Illinois Institute of Technology

CS-584 Machine Learning

Prediction of Youth Employment

Urva Surti
A20516868
usurti1@hawk.iit.edu

Sudarshan Thakur
A20522408
sthakur12@hawk.iit.edu

Bandhavi Parvathaneni
A20516844
bparvathaneni@hawk.iit.edu

Dr. Yan Yan

Contents

CHAPTER's

ABSTRACT

The covid-19 pandemic struck humanity as an unprecedented event. Along with loss of lives there was also a huge loss of jobs. There was a massive job loss around the world. The students who were just about to start their career were in a dismal regarding their placements. This project addresses this issue of the campus placements. Students with a good college gpa, 10-12th marks, projects and work experience are usually considered to have a strong foothold. This project focuses on placement of a college student considering the parameters/features such as 10th -12th percentage marks, graduation gpa, salary expectation, work experience, stream the student belongs to, the board in which he/she studied etc. The machine learning model inputs the features from the user and gives out a result in the form of placed and unplaced. The project tests various machine learning algorithms like Gaussian Naive Bayes, Random Forest, XGBoost and K Nearest Neighbors. The best performing algorithms were chosen for stacking and creating a new hybrid model which was deployed over a web application using flask in python. Based on the result the student can decide about how he/she has to go around for the preparation of campus placements

CHAPTER 1

INTRODUCTION

PURPOSE

The lack of employment for youth is one of the major issues that citizens are facing in this country. And the main reason for this unemployment can be reckoned to be the wanting skillset of the fresh graduate students as well as lack of self-awareness & self-evaluation done by the candidates (students). It's a major issue as youth unemployment directly affects the economy of the nation and has detrimental effects on the student as well as the family of that student. Also, with the advent of the COVID-19 pandemic, which is causing a global financial crisis, this issue of youth unemployment is being faced globally. The main purpose of this project is to assist candidates (students), create a sense of awareness among the youth regarding their employment, and helping the economy flourish by addressing this unemployment problem. If the candidate gets an idea of the area where they're lacking, and which needs to be refined then this problem of rising unemployment can be curbed

SCOPE

The purpose of this project can be achieved by creating a prediction model that will use the data provided by the student to predict whether that student is employable or not. This prediction will be based on various factors such as the academic record of the student and the overall performance of the student across various academic stages in his career. This sort of prediction will help the candidate in understanding where he/she stands in the current scenario which will further assist the candidate to work on various other aspects in order to get placed in the future.

OVERVIEW

With the help of Machine Learning Algorithms and Prediction Modelling, candidates just by providing some data can find out whether they're employable or not and be prepared for the future by enhancing their skills in the domain they wish to be employed into. The best performing model will then be deployed along with GUI which will be user-friendly such that anyone can access it with ease

CHAPTER 2

PROBLEM DISCRIPTION

PROBLEM STATEMENT

The problem statement of this project revolves around predicting the employability i.e. whether a particular candidate participating in the placement process in order to seek employment manages to bag in a job offer or not and provide a better performing prediction model which will outperform all the conventional & previously employed models by the researchers.

PROPOSED SOLUTION

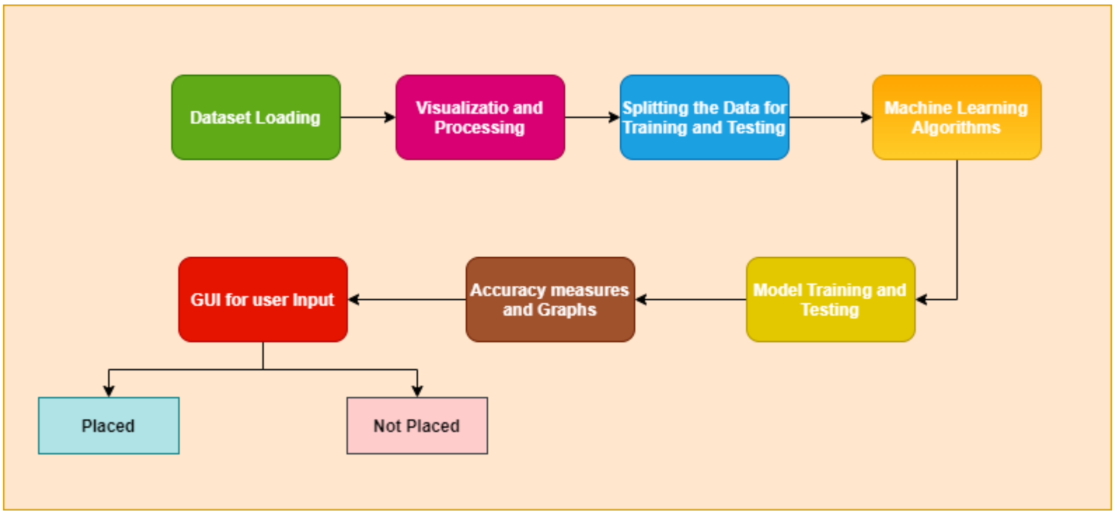
The parameters/features such as 10th -12th percentage marks, graduation gpa, salary expectation, work experience, stream the student belongs to, the board in which he/she studied etc are used as an input data to the model. The machine learning model inputs the features from the user and gives out a result in the form of placed and unplaced. The project tests various machine learning algorithms like Gaussian Naive Bayes, Random Forest, XGBoost and K Nearest Neighbors. The best performing algorithms were chosen for stacking and creating a new hybrid model which was deployed over a web application using flask in python. Based on the result the student can decide about how he/she must go around for the preparation of campus placements

CHAPTER 3

DESIGN

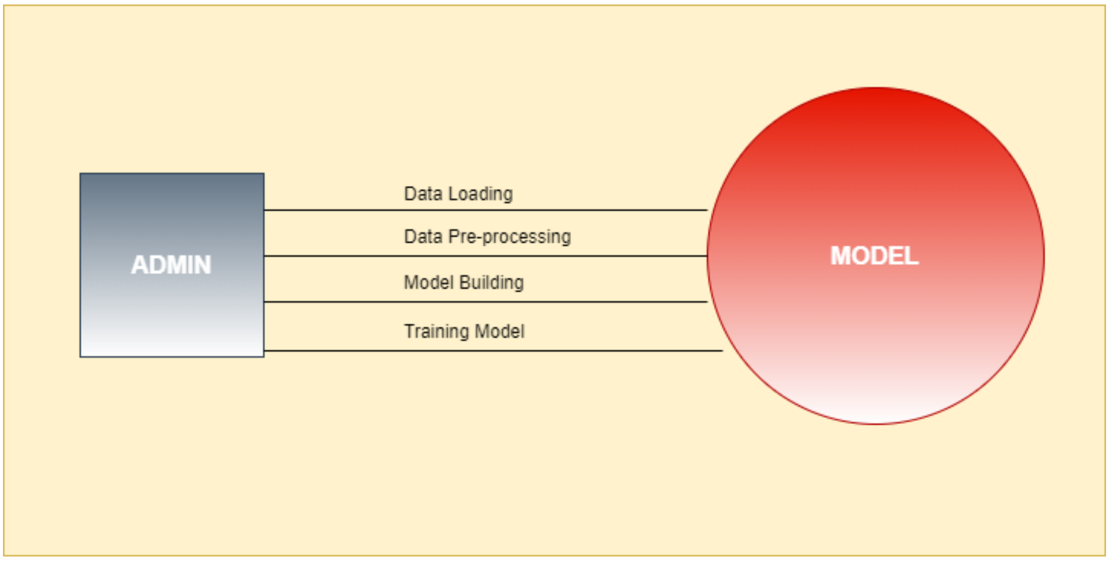
SYSTEM ARCHITECTURE

The system architecture provides a complete overview of the various processes implemented to build and design this complete project. From the figure below, it can be observed that dataset loading is the first process of the project followed by performing visualizations (EDA) and data processing which involves feature elimination, feature encoding and handling null values present in the dataset. The dataset is then split into two i.e., training (70%) and testing (30%) which is used for training and testing the ML model respectively.

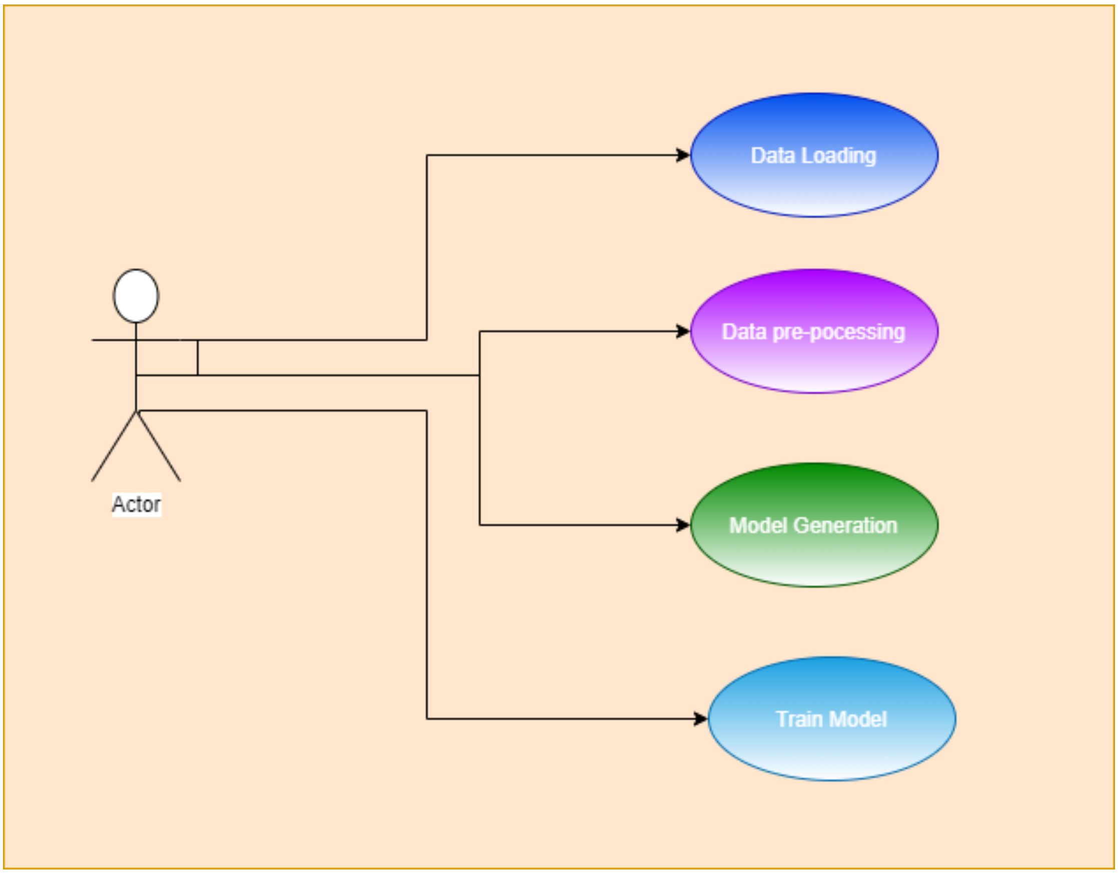

Dataset Loading
Visualizatio and Processing
Splitting the Data for Training and Testing
Machine Learning Algorithms
Model Training and Testing
Accuracy measures and Graphs
GUI for user Input
Placed
Not Placed

After the ML model is trained with the data it is then tested using different inputs and its performance is evaluated on the basis of various performance metrics such as accuracy, misclassification rate, precision, recall etc. Then the best performing model is chosen for deploying over a web application. The model is deployed using flask framework in python which after being deployed provides output after passing on a few sets of inputs to the model.

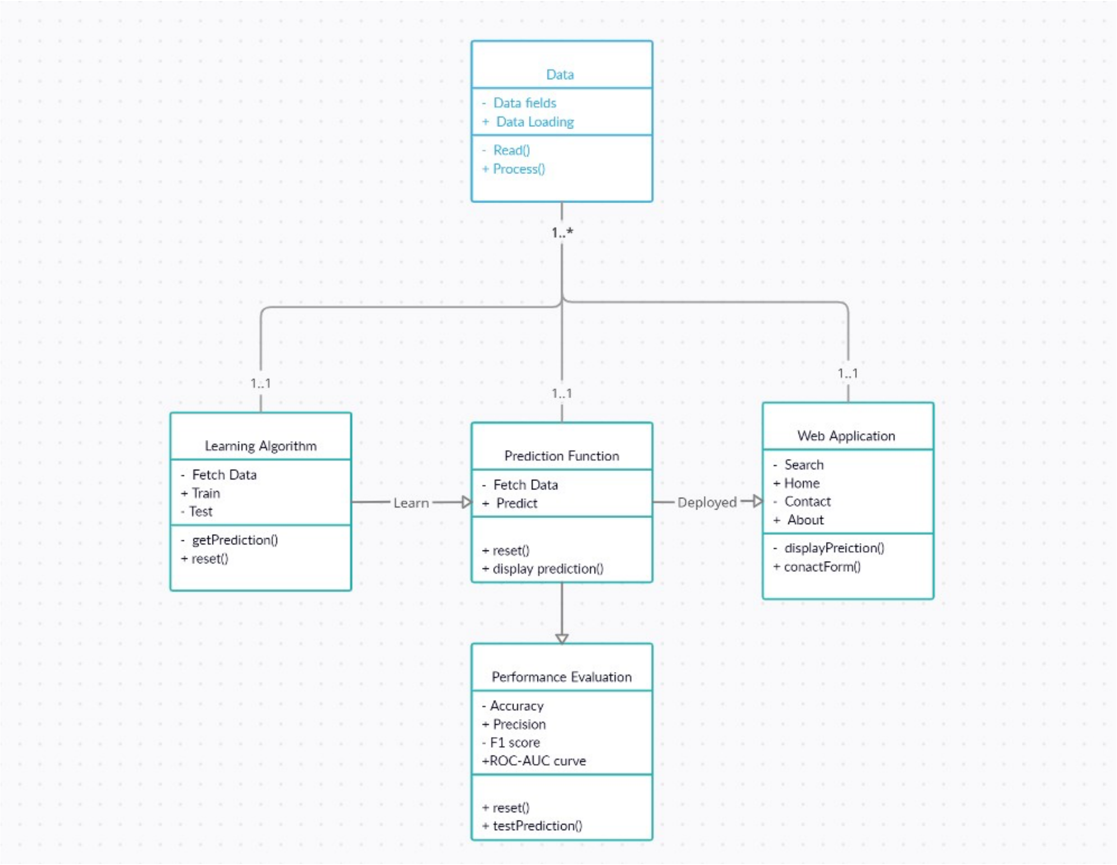
DATA FLOW DIAGRAM



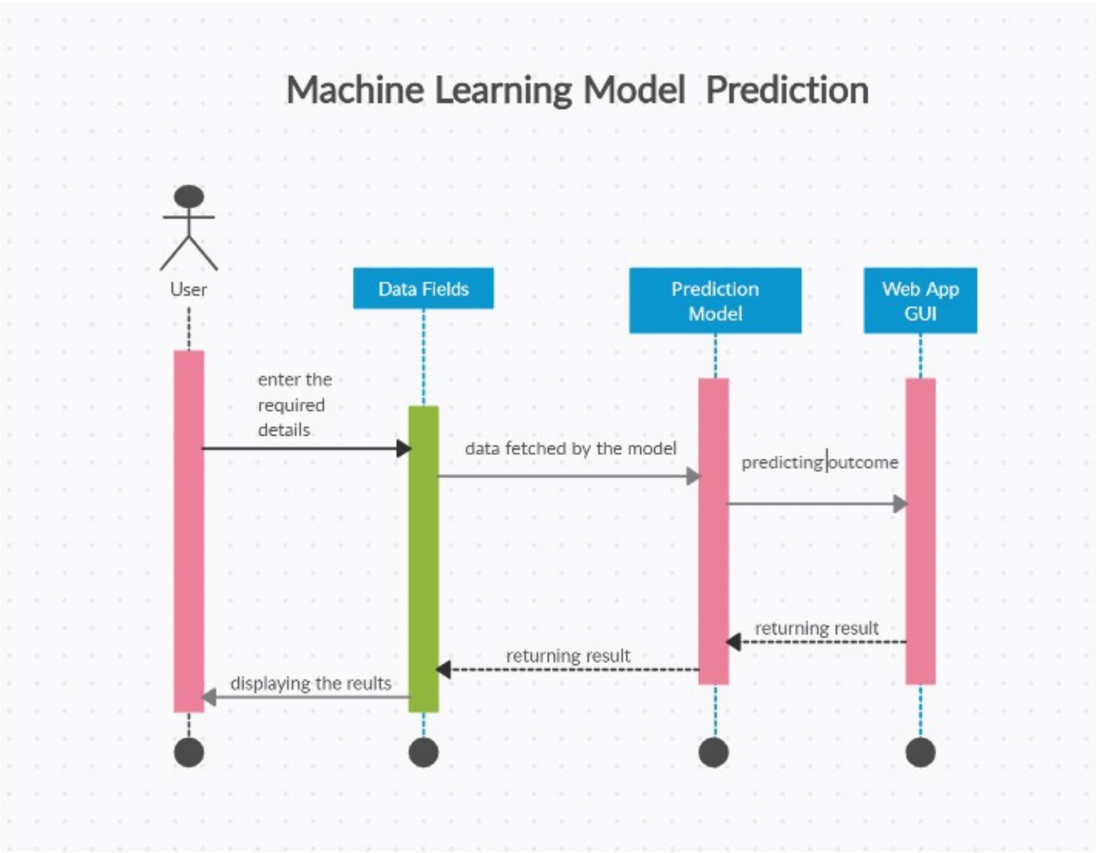
USE CASE DIAGRAM



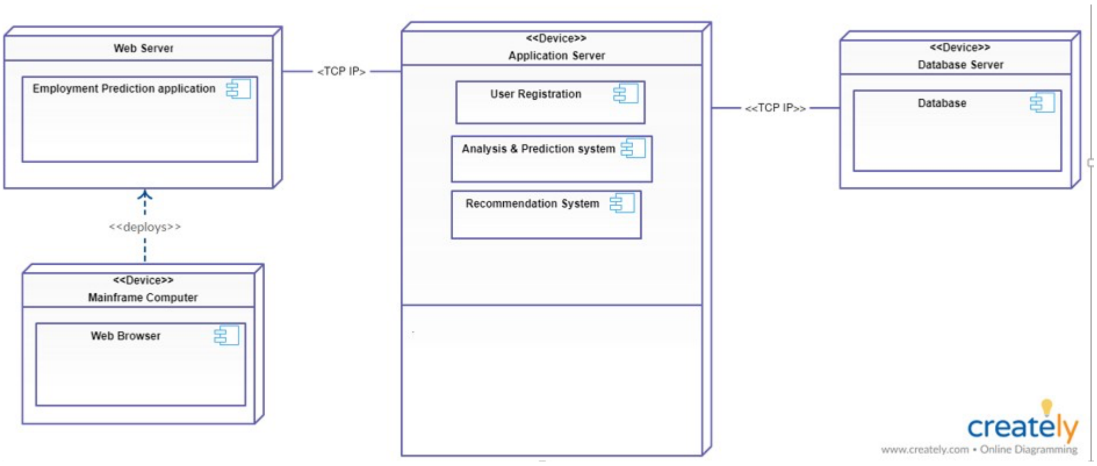
ACTIVITY DIAGRAM



SEQUENCE DIAGRAM



DEPLOYMENT DIAGRAM



CHAPTER 4

RESULT ANALYSIS AND IMPLEMENTATION

MODULE 1

It consists of reading the dataset into the system, pre-processing the data. The pre- processing of data consists of handling null values within the dataset, feature elimination and encoding features into suitable form for the process of modelbuilding.

The table below denotes all the features descriptions and their respective encoded values.

Sr No.	Feature Name	Feature Code	Description	Encoded values
1	Serial Number	sl_no	-	-
2	Gender	gender	Gender of the student	-
3	10th percentage	ssc_p	Percentage marks obtained in 10th	-
4	10th board	ssc_b	Board to which the student belongs in 10th class	central: 1 others: 0
5	12th percentage	hsc_p	Percentage marks obtained in 12th	-
6	12th board	hsc_b	Board to which the student in12th class	central: 1 others: 0
7	12th stream	hsc_s	Stream of the student in class 12th	arts: 1 commerce: 2 science: 3
8	Degree Percentage	degree_p	Percentage marks obtained in graduation	-

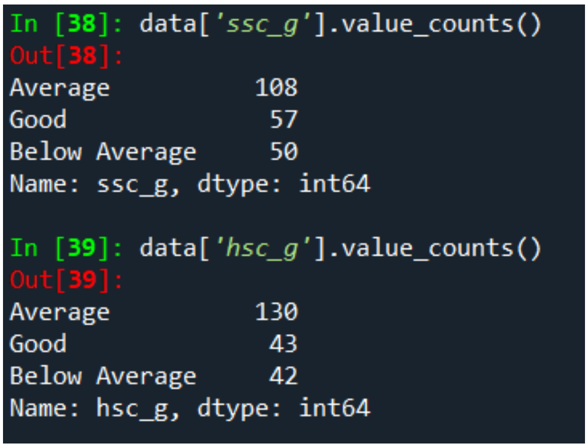
9	Degree Stream	degree_t	Stream of graduation	comm&mgmt: 1 others: 2 sci&tech:3
10	Work Experience	workex	Student’s status of work experience	Yes: 2 No: 1
11	E-test Percentage	etest_p	Percentage marks obtained in pre- placement test	-
12	Postgraduate (MBA) Specialization	specialisation	MBA specialization stream	mkt&hr: 1 mkt&fin: 2
13	MBA percentage	mba_p	Percentage marks obtained during MBA	-
14	Placement Status	status	Status of student’s placement	-
15	Salary	salary	Amount promised as salary	Placed: 1 Not Placed 0

MODULE 2

Consists of exploratory data analysis(EDA) performed on the dataset and also splitting the data in train (70%) and test (30%). 70% of the data is utilized for training the prediction model whereas the rest 30% is utilized for testing the outputs and checking the performance of the model.

The exploratory data analysis (EDA) which gives us important insights and trends within the data which would help in further eliminating or considering the attributes of prime importance when it comes to building a prediction model. EDA has been performed with the help of visualizations using matplotlib and seaborn libraries in python. While performing the EDA, a few new features were introduced where in the students were classified into 3 categories on the basis of their percentage marks obtained in various academic stages such 10th, 12th, graduation, MBA and e_test. The 3 categories were Good, Average & Below Average. The students in the good category had scored 75% and above in their academics, similarly the students in the Average category were those who had scored 60%-74% marks and the students scoring below 60% were categorized as below average.

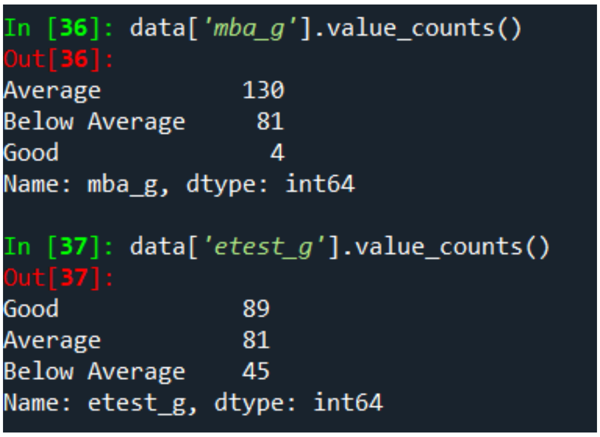
The 2 images below denote the distribution of students across the categories:



```
In [38]: data['ssc_g'].value_counts()
Out[38]:
Average          108
Good              57
Below Average     50
Name: ssc_g, dtype: int64

In [39]: data['hsc_g'].value_counts()
Out[39]:
Average          130
Good              43
Below Average     42
Name: hsc_g, dtype: int64
```

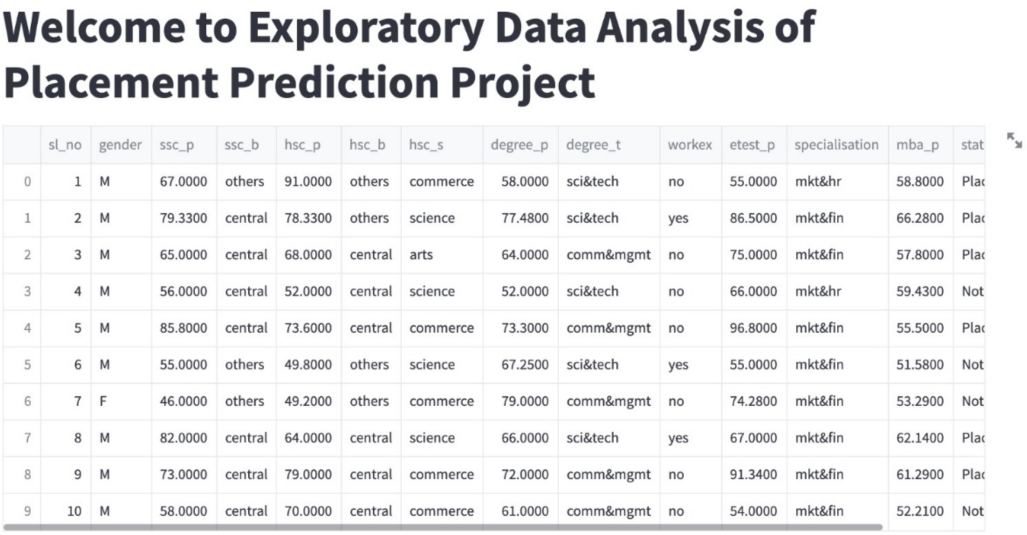
FIG 5.2.1: Student distribution in categories w.r.t 10th and 12th scores



```
In [36]: data['mba_g'].value_counts()
Out[36]:
Average          130
Below Average     81
Good               4
Name: mba_g, dtype: int64

In [37]: data['etest_g'].value_counts()
Out[37]:
Good             89
Average          81
Below Average    45
Name: etest_g, dtype: int64
```

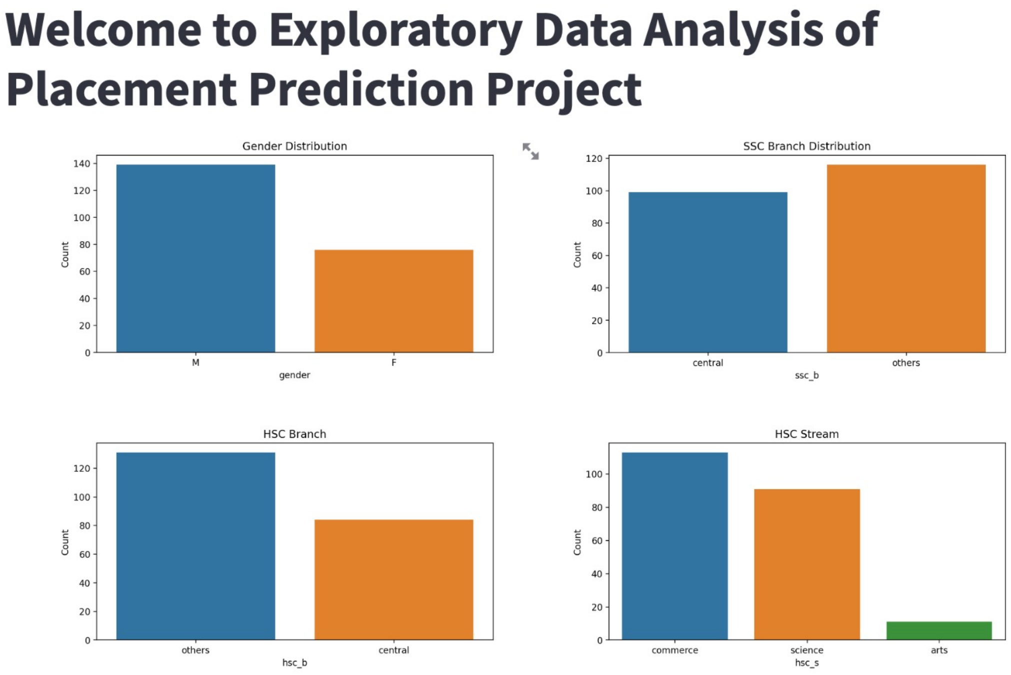
FIG 5.2.2: Student distribution in categories w.r.t MBA and E-Test scores



Welcome to Exploratory Data Analysis of Placement Prediction Project

	sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	stat
0	1	M	67.0000	others	91.0000	others	commerce	58.0000	sci&tech	no	55.0000	mkt&hr	58.8000	Plac
1	2	M	79.3300	central	78.3300	others	science	77.4800	sci&tech	yes	86.5000	mkt&fin	66.2800	Plac
2	3	M	65.0000	central	68.0000	central	arts	64.0000	comm&mgmt	no	75.0000	mkt&fin	57.8000	Plac
3	4	M	56.0000	central	52.0000	central	science	52.0000	sci&tech	no	66.0000	mkt&hr	59.4300	Not
4	5	M	85.8000	central	73.6000	central	commerce	73.3000	comm&mgmt	no	96.8000	mkt&fin	55.5000	Plac
5	6	M	55.0000	others	49.8000	others	science	67.2500	sci&tech	yes	55.0000	mkt&fin	51.5800	Not
6	7	F	46.0000	others	49.2000	others	commerce	79.0000	comm&mgmt	no	74.2800	mkt&fin	53.2900	Not
7	8	M	82.0000	central	64.0000	central	science	66.0000	sci&tech	yes	67.0000	mkt&fin	62.1400	Plac
8	9	M	73.0000	central	79.0000	central	commerce	72.0000	comm&mgmt	no	91.3400	mkt&fin	61.2900	Plac
9	10	M	58.0000	central	70.0000	central	commerce	61.0000	comm&mgmt	no	54.0000	mkt&fin	52.2100	Not

FIG 5.2.3: Dataset Displayed on EDA



Welcome to Exploratory Data Analysis of Placement Prediction Project

FIG 5.2.4: General Distribution of Histogram

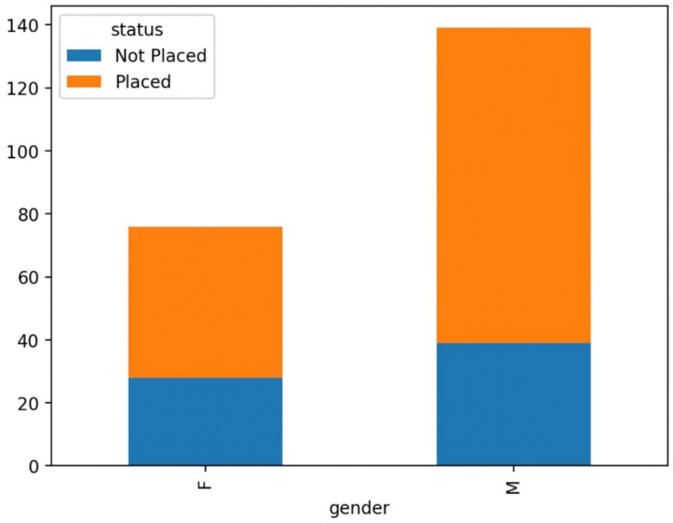


FIG 5.2.5: Histogram of Male Vs Females Placed

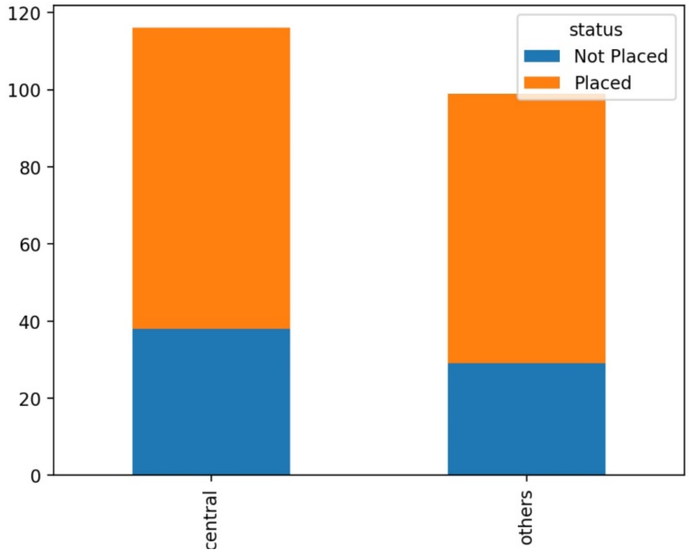


FIG 5.2.6: Histogram of Central Board Vs Other bored

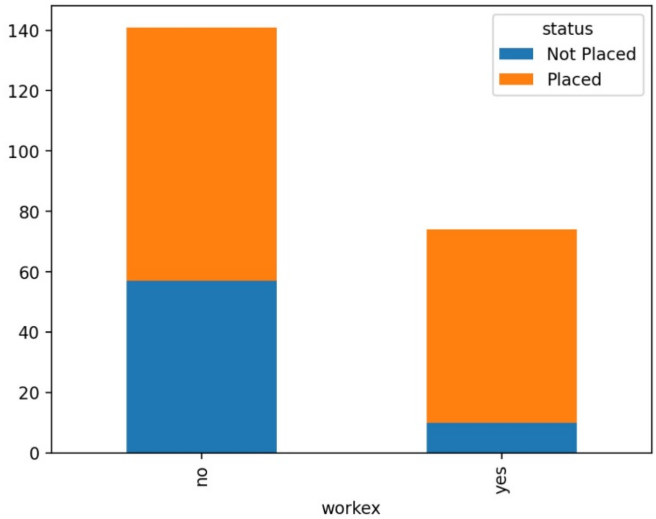


FIG 5.2.6: Histogram of students with Work-Experience Vs No Work-Experience

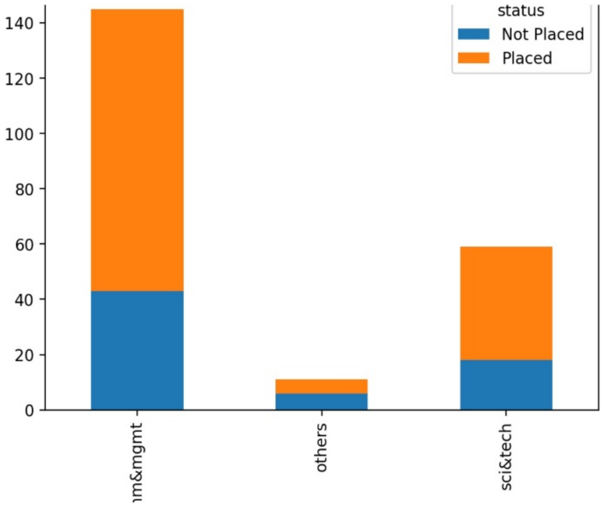
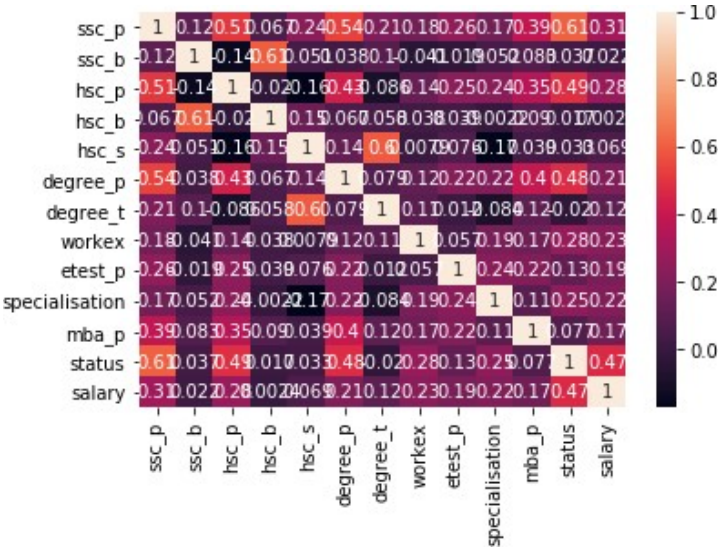


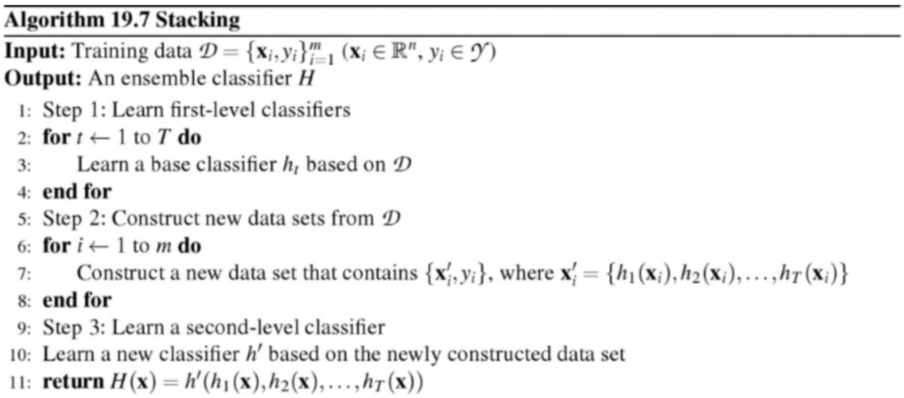
FIG 5.2.6: Histogram of Commers management Vs Other Vs SciTech



5.2.7: Heatmap showing the correlation matrix among the different features

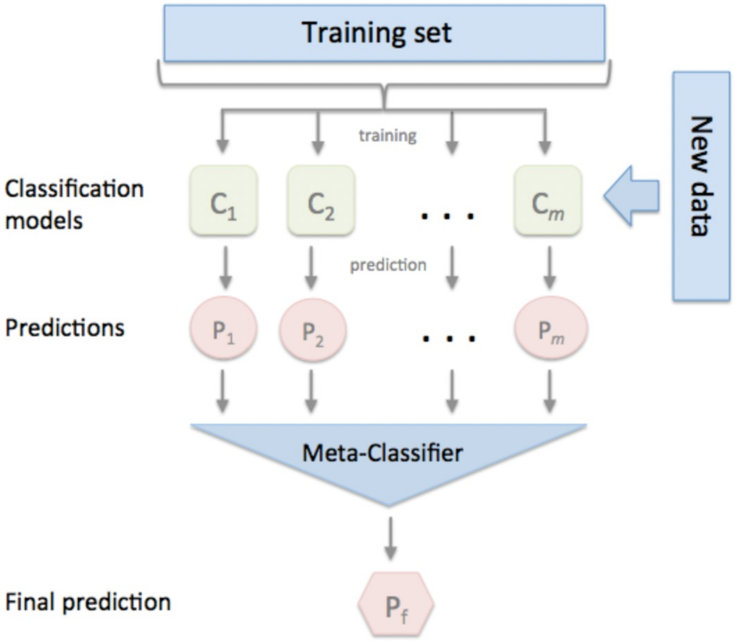
Some of the insights obtained from the visualizations above are:

- Majority of the students have no work experience, whereas very small percentage has some sort of work experience.
- 50.23% of the students fall under the average category when it comes to ssc scores and a rising trend i.e 60.46% of students fall in the average category when it comes to hsc and MBA scores. Similarly in the e-test results 41.39% of the total students have performed good i.e scored above 75% in the e-test.
- The relation between ssc score categories and the placement status, from the histogram an evident observation can be made that 99% of the students who fall in the good category of ssc score have placements whereas 84% of the students falling in the Below Average category score have not been placed yet.
- hsc score categories have a similar relation as the ssc score where 73.8% of students falling in the Average category scores have placements while 76.19% students in the Below Average category are unplaced.



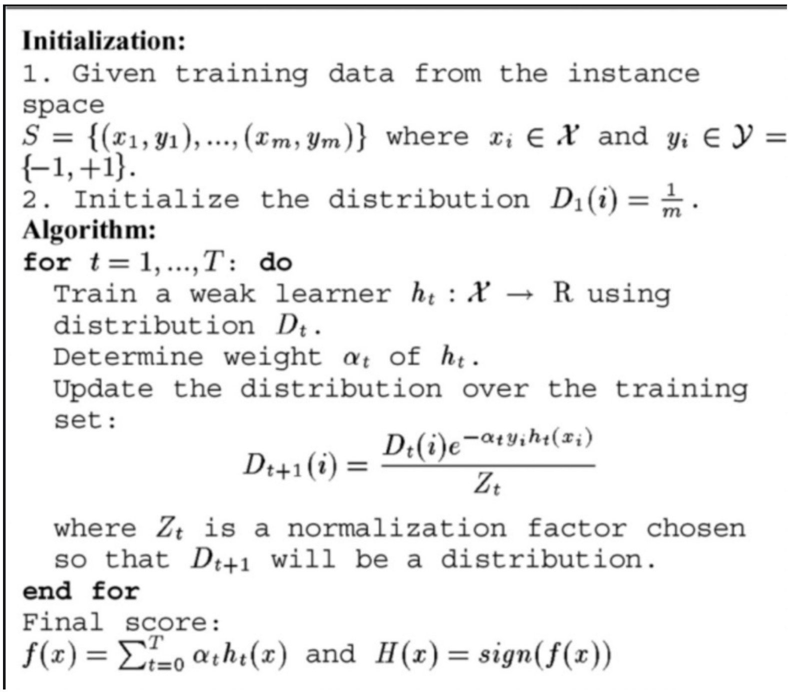
Algorithm 19.7 Stacking

```
Input: Training data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{m}$ ($\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \mathcal{Y}$)
Output: An ensemble classifier $H$
Step 1: Learn first-level classifiers
for $t \leftarrow 1$ to $T$ do
    Learn a base classifier $h_t$ based on $\mathcal{D}$
end for
Step 2: Construct new data sets from $\mathcal{D}$
for $i \leftarrow 1$ to $m$ do
    Construct a new data set that contains $\{\mathbf{x}'_i, y_i\}$, where $\mathbf{x}'_i = \{h_1(\mathbf{x}_i), h_2(\mathbf{x}_i), \ldots, h_T(\mathbf{x}_i)\}$
end for
Step 3: Learn a second-level classifier
Learn a new classifier $h'$ based on the newly constructed data set
return $H(\mathbf{x}) = h'(h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_T(\mathbf{x}))$
```



Stacking is an ensemble learning technique to combine multiple classification models via a meta-classifier. The individual classification models are trained based on the complete training set; then, the meta-classifier is fitted based on the outputs -- meta-features -- of the individual classification models in the ensemble. The meta-classifier can either be trained on the predicted class labels or probabilities from the ensemble.

BONUS



```
Initialization:
1. Given training data from the instance space
$S = \{(x_1, y_1), ..., (x_m, y_m)\}$ where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y} = \{-1, +1\}$.
2. Initialize the distribution $D_1(i) = \frac{1}{m}$.
Algorithm:
for $t = 1, ..., T$: do
  Train a weak learner $h_t : \mathcal{X} \rightarrow \mathrm{R}$ using distribution $D_t$.
  Determine weight $\alpha_t$ of $h_t$.
  Update the distribution over the training set:
      $D_{t+1}(i) = \frac{D_t(i)e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$
  where $Z_t$ is a normalization factor chosen so that $D_{t+1}$ will be a distribution.
end for
Final score:
$f(x) = \sum_{t=0}^{T} \alpha_t h_t(x)$ and $H(x) = sign(f(x))$
```

FIG 5.5 Theoretical Properties of XGBoost

MODULE 3

The module 3 consists of building the model using the below algorithms and evaluating their respecting performance which will assist in finalizing the algorithms for hybridization in the next module.

The algorithms used for model building:

- Gaussian Naïve Bayes
- K-Nearest Neighbors Classifier
- Random Forest Classifier
- XGBoost

The performance evaluation was done for all the above algorithms using metrics like accuracy, confusion matrix, misclassification rate, recall. Out of all the 4 algorithms the XGBoost had the highest accuracy of 89.46%.The lowest accuracy was given by Random forest algorithm which was 83.07%.

MODULE 4

The hybridized algorithm was creating by using the stacking classifier from the MLX tend library and the three algorithms used for stacking were

- XGBoost
- KNN
- Gaussian NB

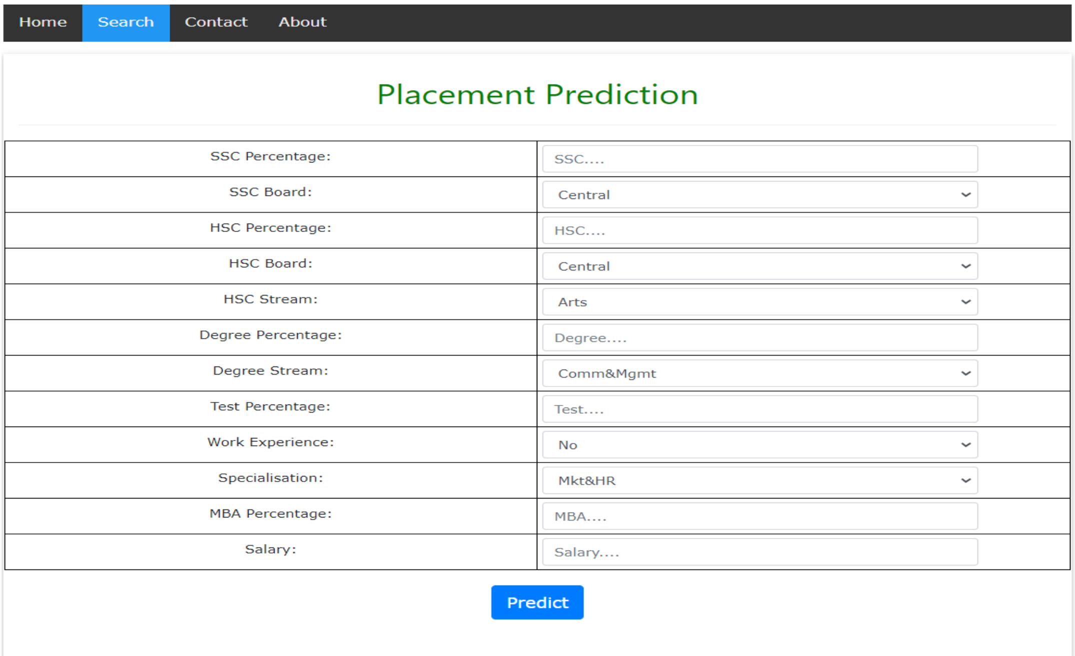
Further this model was tested for performance, and it was observed that the stacked model gave an accuracy of 93.384% which was good enough for deployment over web application.

MODULE 5

The hybrid model was then deployed over a web application using flask framework in python. Which also utilized html, CSS and bootstrap to create templates of different pages of the application and rendered it using flask.

The images below show the GUI of the various pages of the application.

FIG 5.5.1: HomePage



Placement Prediction

SSC Percentage:	SSC....
SSC Board:	Central
HSC Percentage:	HSC....
HSC Board:	Central
HSC Stream:	Arts
Degree Percentage:	Degree....
Degree Stream:	Comm&Mgmt
Test Percentage:	Test....
Work Experience:	No
Specialisation:	Mkt&HR
MBA Percentage:	MBA....
Salary:	Salary....

FIG 5.5.2: Input Fields and Output

CHAPTER 5

TESTING

This document describes the plan for testing the prediction model's performance as well as testing the web application by providing different inputs not merely from the dataset and testing the application.

MODEL PERFORMANCE TESTING

The modules consisted of building the model using the below algorithms and evaluating their respecting performance which will assist in finalizing the algorithms for hybridization.
The algorithms used for model building:

- Gaussian Naïve Bayes
- K-Nearest Neighbors Classifier
- Random Forest Classifier
- XGBoost

The performance evaluation was done for all the above algorithms using metrics like accuracy, confusion matrix, misclassification rate, recall. Out of all the 4 algorithms the XGBoost had the highest accuracy of 89.46%. The lowest accuracy was given by Random Forest algorithm which was 83.07%.

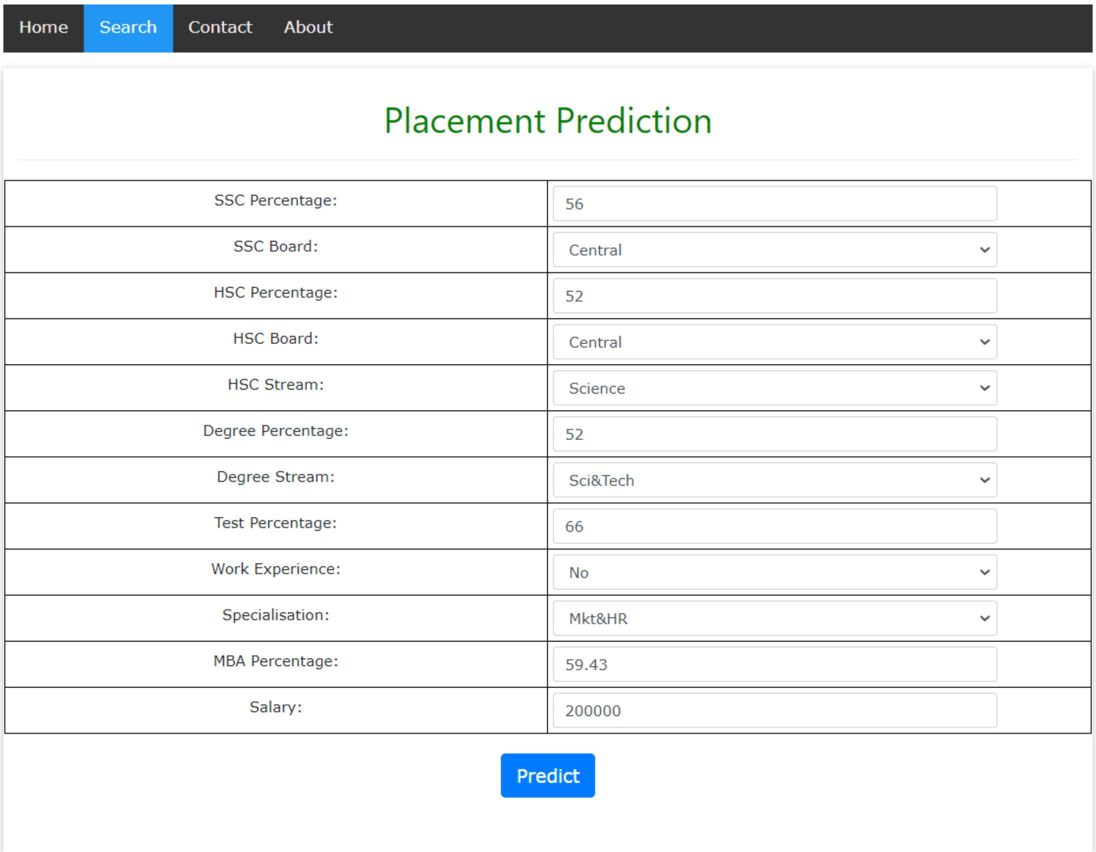
Accuracy is a crucial parameter while dealing with prediction models. It provides the complete picture on how well the model can classify the output into the two respective classes. Confusion matrix gives an idea on how many total records are classified as true positives, false positives, false negatives, and true negatives. The other performance metrics such as precision, recall and misclassification rate can be obtained from the confusion matrix itself. The hybridized algorithm model was created by using the stacking classifier from the MLX tend library and the three algorithms used for stacking were

- XGBoost
- KNN
- Gaussian NB

The hybridized model gives an accuracy of 93.384% which is good for deployment over web application.

INTEGRATION & SYSTEM TESTING

The prediction model which was deployed over the web application using flask was tested using some inputs from the dataset to check the correctness of the output obtained. It was also tested for some values which were not present in the dataset

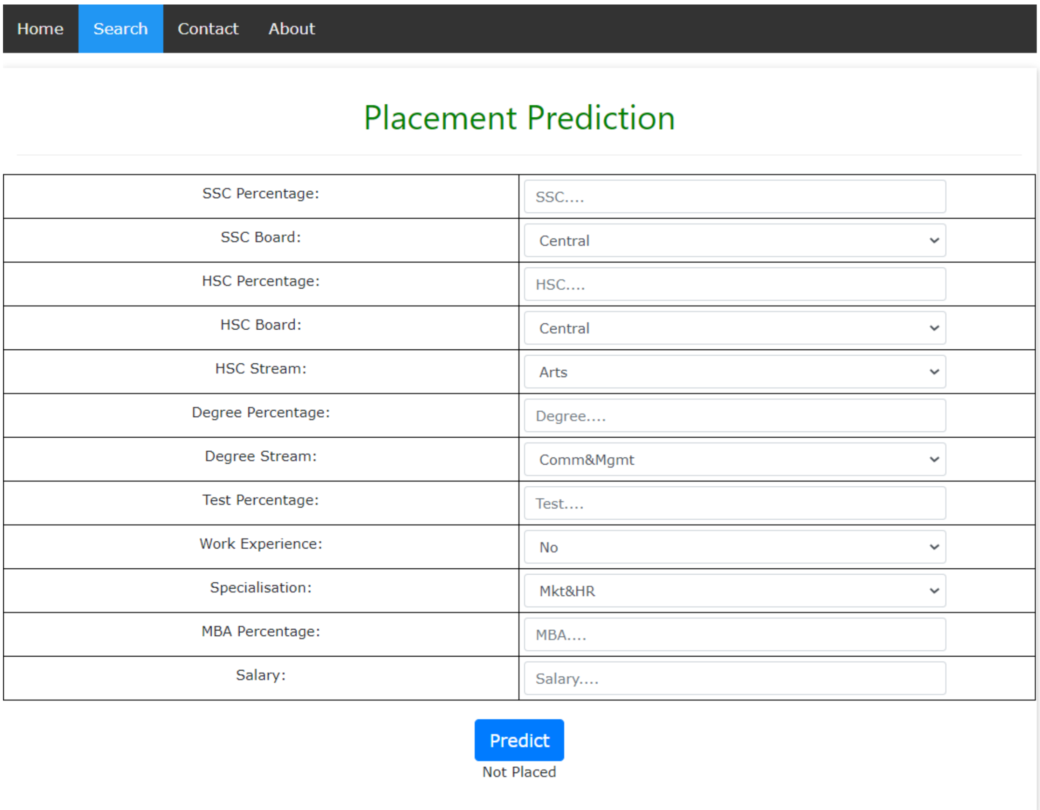


Placement Prediction

SSC Percentage:	56
SSC Board:	Central
HSC Percentage:	52
HSC Board:	Central
HSC Stream:	Science
Degree Percentage:	52
Degree Stream:	Sci&Tech
Test Percentage:	66
Work Experience:	No
Specialisation:	Mkt&HR
MBA Percentage:	59.43
Salary:	200000

FIG 6.2.1: Input taken from the dataset

After providing the above inputs in their respective field values and clicking on the predict button we receive a prediction as shown in the image below. Since, the input values were taken from the dataset, the prediction had to match the value present in the dataset. Since, the output given by the model was identical to the one present in the dataset which indicates that the model was working well and providing great results.

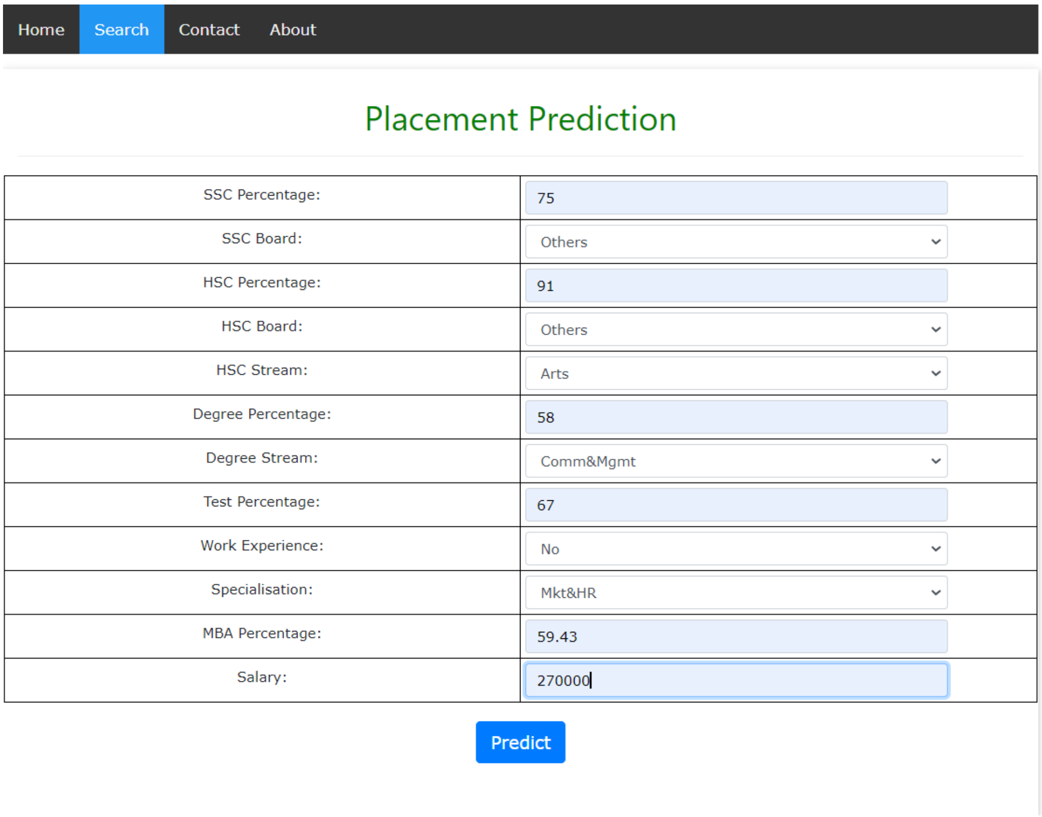


Placement Prediction

SSC Percentage:	SSC....
SSC Board:	Central
HSC Percentage:	HSC....
HSC Board:	Central
HSC Stream:	Arts
Degree Percentage:	Degree....
Degree Stream:	Comm&Mgmt
Test Percentage:	Test....
Work Experience:	No
Specialisation:	Mkt&HR
MBA Percentage:	MBA....
Salary:	Salary....

Not Placed

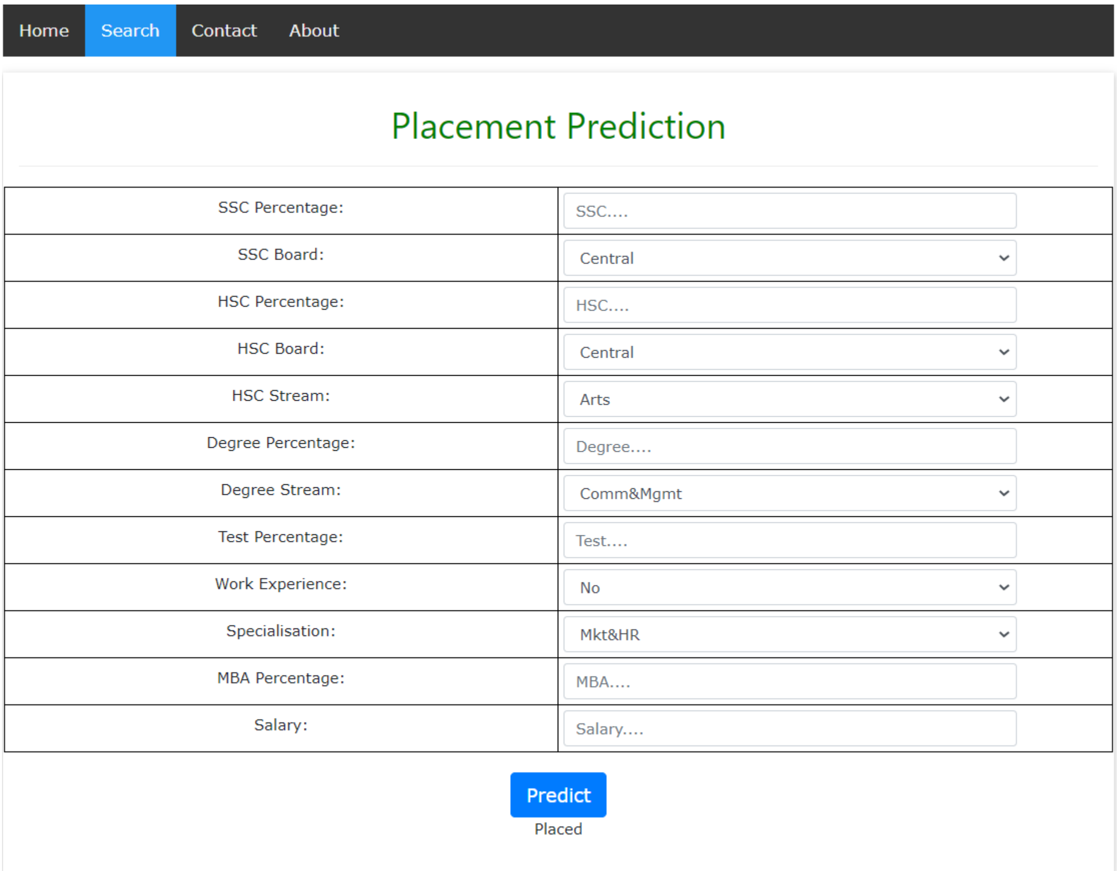
FIG 6.2.2: Output from the model



Placement Prediction

SSC Percentage:	75
SSC Board:	Others
HSC Percentage:	91
HSC Board:	Others
HSC Stream:	Arts
Degree Percentage:	58
Degree Stream:	Comm&Mgmt
Test Percentage:	67
Work Experience:	No
Specialisation:	Mkt&HR
MBA Percentage:	59.43
Salary:	270000

FIG 6.2.3: Random input values taken



Placement Prediction

SSC Percentage:	SSC....
SSC Board:	Central
HSC Percentage:	HSC....
HSC Board:	Central
HSC Stream:	Arts
Degree Percentage:	Degree....
Degree Stream:	Comm&Mgmt
Test Percentage:	Test....
Work Experience:	No
Specialisation:	Mkt&HR
MBA Percentage:	MBA....
Salary:	Salary....

Placed

FIG 6.2.4: Output obtained from the model

CHAPTER 6

CONCLUSION

We have developed a model which can assist students to predict whether they'll be able to get a job or not using various inputs dealing with student's academic record and the field/domain the student comes from. The EDA performed during the making of this project provided our team some important insights within the data itself which will be further used for research purpose. The hybrid model provides a better performance compared to the work of previous researchers we can say that we have achieved our goal.

Since, Machine Learning is quite a vast field there will be researchers coming up with new algorithms which would provide better results and performance compared to the existing work so this process is an ongoing one

CHAPTER 7

REFERENCES

1. K. C. Piad, M. Dumlao, M. A. Ballera and S. C. Ambat, "Predicting IT employability using data mining techniques," 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), Moscow, 2016, pp. 26-30, doi: 10.1109/DIPDMWC.2016.7529358Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).
2. Priam Pillai, Dhiraj Amin,Understanding the requirements Of the Indian IT industry using web scrapping,Procedia Computer Science,Volume 172,2020,Pages 308-313,ISSN 1877-0509
3. Predict the Entrepreneurial Intention of Fresh Graduate Students Based on an Adaptive Support Vector Machine FrameworkJixia Tu,1 Aiju Lin,2 Huiling Chen ,2 Yuping Li ,3 and Chengye Li 3
4. Mishra, Tripti & Kumar, Dharminder & Gupta, Sangeeta. (2016). Students' employability prediction model through data mining. 11. 2275-2282.
5. Y. Bharambe, N. Mored, M. Mulchandani, R. Shankarmani and S. G. Shinde, "Assessing employability of students using data mining techniques," 2017 International Conference on Advances in Computing, Communications, and Informatics (ICACCI), Udupi, 2017, pp. 2110-2114.
6. C. D. Casuat and E. D. Festijo, "Predicting Students' Employability using Machine Learning Approach," 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS), Kuala Lumpur, Malaysia, 2019, pp. 1-5
7. C. D. Casuat and E. D. Festijo, "Identifying the Most Predictive Attributes Among Employability Signals of Undergraduate Students," 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), Langkawi, Malaysia, 2020, pp. 203-206
8. J. Nagaria and S. V. S, "Utilizing Exploratory Data Analysis for the Prediction of Campus Placement for Educational Institutions," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-7
9. C. D. Casuat, A. Sadhiqin Mohd Isira, E. D. Festijo, A. Sarraga Alon, J. N. Mindoro and J. A. B. Susa, "A Development of Fuzzy Logic Expert-Based Recommender System for Improving Students'Employability," 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC), Shah Alam, Malaysia, 2020, pp. 59-62
10. S. G. Thorat, A. P. Bhagat and K. A. Dongre, "Neural Network Based Psychometric Analysis for Employability," 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE), San Salvador, 2018, pp. 1-5

GITHUB LINK

https://github.com/sthakur8417/MLCS584PROJECT.git