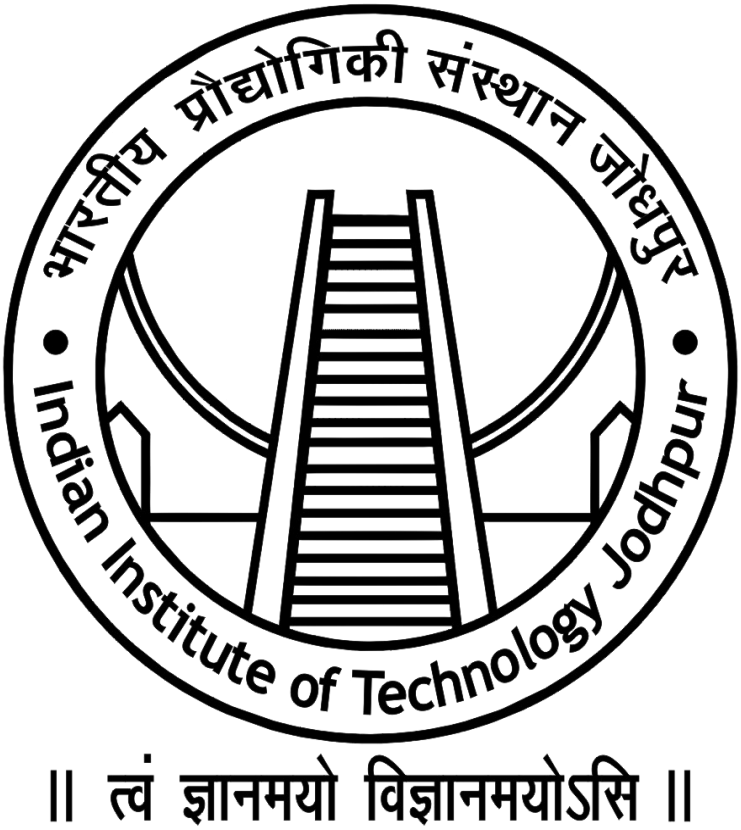
CSL 7550: MACHINE LEARNING ASSIGNMENT 1



Binary Classification Implementation using Linear Programming, Logistic Regression, and Perceptron Algorithm.

Nirdosh (M21MA009)

Indian Institute of Technology Jodhpur

1. Problem Statement

Choose an appropriate dataset of your choice such that every record (example) has at least 5 features which are numeric in nature and there is at least one attribute (feature) which is binary in nature. You can use the binary attribute as the binary target label to be predicted. In case you want to use a target variable which has more than two distinct values, then you can map them into two sets and give label 1 to one of the sets and 0 to the other. Thus, a multiclass classification task can be reduced to binary classification task.

Split your dataset into a training set and a test set. You can try different splits: 70:30 (70% training, 30% testing), 80:20 or 90:10 split.

On the training set, train the following classifiers:

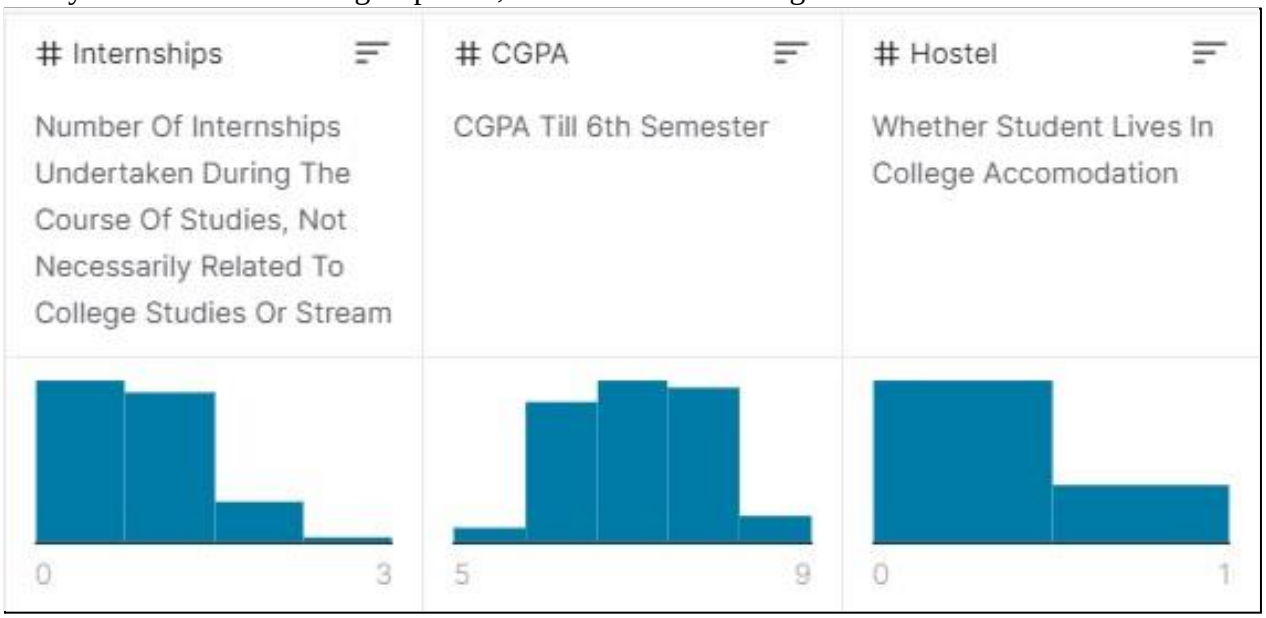
1. Half-Space classifier implemented using LP solver (one such solver is scipy.optimize.linprog)
2. Half-Space classifier implemented using Perceptron Algorithm (implement the iterations)
3. Logistic Regression Classifier

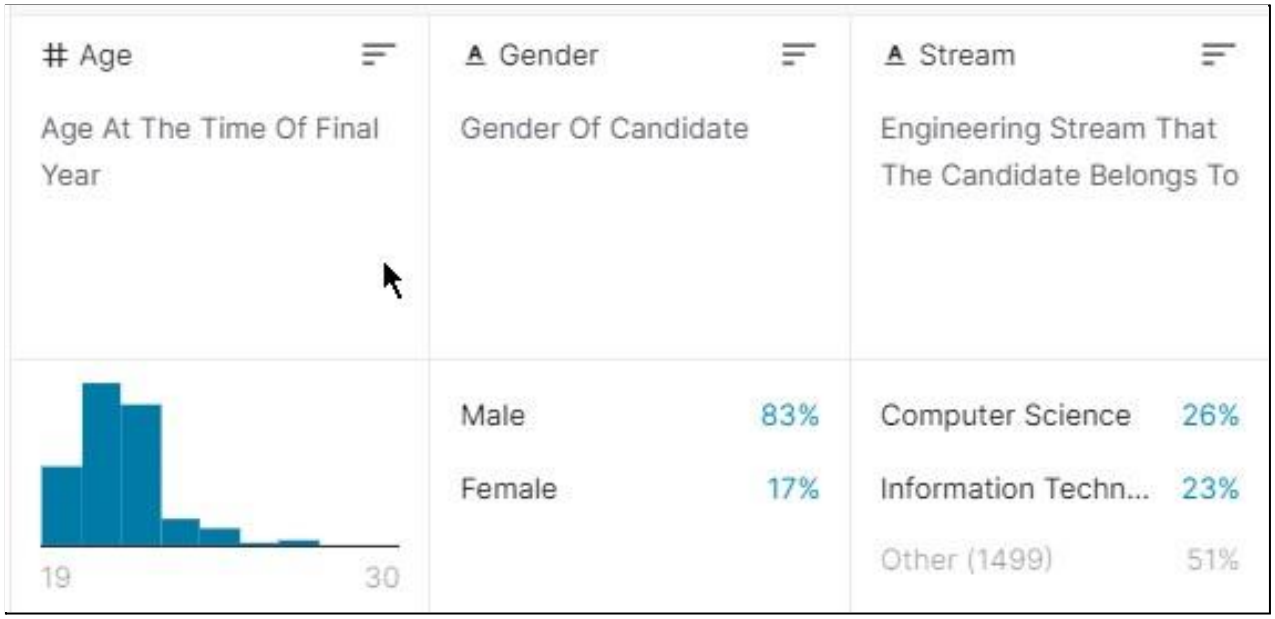
2. Dataset

Description

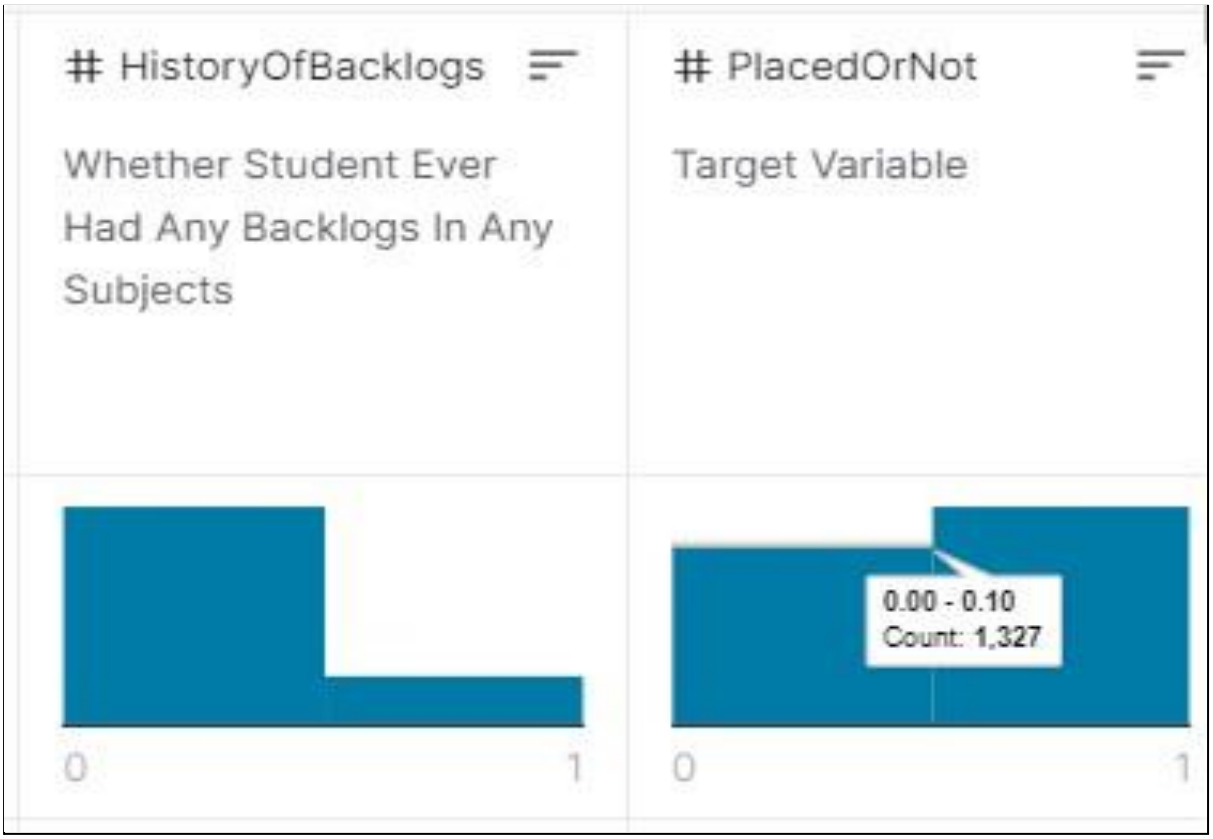
A University made its on-campus placement records public for the world to see. The data is from the years 2013 and 2014.

The following is the college placements data compiled over 2 years. Use this data to predict and analyze whether a student gets placed, based on his/her background.





3



Link

https://www.kaggle.com/tejashvi14/engineering-placements-prediction

CSL 7550: MACHINE LEARNING ASSIGNMENT 1

Classification Task

The task at hand is to predict whether a student will be placed or not given his/her details (features) like *Age*, whether he did an *Internship*, and *No. of backlogs*, etc. We use different methods to design the model. These are *1. Linear Programming (Maximisation Problem)*,
2. Logistic Regression, 3. Perceptron Algorithm

Features and Targets

The features used for the classification task are *Age*, *Gender*, *Stream*, *Internships*, *CGPA*, *Hostel*, *History of Backlogs*.

Here the target is *PlacedOrNot* set consists of binary (boolean) values i.e. (0,1).

```
[171]: data = pd.read_csv('collegePlace.csv')
       data.head() # Here PlacedOrNot is the target label or 'y'
```

[171]:

	Age	Gender	Stream	Internships	CGPA	Hostel	HistoryOfBacklogs	PlacedOrNot
0	22	Male	Electronics And Communication	1	8	1	1	1
1	21	Female	Computer Science	0	7	1	1	1
2	22	Female	Information Technology	1	6	0	0	1
3	21	Male	Information Technology	0	8	0	1	1
4	22	Male	Mechanical	0	8	1	0	1

```
[172]: data['PlacedOrNot'].unique()
```

```
[172]: array([1, 0], dtype=int64)
```

3. Model and Predictions

Accuracy on Different Train-Test-Splits

1. ***LP Solver***
    - ***70-30 Split:*** Accuracy is 50%
    - ***80-20 Split:*** Accuracy is 58.33%
    - ***90-10 Split:*** Accuracy is 58.33%

2. ***Logistic Regression (Self Built)***
    a. ***70-30 Split:***
        i. Train Accuracy is 49.48%
        ii. Test Accuracy is 53.125%
    b. 80-20 Split:
        i. Train Accuracy is 49.48%
        ii. Test Accuracy is 53.125%
    c. ***90-10 Split:***
        i. Train Accuracy is 49.47%
        ii. Test Accuracy is 51.22%
3. ***Logistic Regression (sklearn)***
    a. ***70-30 Split:***
        i. Train Accuracy is 73.09%
        ii. Test Accuracy is 73.52%
    b. ***80-20 Split:***
        i. Train Accuracy is 73.42%
        ii. Test Accuracy is 74.47%
    c. ***90-10 Split:***
        i. Train Accuracy is 73.6%
        ii. Test Accuracy is 78.125%

4. ***Perceptron***

a. ***70-30 Split:***

i. Train Accuracy is 50.67%

ii. Test Accuracy is 48.78%

b. ***80-20 Split:***

i. Train Accuracy is 53.4%

ii. Test Accuracy is 54.167%

c. ***90-10 Split:***

i. Train Accuracy is 50.11%

ii. Test Accuracy is 58.33%