**SUMMER INTERNSHIP – 2020 – DATASCIENCE**

C R DEEPAK KUMAR

2ND YEAR M.Sc. DECISION COMPUTING SCIENCES

COIMBATORE INSTITUTE OF TECHNOLOGY

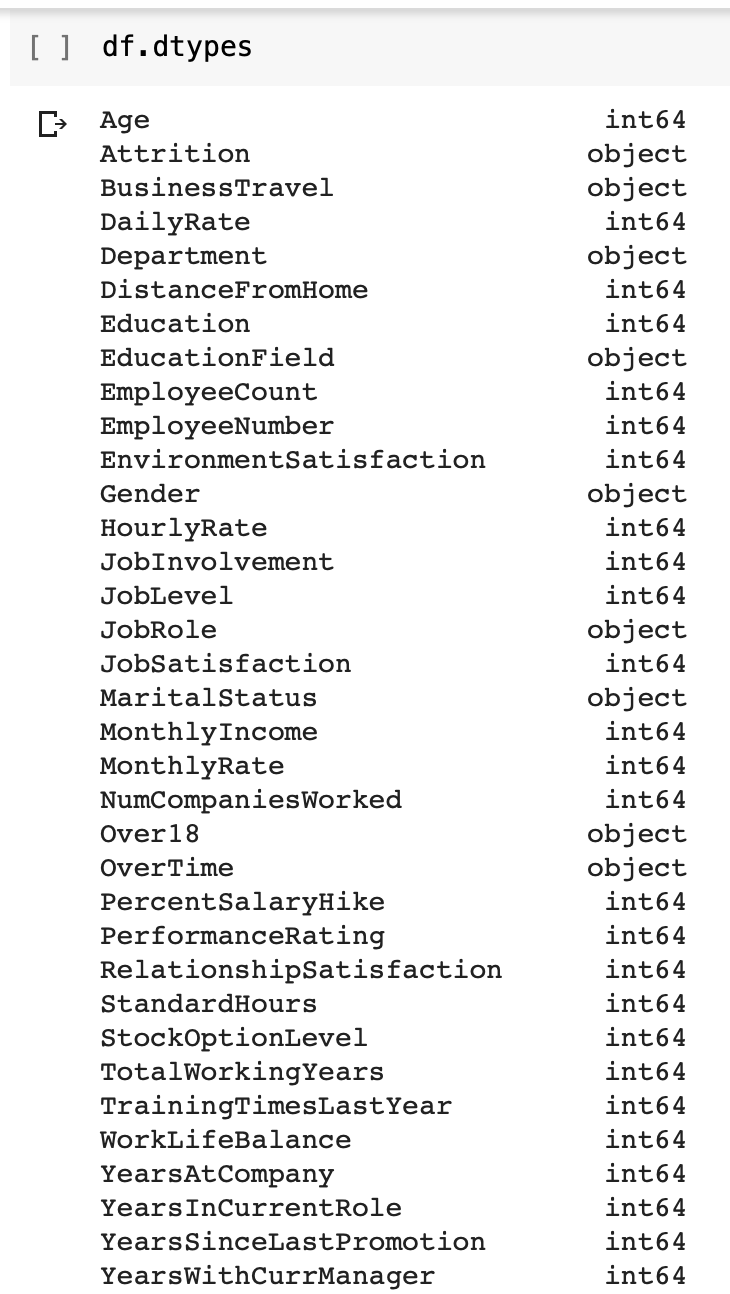
*TASK -1:*

1. *Checking if there are any missing values present in the dataset.*



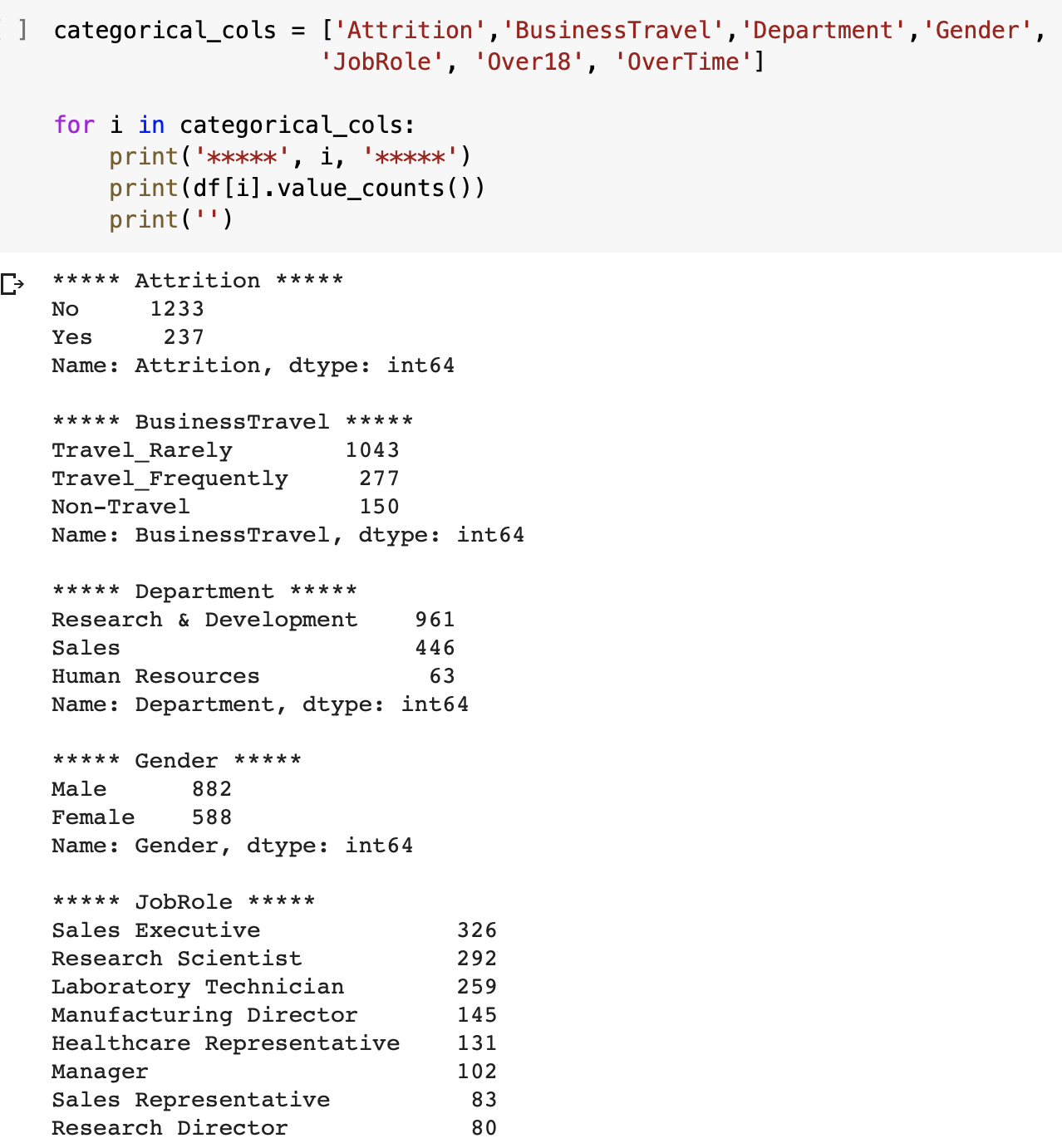
Using ‘df.isnull().sum()’ it shows me the if there any missing values present. The result as shown above is zero for all columns. Therefore, we can say that there are no missing values present in the dataset.

1. *Getting the categorical variables present in the dataset.*



I have to first understand how many categorical variables are there. I can get the object variables using ‘df.dtypes’ which shows me the result as above. The object variables are known as the categorical variables present in the dataset.

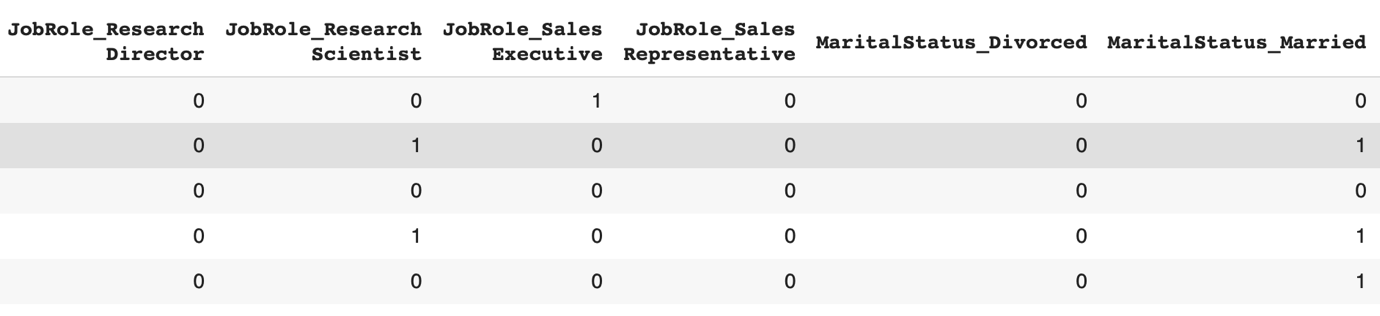
1. *Understang the types of variables present in the categorical coloumns.*



The result shows the types of variables present in each of the categorical columns. We can use this result to make appropriate visualizations of the dataset, in a meaningful manner.

1. *Create the Dummy variables to Categorical to Numerical data*

Using ‘**df=pd.get\_dummies(df)’** we can convert the categorical variables to Numerical data so that the code can process it.



Here is the example of the converted dataset.

*TASK - 2:* ***Attrition\_Visual.ipynb***

When we Visualize the Dataset, we need to understand the Result that we need.

We need to understand why the Employee Attrition occurs in the first place.

The parameters for the occurrence of Employee Attrition are:

1. Less pay compared to Industry Standards
2. Lack of recognition

From the dataset provided to me I chose the following variables to analyze the dataset.

1. Job Level
2. Job Satisfaction
3. Performance Rating
4. Training times Last year
5. Percentage Salary hike converted to ‘%hike<17’ i.e. Percentage of Salary hike less than 17%, I converted it to a True & False format.
6. Monthly Income converted to ‘Income<4919’ since, Rs. 4919 was the median value for Monthly Income received by employees. I chose median & not mean because mean gets affected by the outliers but median does not. I converted it to a True & False format.
7. Total Working Years converted to ‘Work Years<15’ i.e. employees who have worked for less than 15 years. I converted it to a True & False format.
8. Years Since Last promotion converted to ‘Promotion<7’ i.e. employees who haven’t received promotion for the past seven years or less. I converted it to a True & False format.

THE CODE USED TO CREATE THE CHARTS IS:



Only the hue value changes for all the other charts.

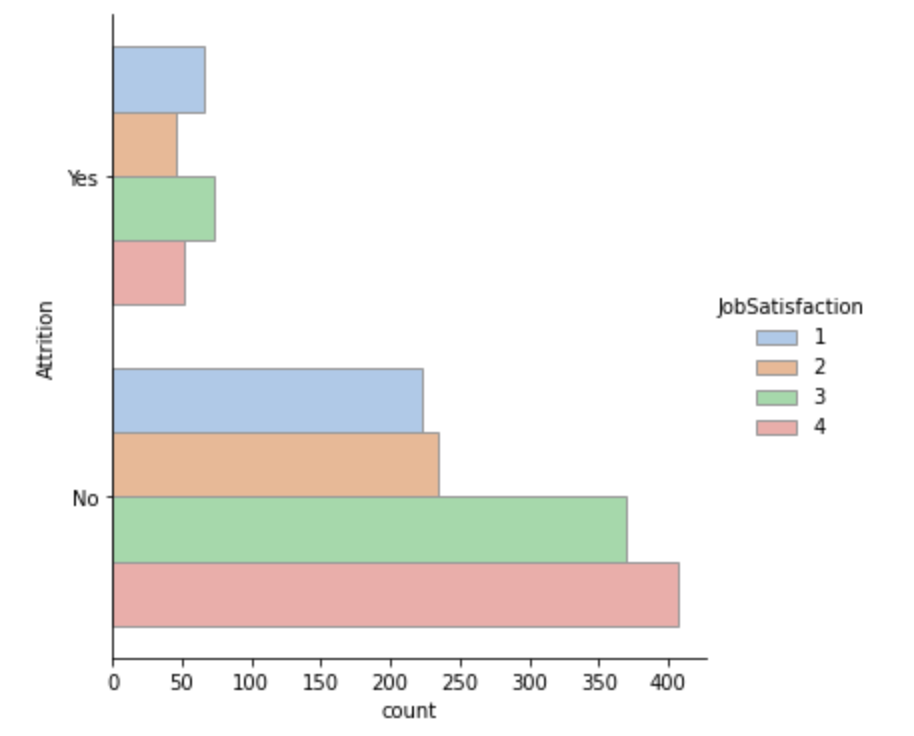
**Catplot** shows frequencies (or optionally fractions or percents) of the categories of one, two or three categorical variables. The first named variable is innermost on the display; that is, its categories vary fastest. Catplot is a wrapper for graph hbar (default) or graph bar or graph dot.

1. ATTRITION VS JOB LEVEL



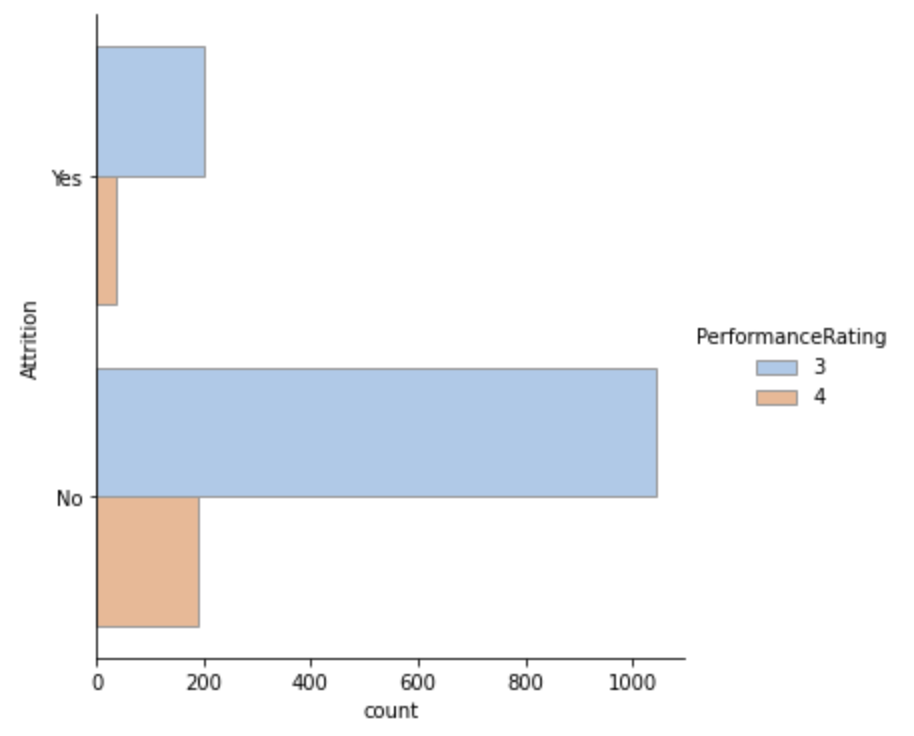
You can see here, we have used Job level to identify and understand which job level was let go and which retained.

Here, Job level-1 was let go the most, while Job level – 2, was retained the most.

1. ATTRITION VS JOB SATISFACTION

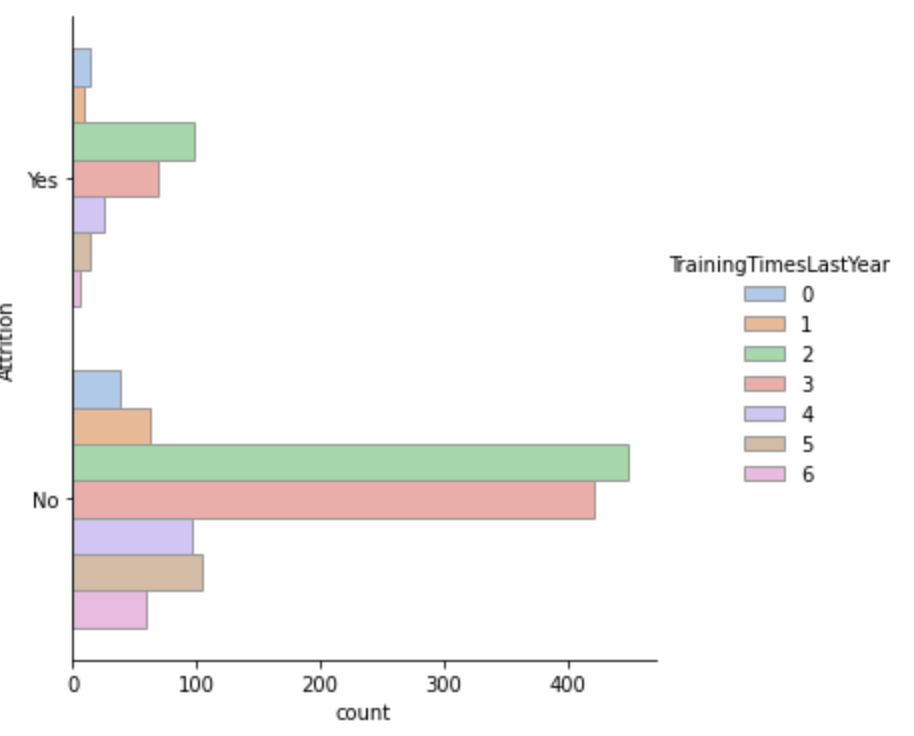
As expected Job Satisfaction with 4 was retained the most while surprisingly Job Satisfaction with 3 was let go the most.

1. ATTRITION VS PERFORMANCE RATING



As expected Performance Rating with 4 was let go the least, we can’t really make a comparison here for the retained employees since the Performance rating with 3 are larger than Performance rating with 4.

1. ATTRITION VS TRAINING TIMES LAST YEAR



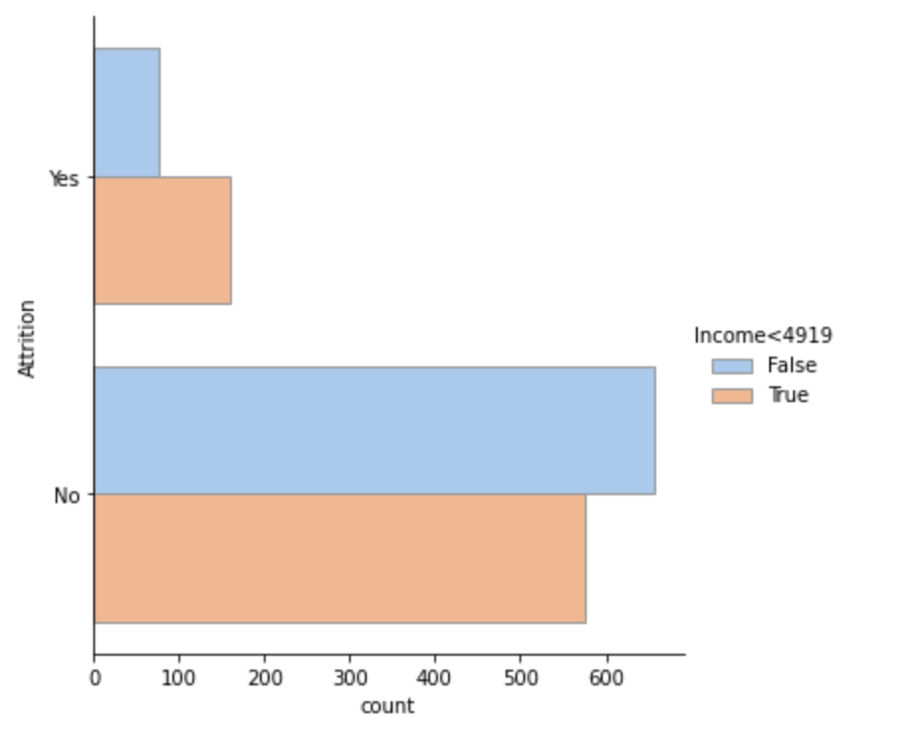
The employees with highest amount of training taken last year have been retained the most and subsequently let go the least.

1. ATTRITION VS %HIKE<17



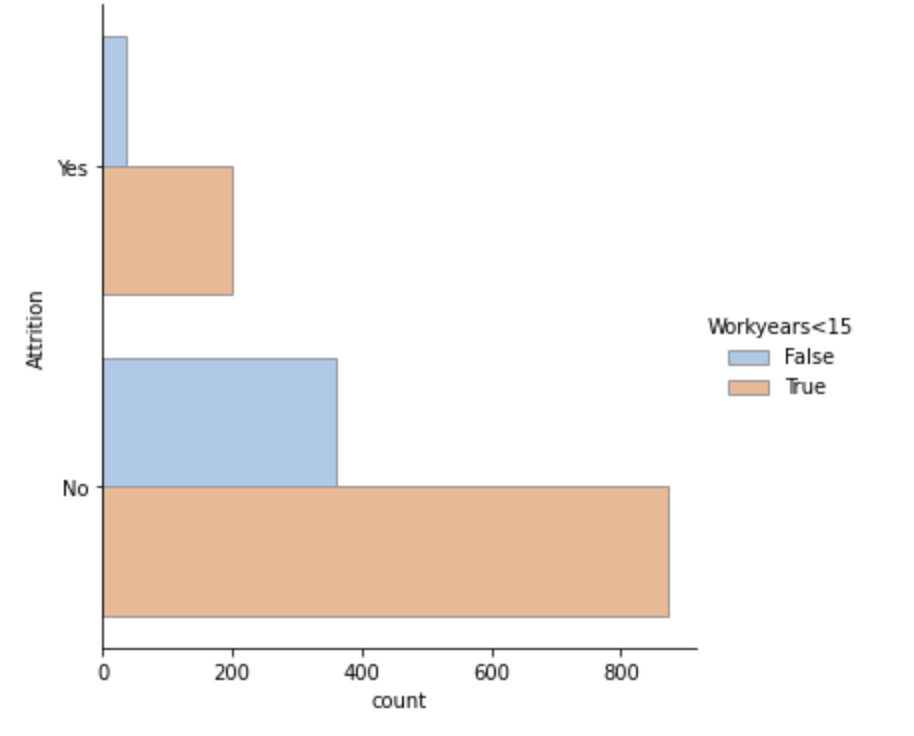
It certainly does not make sense as to why the employees who have received a salary hike of more than 17% be let go. Why the Salary hike then if they were going to leave. These are some of the questions the company can answer through this analysis.

1. ATTRITION VS INCOME<4919



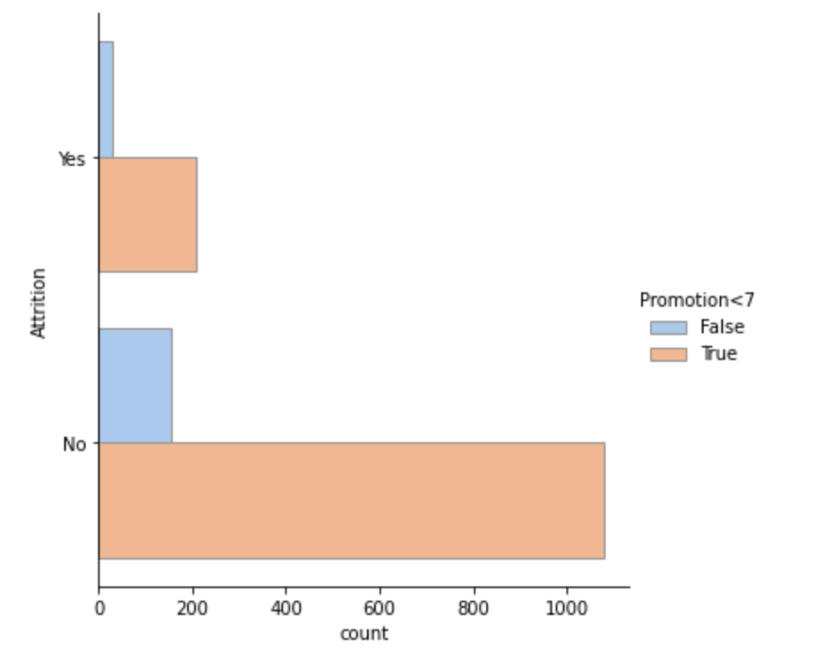
As expected Employees who receive high income have been retained compared to employees who receive less.

1. ATTRITION VS WORK YEARS<15



The employees who have worked for more than 15 years have been let go because many of them could be retiring. While the younger have as expected retained.

1. ATTRITION VS PROMOTION<7

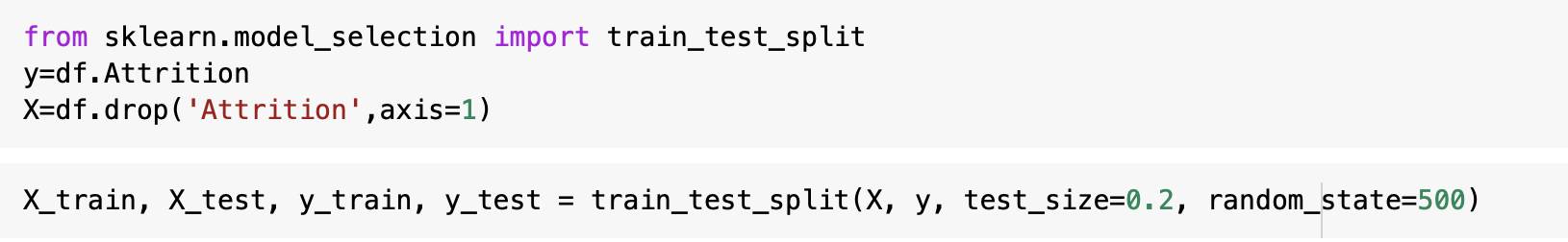


As expected employees who received their promotion recently have been retained while the others have been let go.

*TASK – 3:* ***Attrition\_pred.ipynb***

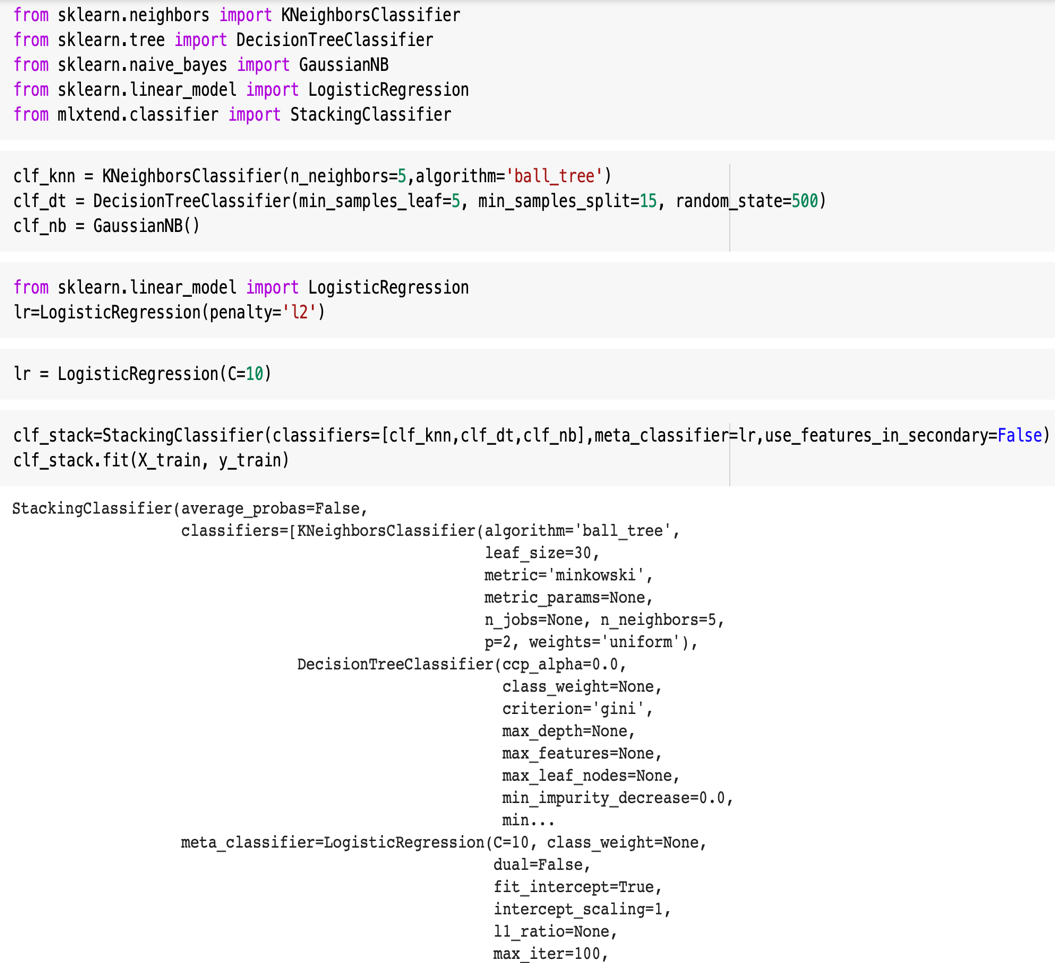
Since all the categorical variables have been converted to Numerical dataset we can now fit the model.

1. VALIDATION TECHNIQUE – TRAIN/TEST SPLIT



I have split the dataset using train/test to validate the model which I will obtain.

1. ENSEMBLE METHODS USED TO CREATE MODEL
   1. STACKING

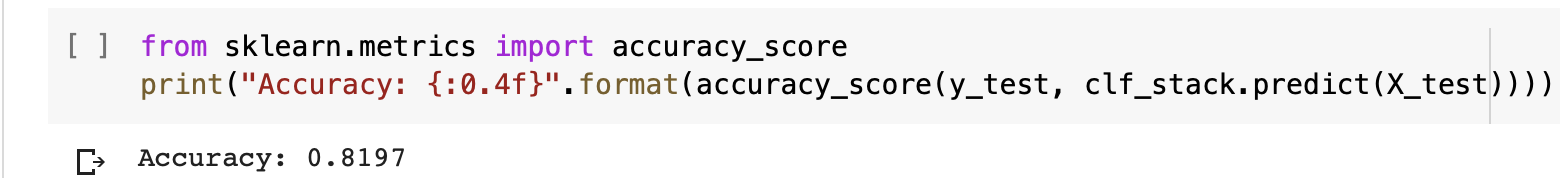


*Stacking:* The idea of stacking is to learn several different weak learners and combine them by training a meta-model to output predictions based on the multiple predictions returned by these weak models. So, we need to define two things in order to build our stacking model: the L learners we want to fit and the meta-model that combines them.

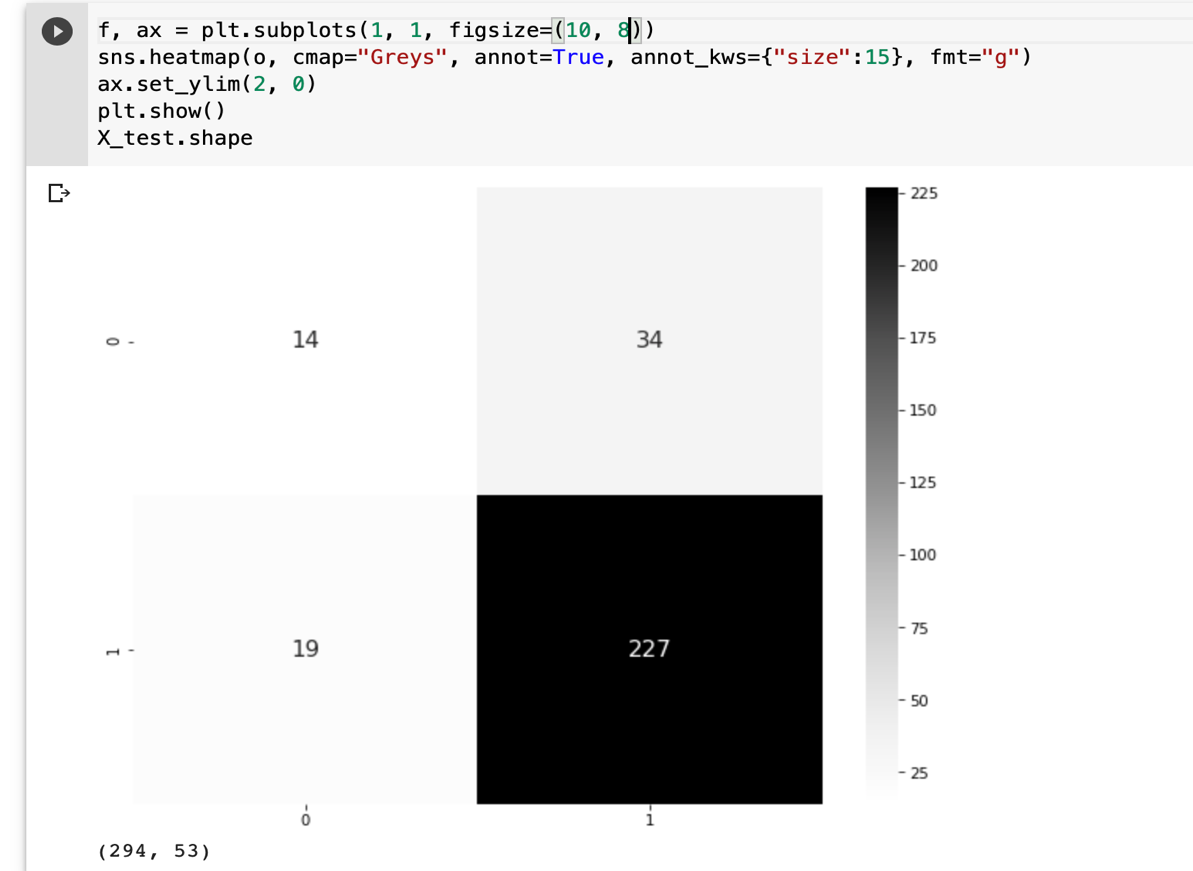
The various models used here are KNN, Decision Tree, Naïve

Bayer’s, and the meta model is Logistic regression.

* VISUALIZATION USING STACKING ENSEMBLE METHOD

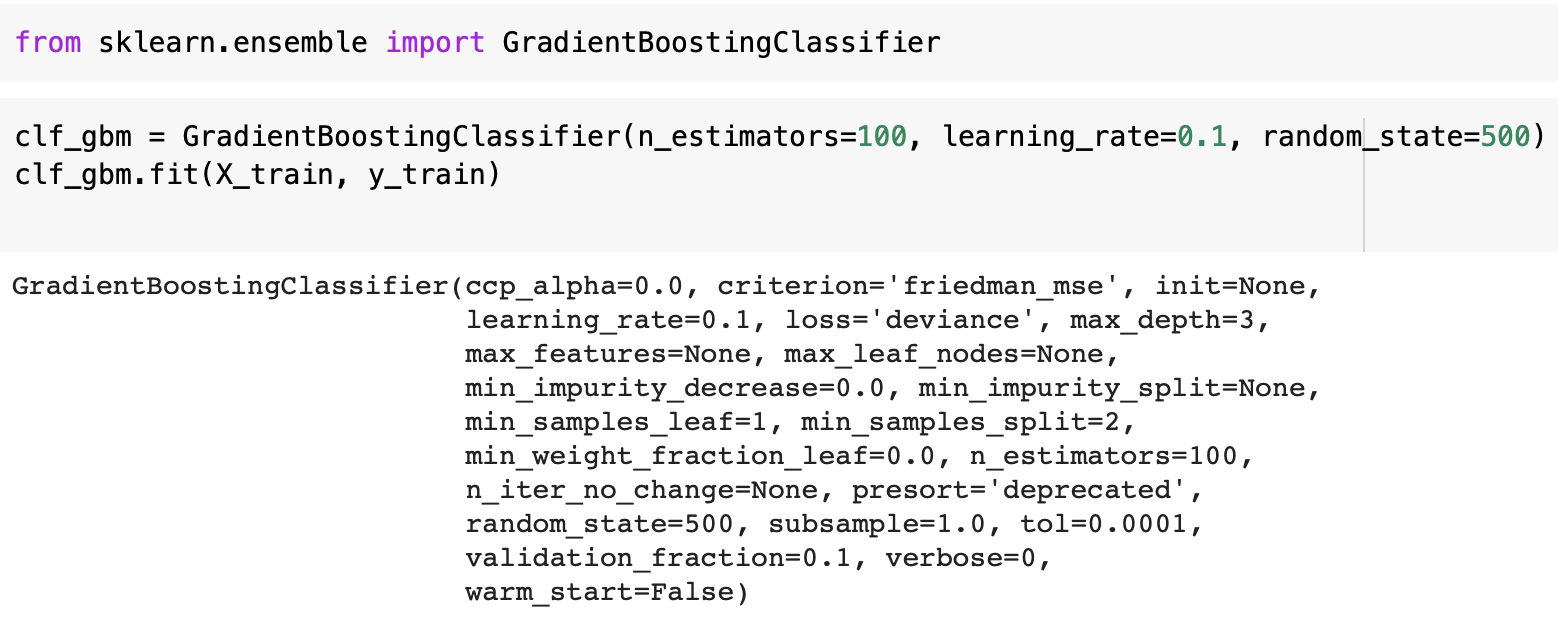


As you can see the accuracy score is 82%.



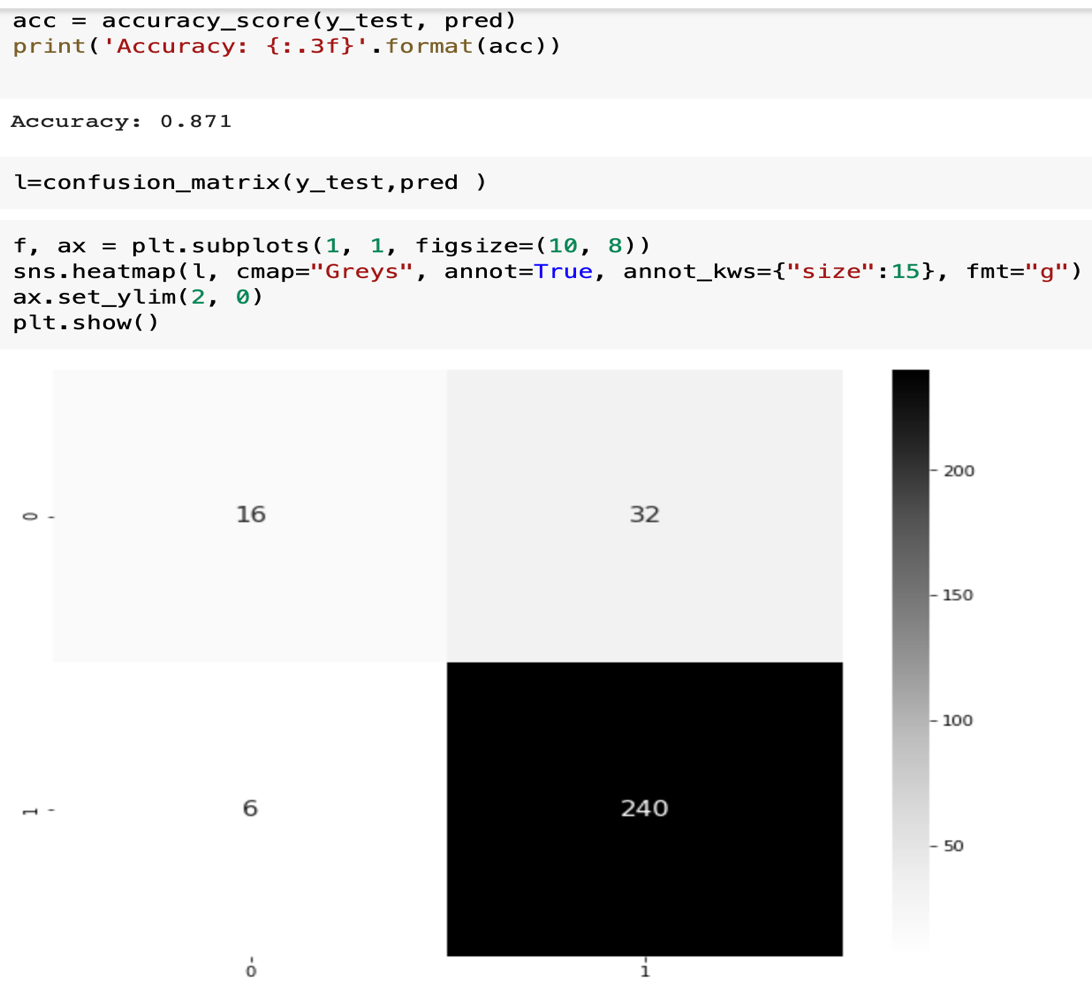
The confusion shows me how many true positives and true negative predictions made.

* 1. GRADIENT BOOSTING ALGORITHM



*Gradient:* It builds trees one at a time and learns from them accordingly to create the best one possible.

* VISUALIZATION OF THE GRADIENT BOOSTING MODEL.



Accuracy score is shown to be 87.1%.

The Confusion Tree also Visualizes the result showing the true positives and true negatives.

*TASK – 4:*

CONCLUSION:

# **FROM THE ACCURACY RESULTS WE CAN CONCLUDE THAT THE GRADIENT BOOSTING METHOD (GBM) IS A BETTER MODEL COMPARED TO STACKING ENSEMBLE MODEL. I TRIED THE RANDOM FOREST ALGORITHM (RFA) AS WELL IT GAVE ME AN ACCURACY OF 87.4% BUT I STILL THINK THAT GBM MODEL IS BETTER BECAUSE GBM LEARNS FROM ITS PREVIOUS TREES WHILE RFA JUST AVERAGES THE RESULTS OBTAINED FROM ITS TREES. GBM MODEL IN THE FUTURE CAN LEARN FROM OTHER DATASETS AND GIVE A MORE CONVINCING PERFORMANCE COMPARED TO RFA**