Detailed Project Report

Campus Placement Prediction

Written By	TRISHIT NATH THAKUR
Version	1.0
Date	01-07-2023

1.1 ABSTRACT

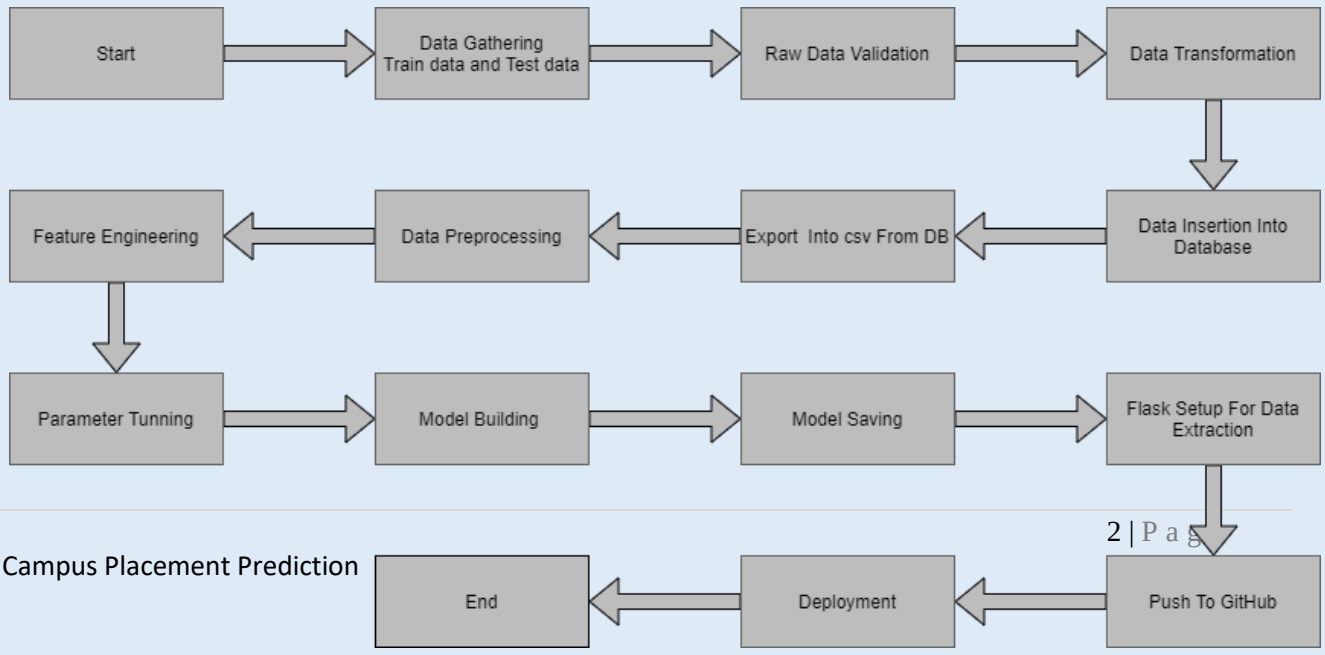
Machine Learning is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and employ algorithms that can receive input data and use statistical analysis to 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 Problem Statement

One of the most crucial goals of an educational institution is student placement. An institution's reputation and yearly admissions are inextricably linked to the placements it offers its students. Because of this, every institution works arduously to enhance their placement department in order to advance the institution as a whole. The capacity of an institution to place its students would be positively impacted by any help in this specific area. Both the institution and the students will always benefit from this.

2. Architecture:

Following workflow was followed during the entire project.



Data gathering the sample data has been collected from our college placement department which consists of all the records of previous year's students. The dataset collected consist of over 1000 instances of students.
Preprocessing Data preprocessing is a technique that is used to convert raw data into a clean dataset. The data is gathered from different sources is in raw format which is not feasible for the analysis. Pre-processing for this approach takes 4 simple yet effective steps.
Attribute selection some of the attributes in the initial dataset that was not pertinent (relevant) to the experiment goal were ignored. The attributes name, roll no, credits, backlogs, whether placed or not, b.tech %, gender are not used. The main attributes used for this study are credit, back-logs, whether placed or not, b.tech %.
Cleaning missing values in some cases the dataset contain missing values. We need to be Page equipped to handle the problem when we come acrossthem. Obviously you could remove the entire line of data but what if you're inadvertently removing crucial information? After all we might not need to try to do that. One in every of the foremost common plan to handle the matter is to require a mean of all the values of the same column and have it to replace the missing data. The library used for the task is called Scikit Learn preprocessing. It contains a class called Imputer which will help us take care of the missing data.
Training and Test data splitting the Dataset into Training set and Test Set Now the next step is to split our dataset into two. Training set and a Test set. We will train our machine learning models on our training set, i.e. our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to examine how accurately it will predict. A general rule of the thumb is to assign 80% ofthe dataset to training set and therefore the remaining 20% to test set.
Feature scaling is the final step of data preprocessing is feature scaling. But what is it? It is a method used to standardize the range of independent variables or features of data. But why is it necessary? A lot of machine learning models are based on Euclidean distance. If, for example, the values in one column (x) is much higher than the value in another column (y), (x2-x1) squared will give a far greater value than (y2-y1) squared. So clearly, one square distinction dominates over the other square distinction. In the machine learning equation.

2.1 Data Description

File descriptions

- train.csv - the training set
- test.csv - the test set
- SampleSubmission.csv - a sample submission file in the correct format Data fields
- gender - sex of the student
- secondary education percentage-marks obtained in secondary education
- higher secondary percentage-marks obtained in higher secondary education
- degree percentage-marks obtained in degree
- Under-graduation(Degree-type)-Field of degree education
- Work-experience • Employability-test-package
- specialization-field of study

Name	Data Type	Measurement
Gender	Integer	Gender of student
ssc_p	Float	Ssc percentage

ssc_b	Object	Ssc board
hsc_p	Float	Hsc percentage
Hsc_b	Object	Hsc board
Hsc_s	Object	Hsc stream
Degree_p	Float	Degree percentage
Degree_t	Object	Graduation stream
Workex	Object	Are they having any work experience
Etest_p	Float	Online test percentage
Specialisation	Object	Specialization the choosed in Mba
Mba_p	Float	Mba percentage
Status	Object	Are they placed or not placed. This the outcome column.

2.2 Data Transformation

In our dataset a lot of categorical values are present, we transformthose attributes into numerical values using one hot encoding. A one hot encoding is a representation of categorical variables as binary vectors. This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.

2.3 Data Preprocessing

Data preprocessing is a technique that is used to convert raw data into a clean dataset. The data is gathered from different sources is in raw format which is not feasible for the analysis. Pre-processing for this approach takes 4 simple yet effective steps. Attribute selection some of the attributes in the initial dataset that was not pertinent (relevant) to the experiment goal were ignored. The attributes name, roll no, credits, backlogs, whether placed or not, b.tech %, gender are not used. The main attributes used for this study are credit, back-logs, whether placed or not, b.tech %. Campus Placement Prediction Cleaning missing values in some cases the dataset contain missing values. We need to be equipped to handle the problem when we come across them. Obviously you could remove the entire line of data but what if you're inadvertently removing crucial information? After all we might not need to try to do that. One in every of the foremost common plan to handle the matter is to require a mean of all the values of the same column and have it to replace the missing data. The library used for the task is called Scikit Learn preprocessing. It contains a class called Imputer which will help us take care of the missing data. to split our dataset into two. Training set and a Test set. We will train our machine learning models on our training set, i.e our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to examine how accurately it will predict. A general rule of the thumb is to assign 80% of the dataset to Training and Test data splitting the Dataset into Training set and Test Set Now the next step is training set and therefore the remaining 20% to test set. Feature scaling the final step of data preprocessing is feature scaling. But what is it? It is a method used to standardize the range of independent variables or features of data. But why is it necessary? A lot of machine learning models are based on Euclidean distance. If, for example, the values in one column (x) is much higher than the value in another column (y), (x2-x1) squared will give a far greater value than (y2-y1) squared. So clearly, one square distinction dominates over the other distinction.

2.4 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.
To handle imbalance data we have performed oversampling using imblearn smote function.
In the same step we also performed feature scaling to bring down every feature on the same page in terms of value range.

2.5 Data Clustering

The campus placement activity is incredibly vital from institution point of view as well as student point of view. In this regard to improve the student's performance, a dataset has been analyzed and predicted using the classification algorithms like KNN algorithm, Random Forest, Naïve Bayes, Logistic Regression, Decision Tree and the SVM algorithm to validate the approaches. Models used: KNN ALGORITHM- K Nearest Neighbor (KNN) is intuitive to understand and an easy to implement the algorithm. It is a versatile algorithm also used for imputing missing values and resampling datasets. Random Forest- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. o Naïve Bayes- Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. o Logistic Regression- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. o Decision Trees- Decision trees use multiple algorithms to decide to split a node into two or more subnodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes. o SVM Algorithm- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. o The goal of the SVM algorithm is to create the best line or decision boundary that can segregate ndimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

2.6 Parameter Tuning

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.

2.7 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.

2.8 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.

PREPROCESSING

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.

```
train.describe()
```

	sl_no	gender	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
count	215.000000	215.000000	215.000000	215.000000	215.000000	215.000000	215.000000	148.000000
mean	108.000000	0.353488	67.303395	66.333163	66.370186	72.100558	62.278186	288655.405405
std	62.209324	0.479168	10.827205	10.897509	7.358743	13.275956	5.833385	93457.452420
min	1.000000	0.000000	40.890000	37.000000	50.000000	50.000000	51.210000	200000.000000
25%	54.500000	0.000000	60.600000	60.900000	61.000000	60.000000	57.945000	240000.000000
50%	108.000000	0.000000	67.000000	65.000000	66.000000	71.000000	62.000000	265000.000000
75%	161.500000	1.000000	75.700000	73.000000	72.000000	83.500000	66.255000	300000.000000
max	215.000000	1.000000	89.400000	97.700000	91.000000	98.000000	77.890000	940000.000000

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 discussed

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.