**Detailed Project Report**

**Campus Placement Prediction**

|  |  |
| --- | --- |
| Written By | Subhodwip Chatterjee, Mihir Kudale,  Madhu Pincha,Hitesh Suthar |
| Version | 1.0 |
| Date | 31-1-2023 |

**Document Change Control Record**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Reviews

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Reviewer** | **Comments** |
|  |  |  |  |

### Approval Status:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Version** | **Review date** | **Reviewed By** | **Approved By** | **Comments** |
|  |  |  |  |  |

**Index**

|  |  |
| --- | --- |
| **Content.** | **Page No.** |
|  |  |
| 1. Introduction | 4 |
| 1.1 Abstract | 4 |
| 1.2 Machine Learning | 4 |
| 1.3 Problem Statement | 4 |
| 2. Architecture | 5 |
| 2.1 Data gathering | 5 |
| 2.2 Raw Data Validation | 5 |
| 2.3 Data Transformation | 5 |
| 2.4 Data preprocessing | 5 |
| 2.5 Feature Engineering | 6 |
| 2.6 Over sampling | 6 |
| 2.7 Parameter tunning | 6 |
| 2.8 Model building | 6 |
| 2.9 Model saving | 6 |
| 2.10 Git Hub | 7 |
| 2.11 Deployment | 7 |
| 3. Data set description | 7 |
| 4. Implementation and Results | 9 |
| 4.1 Implementation Platform and Language | 9 |
| 4.2 Metrics for Data Modelling | 9 |
| 4.3 Prediction | 10 |
| 5. Conclusion | 10 |
| 7. Q & A | 10 |

# **1. Introduction**

# **1.1 Abstract**

Machine Learning is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and employ algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. These models can be applied in different areas and trained to match the expectations of management so that accurate steps can be taken to achieve the organization’s target. In this paper, the educational records of students are used to predict wheter they get placement or not. Taking various aspects of a dataset collected for Educational instituions, and the methodology followed for building a predictive model, results with high levels of accuracy are generated, and these observations can be employed to strengthen their placement department so as to improve their institution on a whole.

**1.2 Machine Learning**

The data available is increasing day by day and such a huge amount of unprocessed data is needed to be analyzed precisely, as it can give very informative and finely pure gradient results as per current standard requirements. It is not wrong to say as with the evolution of Artificial Intelligence (AI) over the past two decades, Machine Learning (ML) is also on a fast pace for its evolution. ML is an important mainstay of IT sector and with that, a rather central, albeit usually hidden, part of our life. As the technology progresses, the analysis and understanding of data to give good results will also increase as the data is very useful in current aspects.

In machine learning, one deals with both supervised and unsupervised types of tasks and generally a classification type problem accounts as a resource for knowledge discovery. It generates resources and employs regression to make precise predictions about future, the main emphasis being laid on making a system self-efficient, to be able to do computations and analysis to generate much accurate and precise results. By using statistic and probabilistic tools, data can be converted into knowledge. The statistical inferencing uses sampling distributions as a conceptual key.

ML can appear in many guises. In this paper, firstly, various applications of ML and the types of data they deal with are discussed. Next, the problem statement addressed through this work is stated in a formalized way.

## 1.3 Problem Statement

The main goal is to predict whether the student will be recruited in campus placements or not based on the available factors in the dataset

**2. Architecture:**

Following workflow was followed during the entire project.



**2.1 Data gathering:**

Data source: <https://www.kaggle.com/competitions/ml-with-python-course-project/data>

Train and Test data are stored in .csv format.

**2.2 Raw Data Validation:**

After data is loaded, various types of validation is required before we proceed further for any operation. Validations like checking for zero standard deviation for all the columns, checking for complete missing values in any columns, etc. These are required because The attributes which contains these are of no use. It will not play role in predicting placement status for any student.

Like if any attribute is having zero standard deviation, it means that’s all the values are same, its mean is zero. Which indicate that either the Student is placed or not that attribute will remain the same. Similarly, if any attribute is having full missing values, then there is no use of taking that attribute into an account for operation. It’s unnecessary increasing the chances of dimensionality curse.

**2.3 Data Transformation**

Before sending the data into the database, data transformation is required so that data are converted into such form with which it can easily insert into the database. Here there is no missing values in the training data. If there will be any missing values, they are filled in both train set as well as test set with supported appropriate data types.

**2.4 Data preprocessing**

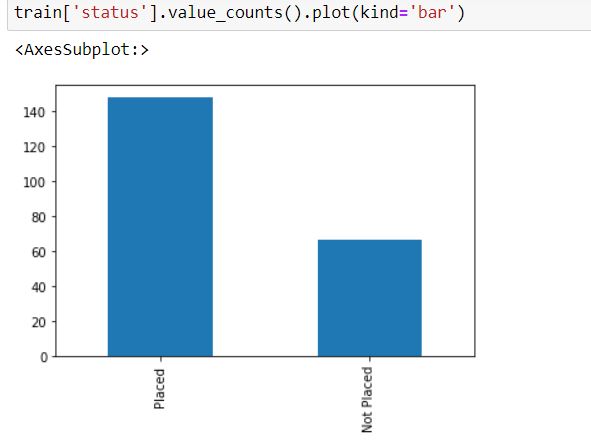
In data preprocessing all the processes required before sending the data for model building are performed. Here, I removed all those columns who standard deviation is zero because they are not contributing anything and those columns also which are of no use in predicting the status of Placement. After that I separate label and features columns from the train and test data.

**2.5 Feature Engineering:**

After preprocessing it was we saw that there are few categorical columns present in the data. For converting the categorical column, I have created the column transformer object and inside it I performed one hot encoding on categorical columns. Afterwars I save the transformer object so that it can be used for transforming the test and prediction data in same way as train data. I have converted the target column categorical values into numerical using label encoder.

**2.6 Over sampling:**

The training data given is imbalance because there are more no. of rows for a particular target class, so this will create situation of imbalance data and will affect the prediction. To handle imbalance data we have done oversampling using smote function to create equal no. of rows for each classes.



**2.7 Parameter tunning:**

I have used sklearn pipeline library with grid search cv to peform hyper parameter tuning. I have used different algorithms in pipeline and set different parameters for their important attributes.

**2.8 Model building:**

After doing all kinds of preprocessing operations mention above and performing scaling and hyperparameter tuning, the data set is passed into 3 models, SVC, XGB and Random Forest. It was found that Radom forest classifier performs best with the highest accuracy score equals 0.89. So ‘Random forest classifier’ performed well in this problem.

**2.9 Model saving:**

Model is then saved using pickle library.

**2.10 Git Hub**

Whole project directory will be pushed into GitHub repository.

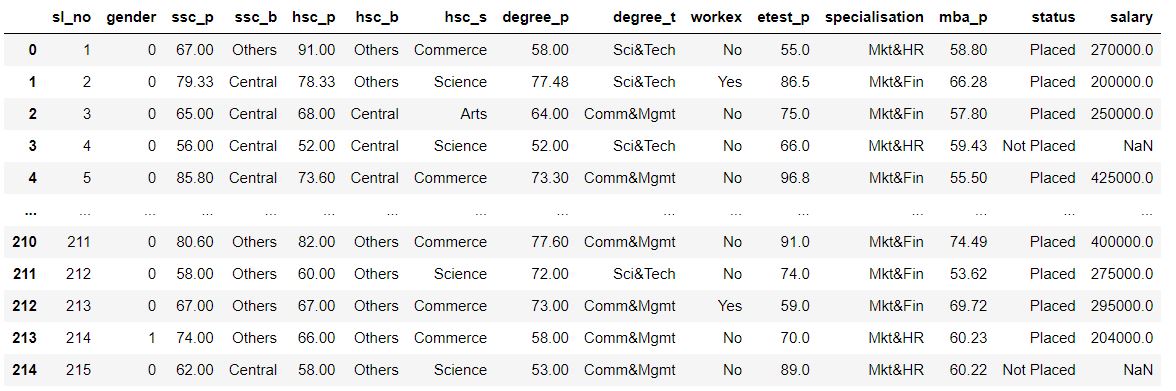
**2.11 Deployment:**

Cloud environment was set up and project was deployed form GitHub into Heroku cloud platform.

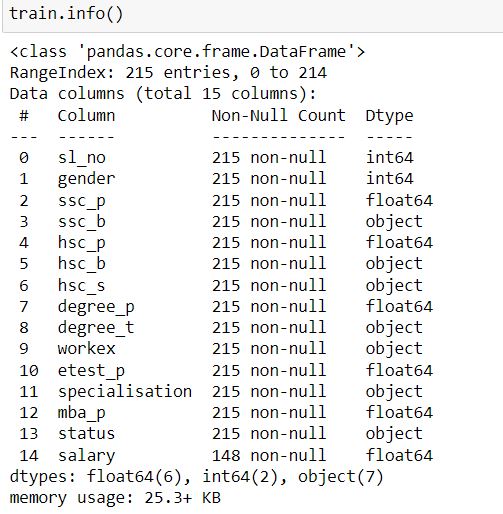
App link[- http://54.236.247.145:8501/](http://54.236.247.145:8501/" \t "_blank)

**3. Data set description**

Data of students for predicting campus placement is collected from various institutions. This data about students contains their percentage at different educational level and which stream they choose in graudation and post graduation. The dataset looks like as follow:



The data set consists of various data types from integer to float to object as shown in Fig.



In the raw data, there can be various types of underlying patterns which also gives an in-depth knowledge about subject of interest and provides insights about the problem. But caution should be observed with respect to data as it may contain null values, or redundant values, or various types of ambiguity, which also demands for pre-processing of data. Dataset should therefore be explored as much as possible.

Various factors important by statistical means like mean, standard deviation, median, count of values and maximum value etc. are shown below for numerical attributes.



Preprocessing of this dataset includes doing analysis on the independent variables like checking for null values in each column and then replacing or filling them with supported appropriate data types, so that analysis and model fitting is not hindered from its way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tells about variable count for numerical columns and model values for categorical columns. Maximum and minimum values in numerical columns, along with their percentile values for median, plays an important factor in deciding which value to be chosen at priority for further exploration tasks and analysis. Data types of different columns are used further in label processing and one-hot encoding scheme during the model building.

# **4. Implementation and Results**

In this section, the programming language, libraries, implementation platform along with the data modeling and the observations and results obtained from it are discusse

## 4.1 Implementation Platform and Language

Python is a general purpose, interpreted-high level language used extensively nowadays for solving domain problems instead of dealing with complexities of a system. It is also termed as the ‘batteries included language’ for programming. It has various libraries used for scientific purposes and inquiries along with number of third-party libraries for making problem solving efficient.

In this work, the Python libraries of Numpy, for scientific computation, and Matplotlib, for 2D plotting have been used. Along with this, Pandas tool of Python has been employed for carrying out data analysis. Random forest is used to solve tasks by ensembling random forest method. As a development platform, Jupyter Notebook, which proves to work great due to its excellence in ‘literate programming’, where human friendly code is punctuated within code blocks, has been used.

## 4.2 Metrics for Data Modelling

* Classification accuracy involves first using a classification model to make a prediction for each example in a test dataset. The predictions are then compared to the known labels for those examples in the test set. Accuracy is then calculated as the proportion of examples in the test set that were predicted correctly, divided by all predictions that were made on the test set.

Accuracy = Correct Predictions / Total Predictions

* AUC means area under the curve so to speak about ROC AUC score we need to define ROC curve first.

It is a chart that visualizes the tradeoff between true positive rate (TPR) and false positive rate (FPR). Basically, for every threshold, we calculate TPR and FPR and plot it on one chart.

Of course, the higher TPR and the lower FPR is for each threshold the better and so classifiers that have curves that are more top-left-side are better.

* Balanced accuracy is a machine learning error metric for binary and multi-class classification models. It is a further development on the standard accuracy metric whereby it's adjusted to perform better on imbalanced datasets, which is one of the big tradeoffs when using the accuracy metric. It is therefore often seen as a better alternative to standard accuracy.

## 4.3 Prediction

The user has give the necessary inputs in the web page like gender,hsc\_p,hsc\_b, degree\_p,degree\_t, workex, etest\_p,specialisation, e.t.c. On the basis of the input from the user the output which is whether the student is palced or not placed is displayed on the screen.

## 5. Conclusion

In this project, basics of machine learning and the associated data processing and modeling algorithms have been described, followed by their application for the task of placement prediction in Educational instituions. On implementation, the prediction results show what are the educational records and marks they to get job placement, Any assistance in this particular area will have a positive impact on an institution’s ability to place its students. This will always be helpful to both the students, as well as the institution.

**6. Q & A:**

**Q1) What’s the source of data?**

Ans. The data for training is provided by the client from:

<https://www.kaggle.com/competitions/ml-with-python-course-project/data>

**Q 2) What was the type of data?**

Ans. The data was the combination of numerical and Categorical values.

**Q 3) What’s the complete flow you followed in this Project?**

Ans. Refer the Architecture section for this.

**Q 4) After the File validation what you do with incompatible file or files which didn’t pass the validation?**

Ans. Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.

**Q 5) What techniques were you using for data pre-processing?**

* + Removing unwanted attributes
  + Visualizing relation of independent variables with each other and output variables
  + Checking and changing Distribution of continuous values
  + Removing outliers
  + Cleaning data and imputing if null values are present.
  + Converting categorical data into numeric values.
  + Scaling the data

**Q 6) How training was done or what models were used?**

* The training and validation data were divided.
* The scaling was performed over training and validation data
* Algorithms like SVC, Random forest, Decision Tree, Logistic Regression, Adaboost

were used.

**Q 7) How Prediction was done?**

Ans. User will fill the required data in the web page on the basis of that prediction is shown on the web page whether he is placed or not palced.

**Q 8) Where the model was deployed?**

Ans. When the model is ready, we deploy our project in streamlit and hosted it on AWS platform. This model is an web application where user can enter the data and these data gets extracted in the backend and user gets the prediction result.