | M | 56.00 | Central | 52.00 | Central | Science | 52.00 | Sci&Tech | No | 66.0 | Mkt&HR | 59.43 | Not Placed | NaN |
| **4** | 5 | M | 85.80 | Central | 73.60 | Central | Commerce | 73.30 | Comm&Mgmt | No | 96.8 | Mkt&Fin | 55.50 | Placed | 425000.0 |



Model -1: Logistic Regression

We are able to achieve **83%** accuracy.

Model -2: KNN

We are able to achieve **76.19%** accuracy.

Model - 3: Naive Bayes

We are able to achieve **72%** accuracy.

Model - 4: Decision Tree Classifier

We are able to achieve **73.8%** accuracy

Model -5: Random Forest Classifier

We are able to achieve **78.6%** accuracy

## 6. ADVANTAGES AND DISDVANTAGES

Some of the advantages of the project is as follows:

1. Comprehensive data exploration and visualization
2. Data pre-processing for model readiness
3. Evaluation of multiple machine learning algorithms

It calculates essential metrics, reveals feature importance, documents each step clearly, and employs modular design for reusability. Leveraging popular Python libraries for efficiency, it incorporates statistical analysis and presents results visually. Altogether, it demonstrates a systematic, well-documented, and modular approach to data analysis and modelling, aiding efficient decision-making and providing a valuable resource for similar tasks.

Some of the disadvantages are as described: It may lack extensive error handling, potentially leading to issues if unexpected data or errors occur. The code's complexity, especially in the machine learning model section, can be challenging for beginners to understand and modify. Furthermore, it may not address specific domain intricacies, requiring customization for certain applications. The reliance on external libraries could result in version compatibility problems or dependencies that may become obsolete over time.

## 7. APPLICATIONS

This project's applications span various sectors, aligning with placement trends in fields such as artificial intelligence, data science, and computer vision. Proficiency in these areas can open doors to positions in research, software development, and data analysis, reflecting the

evolving demands of the job market.

## 8. CONCLUSION

Therefore, the given campus placement dataset has been studied, and the relationship between various attributes have been analyzed using mathematical models. It can be concluded from the obtained results that we are able to achieve highest accuracy using Logistic Regression to determine whether the student is placed or not.

## 9. FUTURE SCOPE

The next version of the code will optimize performance with techniques such as model quantization and transfer learning, improve user interaction with easy-to-use interfaces and APIs, diversify pre-trained models, implement parallelism and adding good improvement feedback. This mechanism provides comprehensive documentation, enhances security, allows customization of hyper parameters, and supports multiple data formats. These improvements make the code more versatile, efficient, and accessible, enabling it to address a wider range of applications and user needs, such as image classification.

## 10. BIBLIOGRAPHY


- https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf
- https://balasahebtarle.files.wordpress.com/2020/01/machine-learning-algorithms_text-book.pdf
- https://www.javatpoint.com/machine-learning-algorithms
- https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML


## 11. APPENDIX

Original file is located at: https://colab.research.google.com/drive/1ul-gH600wC92Qakbv0fIlNtV9NECgYtp

#Dataset used: https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement