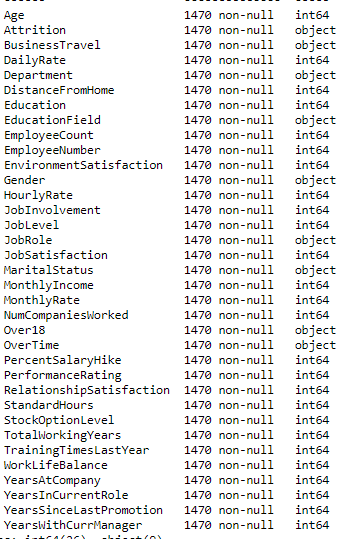
PROBLEM: We need to find how does HR Analytics help in analyzing attrition, Attrition in human resources refers to the gradual loss of employees over time. In general, relatively high attrition is problematic for companies.  We need to classify on the basis of other attributes if a employee will stay or leave the organization.

Dataset: There are 1470 rows and 35 columns,



We can see that dataset is a combination of both categorical and continuous variables will no null values.

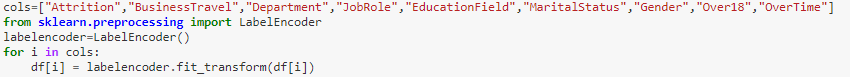
EDA:

After visualizing we draw the below conclusions:

* People who are in late 20's and mid 30's are more prone to stay in the organization.
* People who are at job level 4 and 5 (i. e higher position) are to have higher working years also peopl leaving the organization are mostly have higher job level and working years
* employees with higher work life balance are more prone to stay than the one with less work life balance
* People who travel rarely are to stay more in the organization
* people who are married are more longer to stay they need more job security and the one leaving the organization consist more singles
* Higher the salary, more prone to stay.
* people who stretch are to to leave easily than the one who dont

We used libraries like seaborne and matlpotlib for our visualization.

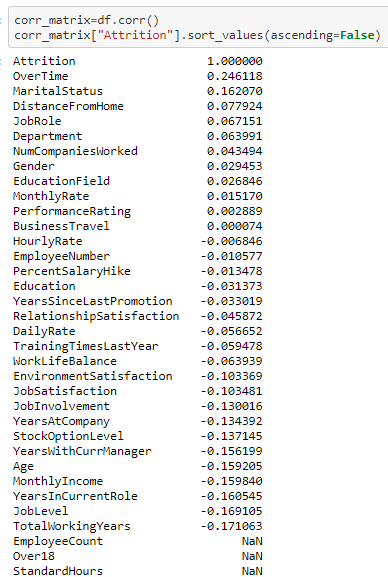
Encoding:Now encoding the categorical columns which to use to find the correlation .



As there is more than one column to encode so we can run a loop for the same.

PREPROCESSING:

Correlation: Find the relationship of other columns with the target variable,”Attrition”.

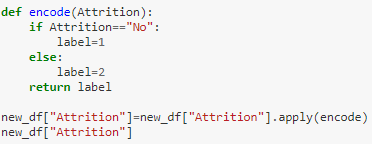


From above we will drop columns which have no correlation or very less correlation with the target variable, ’Attrition’.

We are creating a new data frame with useful columns.

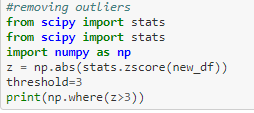
Screenshot 2021-01-14 203349.png

Now, we need to convert the categorical value ‘Attrition’ to find the correlation and model evaluation.



Skewness- Now we will treat skewness, by using log function .

Outliers- We need to find the vales which have z- score above 3.

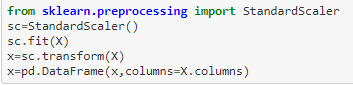


Drop all the values with z-score greater than 3.

Screenshot 2021-01-14 204133.png

Dividing the dataframe into input and output, for input datatframe we will have all columns except the target variable and for the output we will just have the target variable.

Standard Scalar- Using standard scalar to scale all the columns at the same level which will help in better classification.

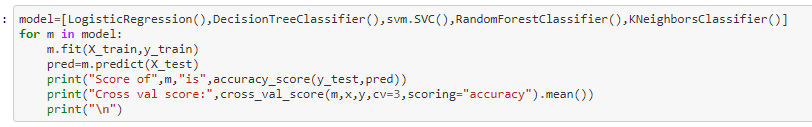


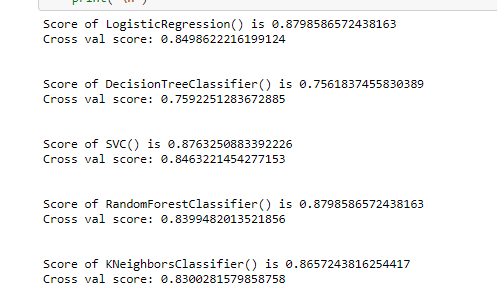
BUILDING MODEL:

Splitting x and y into train and test dataset training dataset will fit the dataset and test dataset will help us for prediction and measuring output. We usually split the dataset into 80 and 20 where 80% is the train dataset and 20% is the dataset.

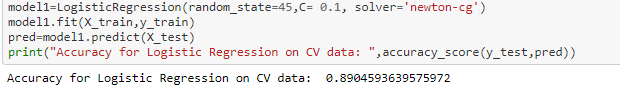
Importing model – Importing different models and then find which has better accuracy score and use it for model evaluation

Creating a loop with different classification models.

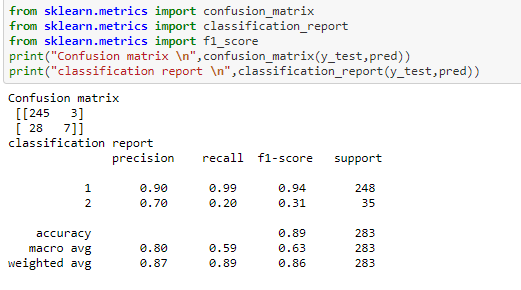




The scores for Random Forest Classifier and Logistic Regression, however cross val score is better for Logistic Regression. Now we will use Grid Search CV and will find the best parameters which will increase the accuracy.



Using hyperparmeter tuning our accuracy scores increased to 89 , we will save this model and use metrics like Confusion\_matrix, Classification\_Report for evaluation our predictions.



From the above we can see that the true positive is 245, true negative is 7, false positive is 3 and false negative is 28.

REMARKS:

This was the classification problem with 35 columns to find whether an employee is to stay in the organistaion or not, we converted the categorical value to continuous and then find columns which has better correlation with the target variable and use it further for prediction. Spllitted the dataset into x and y after treating skewness and outliers and then further spliited the dataset into train and test. We finalized Logistic Regression as it has better accuracy and cross val score. We save the predicted output in a csv file and the model using joblib.