Campus Recruitment Analysis

BY TEAM ANALYTICA

Anish S
PES2UG19CS045
Student of
PES UNIVERSITY
EC CAMPUS
Electronic City, Bengaluru – 100, Karnataka, India
anishsunil01@gmail.com

Kevin Thomas
PES2UG19CS184
Student of
PES UNIVERSITY
EC CAMPUS
Electronic City, Bengaluru -100, Karnataka, India
kevinmt.727@gmail.com

Preethika K
PES2UG19CS172
Student of
PES UNIVERSITY
EC CAMPUS
Electronic City, Bengaluru – 100, Karnataka, India
preethika1102@gmail.com

Sneha Sujit Saha
PES2UG19CS393
Student of
PES UNIVERSITY
Electronic City, Bengaluru -100, Karnataka, India
Sneha.pes19@gmail.com

***Abstract*—this is a report of the data analysis performed on the Campus Recruitment dataset. This report helps our readers get a contextual understanding of the methods and models built to better interpret the data provided by the source dataset.**

Keywords—Random Forest, SVM, KNN, ID3

I. Introduction

Campus Recruitment is one of the events that occur in almost every higher educational institution that offers degree courses to the enrolled students. This event can be further analysed to improve upon the current system of manual recruitment or add to the current trend of automation to build machine learning based AIs to handle this process smoothly.

Our team has focused its efforts to build simple yet accurate ML models to predict the outcome of recruitments by first visualizing the data, interpreting it, deducing a problem statement out of it, studying the literature to gain knowledge on previously experimented and worked on projects and the results they yielded and tackling the said problem

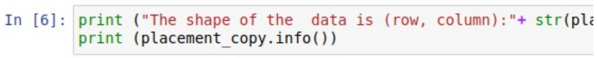
II. Exploratory Data Anaysis and Visualization

This is one of the most crucial steps in data handling as this helps the analysts to get a better insight of the true contents of the data and the relationships (statistical or otherwise) between the different components of the dataset[1]. Have a look at our EDA and visualizations, as to how this step helped us infer information about the dataset [1], create a problem statement and finally how this helped us construct models to better explore the concepts of data analysis.

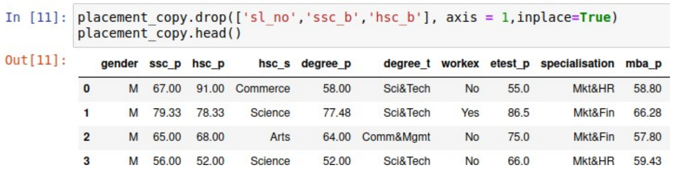
A. EDA

The data was explored and the following information was gained from it:

- Shape of the data is 215 rows x 15 attributes/columns.



```
In [6]: print ("The shape of the  data is (row, column):"+ str(pla
        print (placement_copy.info())
```

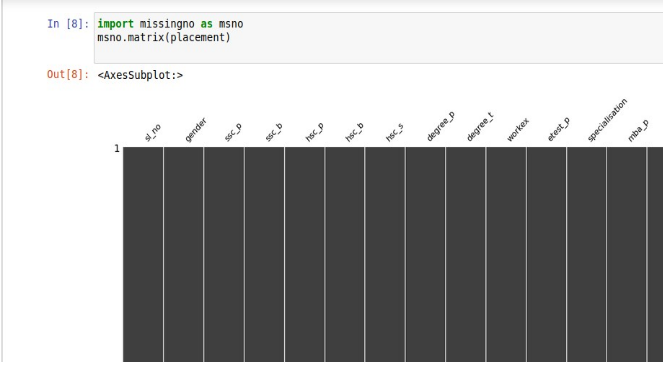


```
In [11]: placement_copy.drop(['sl_no','ssc_b','hsc_b'], axis = 1,inplace=True)
         placement_copy.head()

Out[11]:
```

	gender	ssc_p	hsc_p	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p
0	M	67.00	91.00	Commerce	58.00	Sci&Tech	No	55.0	Mkt&HR	58.80
1	M	79.33	78.33	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28
2	M	65.00	68.00	Arts	64.00	Comm&Mgmt	No	75.0	Mkt&Fin	57.80
3	M	56.00	52.00	Science	52.00	Sci&Tech	No	66.0	Mkt&HR	59.43

- There was 1 attribute, namely ‘Salary’, that consisted of 67 missing data points.

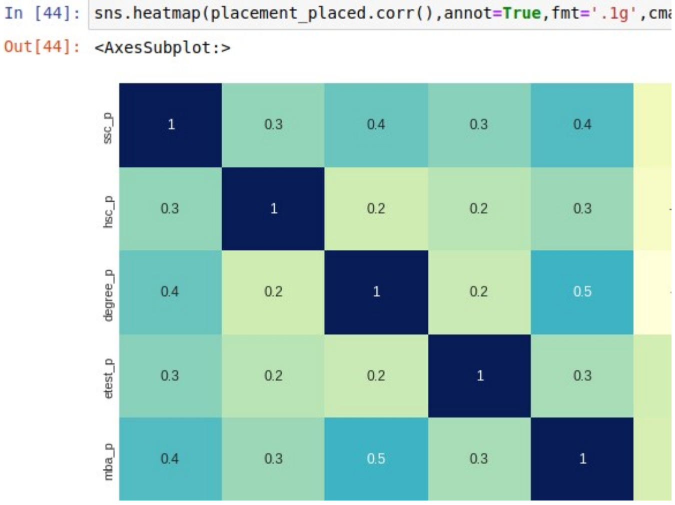


```
In [8]: import missingno as msno
        msno.matrix(placement)

Out[8]: <AxesSubplot:>
```

- The attributes ‘ssc_b’, ‘hsc_b’ are incomplete data as there are many other boards other than central and the attribute ‘sl_no’ is not contributing any new value to the existing dataset. Such attributes are appropriately dealt with.

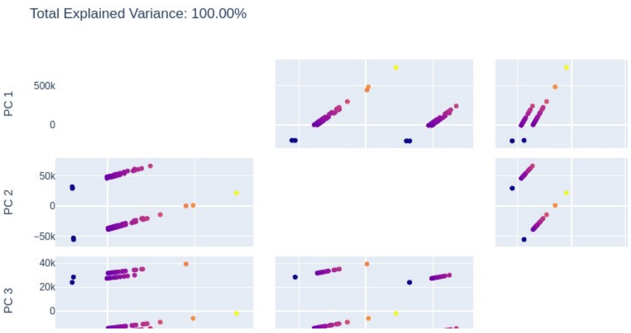
- The attributes within the dataset are correlated as seen below:



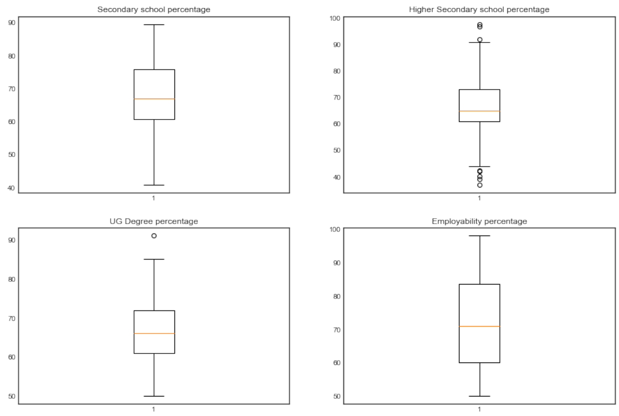
```
In [44]: sns.heatmap(placement_placed.corr(),annot=True,fmt='.1g',cma

Out[44]: <AxesSubplot:>
```

- PCA Analysis: We apparently found 3 principal components.



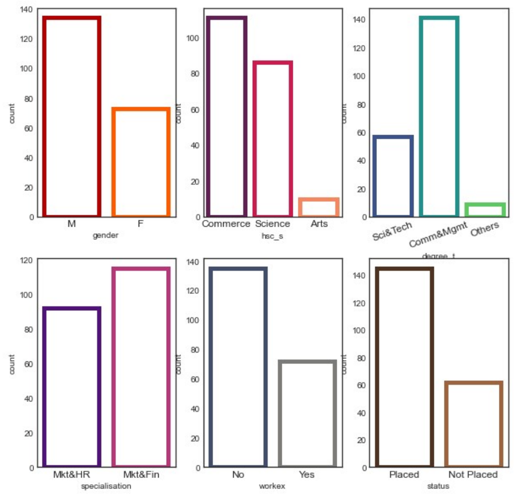
- Checking for outliers in the given dataset: It was observed that only one attribute namely ‘hsc_p’ or higher secondary school percentage has some outliers.



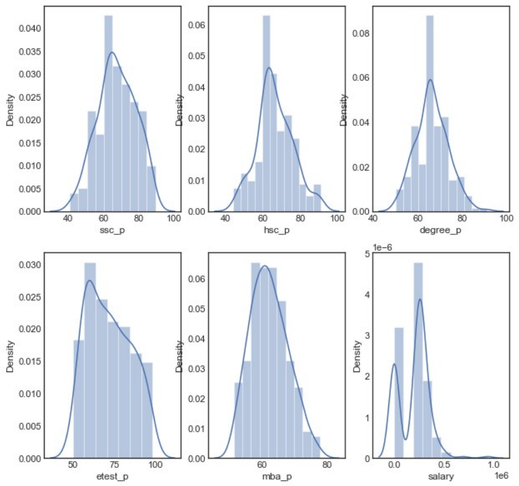
B. Visualizations

The following visualizations of the data are done:

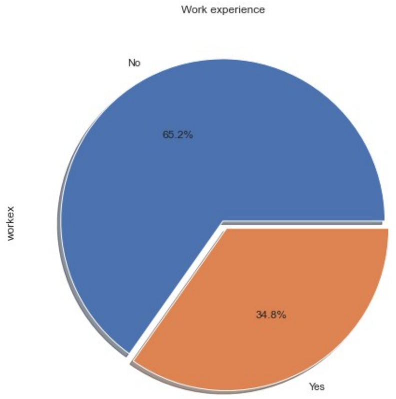
- Bar-Plots



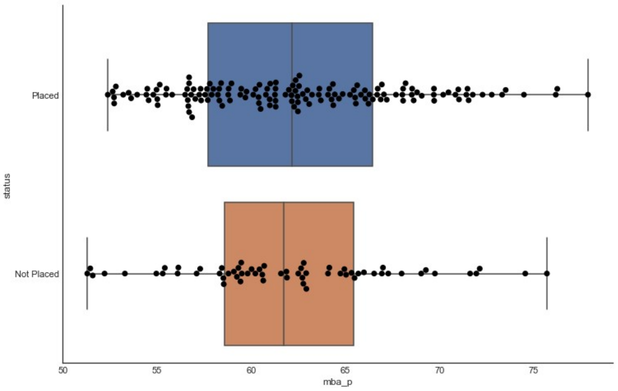
- Histograms

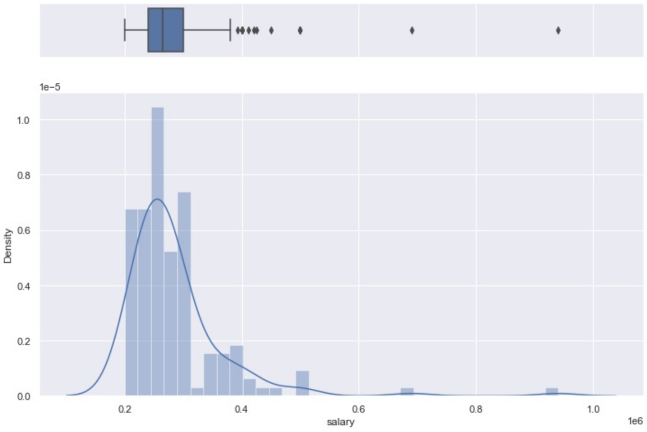


- Pie-chart

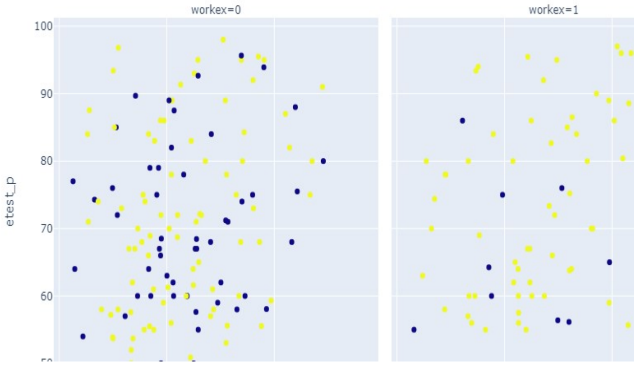


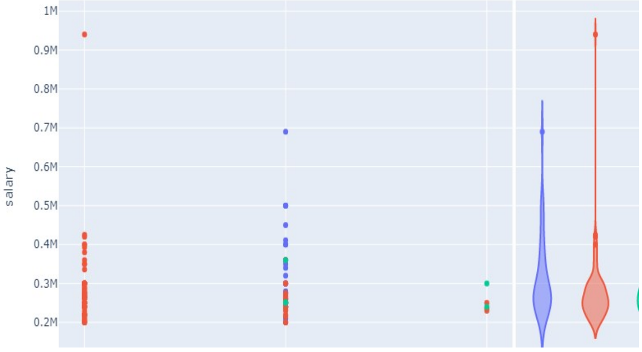
- Box-plot



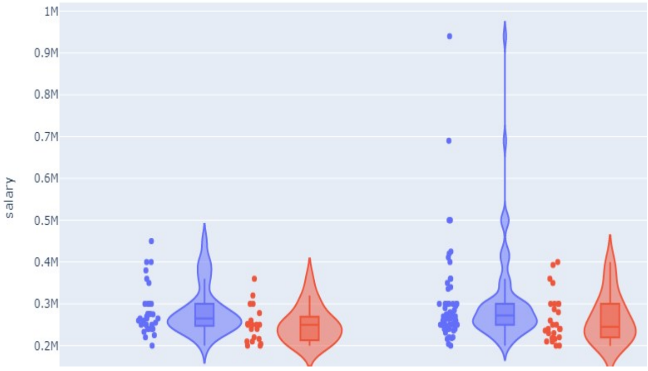


- Scatter-plot

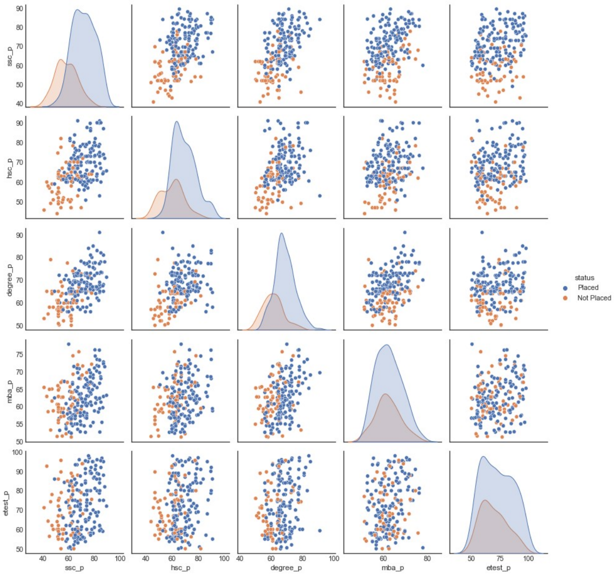




- Violin plot



- Pair plot



C. *Inferences*

From observing the above displayed visualizations and via EDA, our team arrived at the following drawn conclusions:

1. From the Bar plots: The candidates who have got placed are mostly men; All the candidates who have applied are from the Marketing and Finance or Marketing and Human resources specialization; Many of the candidates who have gotten placed have 0 work experience.
2. From the Boxplots: MBA Score or percentage does have an influence on the placement status; Many candidates received a package between 2lakhs to 4 lakhs per annum salary; Only one candidate has got around 10 lakhs salary package; The avg salary obtained is around 2LPA.
3. From the Histograms: Almost all the distributions of attributes follow normal except for 'Salary'. Most candidates' academic performances lay around 60%-80%.
4. From the Pie-chart: Nearly 66.2% of the candidates have no work experience.
5. From the scatter plots: There's no relation between employability test and mba percentage; Most of the students, regardless of their previous work experience, have gotten placed if they've performed better in the employability test though; People from the Science and Tech sectors on an average earn slightly more than those from the Commerce and Management backgrounds; However, the highest salary is bagged from a student in Commerce and Management.
6. From the Violin Plot: Only Male candidates have bagged the top salaries.
7. From the Pair-plot: candidates with good performance in higher secondary school have gotten placed.

III. Data Pre-Processing

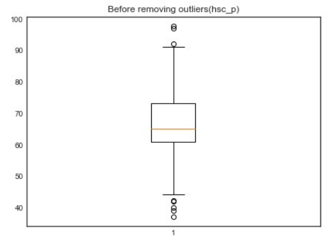
Data Pre-processing refers to manipulating and dropping of data before it is used in order to ensure or enhance performance. The phrase ‘garbage in, garbage out’ can be easily applied to data mining and machine learning projects; pre-processing helps avoid this obstacle by making to it easier to get better and easily manipulative data from our raw dataset. It explicitly helps in dropping out-of-range values, missing values and null values to expertly improve the quality of data analysis.

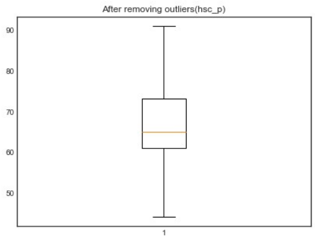
In our project we have performed the following data pre-processing techniques:

A. Data Cleaning – Handling missing values and removal of Outliers

As seen in the EDA, the dataset under consideration has one attribute called ‘Salary’ with 67 missing values. This has been dealt by replacing the missing values with ‘0’.

As for outliers, the one attribute ‘Higher-Secondary School Percentage’ is the only one to have outliers. This also has been dealt with by removing any value that lies outside the inter-quartile range in the box-plot representation of this attribute’s data.





B. Label Encoding

Label encoding refers to conversion of labels to numeric form so as to make it easier to apply machine learning algorithms onto them later in the data processing step.

The four attributes that underwent this process in our dataset were: ‘gender’, ‘workex’, ‘specialization’ and ‘status’ with following conventions:

1. gender (1,0) -> (male, female)
2. workex(1,0) -> (yes, no)
3. specialization(1,0) -> (Mkt & HR, Mkt & Fin)
4. status(1, 0) -> (placed, not placed)

C. One Hot Encoding

One hot encoding is the process of conversion of categorical variables to a form that could help ML algorithms performs better.

The attributes that underwent this process in our dataset were: ‘hsc_s’ and ‘degree_t’ with the following distributions:

1. hsc_s -> temp_science, temp_arts, temp_commerce
2. degree_t -> temp_Sci&tech, temp_Comm&Mgmt, temp_Others

D. Training and Testing Split(80:20)

The data was further pre-processed by splitting it to training and testing datasets. This was done to calculate the accuracy of the models working on them in the later stages of the project. Also, the ratio of 80:20 was chosen to split the data as this is one of the most common and fondly used ratio values for ‘train and test split’.

IV. Literature Survey

Literature Survey is a term that refers to the process of studying different research papers and reports that constitutes information and results relevant to the project or experiment being performed by the surveyors.

Our team has surveyed over 12 different papers, most of which were published by the famous IEEE to gain information on the various models that can be used to act upon data almost similar to the one currently being worked upon. The link to the various papers surveyed by our team is listed in the references section [2].

To paint a brief picture of the inferences and our understanding of the content studied during the literature survey, the following are listed:

We concentrated upon building 5 different models to compare their accuracies on the dataset [1]. For that purpose we concentrated upon the following papers:

1. Student Placement Prediction using SVM[3]
2. Campus Placement Prediction Using Supervised Machine Learning Techniques[4]
3. Recruitment System with Placement Prediction[5]
4. Students’ Performance and Employability Prediction through Data Mining: A Survey[6]
5. A Comparative Study on Machine Learning Algorithms for Predicting the Placement Information of Under Graduate Students[7]

The one different approach or rather one main difference between our project and the ones that we referred is that the models that we’ve chosen to compare the results have not been considered together in any of the researches conducted so far on this particular form of data.

V. Problem Statement

This here is the problem that we through this project wish to solve and observe the results off of:

The objective of the project is

- To determine whether a student gets placed in a company or not using 5 different supervised models and compare all of them and find out which model provides the highest accuracy;
- To determine the range/tier of the placed student’s salary in order to predict this value beforehand for new students;
- To determine if there exists some form of gender bias in the recruitment process.

VI. Data Processing - I

A. Basic Concepts

1. KNN

In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. It takes for input the k closest training examples from the dataset. [8]

In *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor. [8]

In k-NN regression, the k value is determined by the elbow method.

2. Support Vector Machine(SVM)

Support vector machines are machine learning models with associated learning algorithms that analyze data for classification and regression analysis. [9]

For a training dataset, each classified to one of two classes, an SVM training algorithm builds a model that assigns new values to one class or the other, making it a linear classifier.SVM maps training data to points in place so as to maximize the width of the hyper plane between the two classes. [9]

The new training dataset is mapped in the same space and to predict which class they belong to. [9]

3. Random Forest Algorithm

Random forest is an ensemble type classification or regression algorithm that constructs a multitude of decision trees at training time. [10]

For classification problems, the output of the random forest is the class selected by most trees. For regression problems, it finds the mean or average prediction of the individual trees. [10]

Random forests don't undergo the problem of overfitting to their training set. Random forests always outperform decision tree classifiers; however their accuracies are lower than gradient boosted trees. [10]

4. Naïve Bayes Classifier

Naïve Bayes Classifiers are a family of simple 'probabilistic classifiers' based on applying Bayes' Theorem with strong independence assumptions between the features. They are the simplest Bayesian network models that can achieve higher accuracy levels. [11]

Naïve Bayes Classifiers are highly scalable requiring a number of parameters to be linear. Maximum-likelihood training can be done by evaluating a closed-form expression, rather than by expensive iterative approximation as used for many other types of classifiers. [11]

5. ID3 Algorithm

ID3 (Iterative Dichotomiser 3) is one of the oldest algorithms to produce decision trees. ID3 is the predecessor of C4.5 algorithm, and is typically used in the machine learning and natural language processing. [12]

6. K-fold Cross Validation

In K-fold cross validation, the original sample is randomly split into 'k' equal sized parts. Of the 'k' parts, a single part is retained as the validation data for testing the model, and the remaining 'k-1' parts are used as training data. This process is then repeated 'k' times, with each of the 'k' parts used exactly once as the validation data. The mean of the 'k' results is then used to find out the estimation. [13]

B. Models Built

The following models were built after the pre-processing step and their accuracies compared and here are the results:

MODEL	ACCURACY
Random Forest	80.95238095238095 %
ID 3	61.904761904761905 %
SVM	83.33333333333334 %
KNN	76.19047619047619 %
Naïve Bayes Classifier	67.1190476190476 %

From the table it's easy to infer that SVM gives the best accuracy when compared to other models. This proves the flexibility and the versatility of the SVM to produce high accuracies in many different kinds of conditions. It is to be noted that based on the literature survey conducted, we expected the Random Forest model to give the highest accuracy compared to others. While it still came in at a close second position, this may have been due to the smallness (in size) of the dataset that was worked upon and one can expect this model to give out better results for larger and more complete datasets.

From the table, we can see that KNN came at third position compared to the rest and once again this can be attributed to the fact that the dataset under consideration is not of too good a quality in terms of diversity and richness. To get better results, one can definitely venture into building ensemble models constituting of strong classifiers such as SVM, KNN and Random Forest.

Also another point of note would be the Python Modules used to generate these models. **Sklearn** is perhaps the most useful library there exists to easily generate ML models in python language. We have made use of Sklearn to produce over 4 out of 5 of our ML models, excluding ID3.

Finally, this concludes the model building part of our project and part 1 of our problem statement.

VII. Data Processing – II

Predicting Salaries based on Academic performances and checking whether it can accurately predict tiers. The academic performances considered in the case are 10th percentage, MBA percentage and degree percentage for prediction of salary using Random forest. The predicted Salary tier is compared with the actual company tier. In our case tiers are classified based on salaries the company offers, tier 1 is considered to be above 500K, tier 2 is above 300K and tier3 below 300K.

The predicted salaries in this model accurately classify the tiers without any false classification

VIII. Data Processing – III

Comparing female and male salaries with respect to their respective academic performances .The dataset is grouped based on gender on the top 20 male and female students and their marks and salaries are retrieved. The average marks, salaries and the correlation between marks and salary is found. It is found that the academic performance is relatively the same for the top 20 students of the class but the academic performance seems to be highly correlated to salaries of women whereas for men it doesn't. Similar attributes is retrieved for students who earn more than 300K and less (Classified in this case as tier 1 and tier2 respectively). Similar trend as above is seen in both the tiers, and the representation of female students is less in tier companies despite having a better degree percentile. However representation of female students is relatively same as male students in tier 2.

CONCLUSIONS

Team Member	Contribution
Anish S	Making of the report, Reviewed 4 papers for literature survey. Contributed to writing the code (ID3, Random forest, naïve Bayes).
K Preethika	Making of the report, Reviewed 3 papers for literature survey. Contributed to writing the Code (SVM, KNN, gender discrimination and salary prediction).
Kevin Thomas	Reviewed 4 papers for literature survey
Sneha Sujit Saha	Reviewed 1 paper for literature survey

For the first part of data processing we compared 5 models with training and testing datasets and it is found that SVM has the highest accuracy and ID3 has the lowest.

The second part we predicted the company tier which the student got placed at based on their respective academic performance and an accuracy of 100 percent was achieved.

For the third part we analyzed whether gender discrimination exists and it was found that in all categories men relatively earned higher packages and it had a negative correlation to their marks, whereas for women they were less represented in tier 1 companies despite having better average marks and also marks are directly correlated to the salary packages. Therefore we conclude that gender discrimination does exist while hiring.

For complete information on the project undertaken by our team please refer to the github [14] link attached in the references.

References

The following if the list of the materials referred to complete the project:

The first two links here are the dataset link and the list of literature survey papers link respectively.

[1] https://github.com/Sage101201/Campus-Placement/blob/main/Placement_Data_Full_Class.csv

[2] https://drive.google.com/drive/folders/1BMBS5RhkisbKVp9HRk0L64_iLxCkydeH?usp=sharing

The next five (3-7) are the links of the most referred research papers.

[3] https://drive.google.com/file/d/1A1euAN6RgrAs2_D845g-vtKMLO3s-Jyt/view?usp=sharing

[4] https://www.ripublication.com/ijaer19/ijaerv14n9_19.pdf

[5] https://drive.google.com/file/d/1ly2imSuK4EzSEFhkJEKgsmiNQIqdk5IU/view?usp=sharing

[6] https://drive.google.com/file/d/1Sv_8AyIj0-q_pqI55zOFzIdnlYfwgWVu/view?usp=sharing

[7] https://drive.google.com/file/d/1aKqGA6XQkJNTfAbiUGADRGXxn9ECOtue/view?usp=sharing

The next six (8-13) are the links of the concepts involved behind the models.

[8] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[9] https://en.wikipedia.org/wiki/Support-vector_machine

[10] https://en.wikipedia.org/wiki/Random_forest

[11] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[12] https://en.wikipedia.org/wiki/ID3_algorithm

[13] https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation_with_validation_and_test_set

This last one is the link of the Github repository where all the information related to the project is stored.

[14] https://github.com/Sage101201/Campus-Placement