**Employability Prediction for Graduates: An Overview**

**1)Abstract:**

Nowadays, Graduates' employability is a major concern for the universities, and predicting their employability can help take timely actions to increase the institutional placement ratio. We predict the placement status of a student based on various student attributes using Logistic regression algorithm. Knowing your weaknesses before appearing for placements can help students work in areas that they need to improve to best match the company's skillset.

**2) Introduction:**

This system predicts if a student would be placed or not based on the student’s qualifications, historical data, and experience. This predictor uses a machine-learning algorithm to give the result. The algorithm used is logistic regression. Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X. Talking about the dataset, it contains the secondary school percentage, higher secondary school percentage, degree percentage, degree, and work experience of students. After predicting the result its efficiency is also calculated based on the dataset

**3) Main Steps For Employability Prediction:**

a) Finding suitable dataset: The dataset contains the secondary school percentage, higher secondary school percentage, degree percentage, degree, and work experience of students. After predicting the result its efficiency is also calculated based on the dataset. The dataset used here is in .csv format

b) Dataset preparation and pre-processing: Import the required modules. Using pd.read\_csv to read the dataset then dropped the data columns that had no use in our model. For example, one column that was dropped was Salary. While it can be an important column for other projects, we do not care about how much the students are making when they are employed, we are care about whether they are employed or not. Now before moving forward, we need to pre-process and transform our data. For that, we will use astype() method on some columns and change the datatype to category. Now we will apply codes on some of these columns to convert their text values to numerical values.

c) Split Database:  Now we will split the dataset into train and test data which will be used to check the efficiency later.

d) Train Model:  Now we need to train our model for which we will need to import a file, and then we will create a classifier using sklearn module. Then we will check the accuracy of the model. Once we have trained the model, we will check it giving some random values. To gain a more nuanced understanding of our model’s performance we need to make a confusion matrix. A confusion matrix is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives. To get the confusion matrix it takes in two arguments: The actual labels of your test set y\_test and predicted labels

**4)Employability prediction features:**

Features are identified as factors influencing the success of internship programs. We distinguished three categories of features as follows.

a) Adopted features: we classified the different adopted features into six categories: Hard skills, Soft skills, Demographic features, Extracurricular/co-curricular activities, Professional experience and Internship.

b) Output features: eleven of twenty studies focused on placement, employability, hiring , recruitment, getting a job and working . Rare are those studies that concentrated on employability rate, company or graduation.

c) Sensitive features are the selected features. Most of the studies did not identify the most sensitive features. Internship is considered as sensitive in only three studies. In addition, internship is considered in as the most sensitive variable, followed by specific majors and co-curricular activities. Extracurricular activities are considered in as a sensitive feature. Moreover, mental alertness, manner of speaking, ability to express ideas and self-confidence are sensitive features in. In, the most predictive features are as follows: aptitude and reasoning skills, communication skills, family income status, mentor and quality of teaching in the college

**5)Methods/Techniques for determining sensitive features**:

A number of techniques are used to determine the sensitive features such as

a) Univariate Feature Selection technique,

b) ID3 algorithm,

c) Random Forest,

d)Logistic regression (LR) ,

e) Pearson correlation method and Kandel correlation method, and WEKA feature selection.

**5.1) Random Forest Approach:**

The random forest algorithm can also be thought of as an ensemble method in machine learning. The input to a random forest algorithm is a dataset consisting of records, with attributes. Random subsets of the input are created. On each of the random subset created, a decision tree will be constructed. The final class of a test record will be decided by the algorithm which uses the majority vote technique. Random forest algorithm makes use of the out of bag error technique. Each tree is constructed using the following algorithm:

1. Let the number of training cases be N, and the number of variables in the classifier be M.

2. We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much less than M.

3. Choose a training set for this tree by selecting N times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.

4. For each node in the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.

5. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

**5.2) ID3 algorithm:**

The decision tree technique involves constructing a tree to model the classification process. Once a tree is built, it is applied to each tuple in the database and results in classification for that tuple. The following issues are faced by most decision tree algorithms:

• Choosing splitting attributes

• Ordering of splitting attributes

• Number of splits to take

• Balance of tree structure and pruning

• Stopping criteria

The ID3 algorithm is a classification algorithm based on Information Entropy, its basic idea is that all examples are mapped to different categories according to different values of the condition attribute set; its core is to determine the best classification attribute form condition attribute sets. The algorithm chooses information gain as attribute selection criteria; usually the attribute that has the highest information gain is selected as the splitting attribute of current node, in order to make information entropy that the divided subsets need smallest . According to the different values of the attribute, branches can be established, each branch to create other nodes and branches until all the samples in a branch belong to the same category. To select the splitting attributes, the concepts of Entropy and Information Gain are used

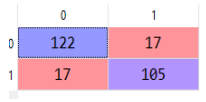
**5.3) Logistic Regression:**

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for given set of features (or inputs), X. Talking about the dataset, it contains the secondary school percentage, higher secondary school percentage, degree percentage, degree, and work experience of students. After predicting the result its efficiency is also calculated based on the dataset

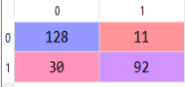
**6) Evaluating the Employability Prediction Systems:**

The campus placement activity is incredibly a lot of vital as institution point of view as well as student point of view. In this regard to improve the student’s performance, a work has been analyzed and predicted using the classification algorithms Decision Tree and the Random forest algorithm and Logistic Regression to validate the approaches.

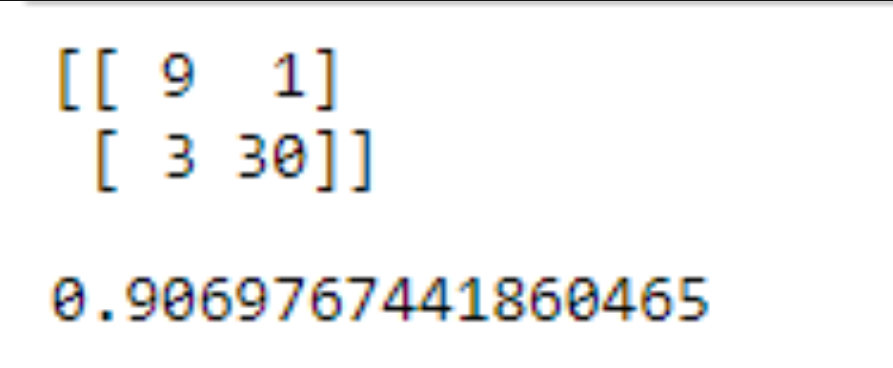
Confusion Matrix for Random Forest:



Confusion Matrix for Decision Tree Algo:



Confusion Matrix for Logistic Regression:



|  |  |
| --- | --- |
| Algorithm | Accuracy |
| Random Forest | 86% |
| Decision Tree | 84% |
| Logistic Regression | 90.69% |

**7) Conclusion:**

The algorithms are applied on the data set and attributes used to build the model. The accuracy obtained after analysis for Decision tree is 84% and for the Random Forest is 86% while Logistic Regression has 90.69% accuracy. Hence, from the above said analysis and prediction it’s better if the Logistic Regression algorithm is used to predict the placement results.