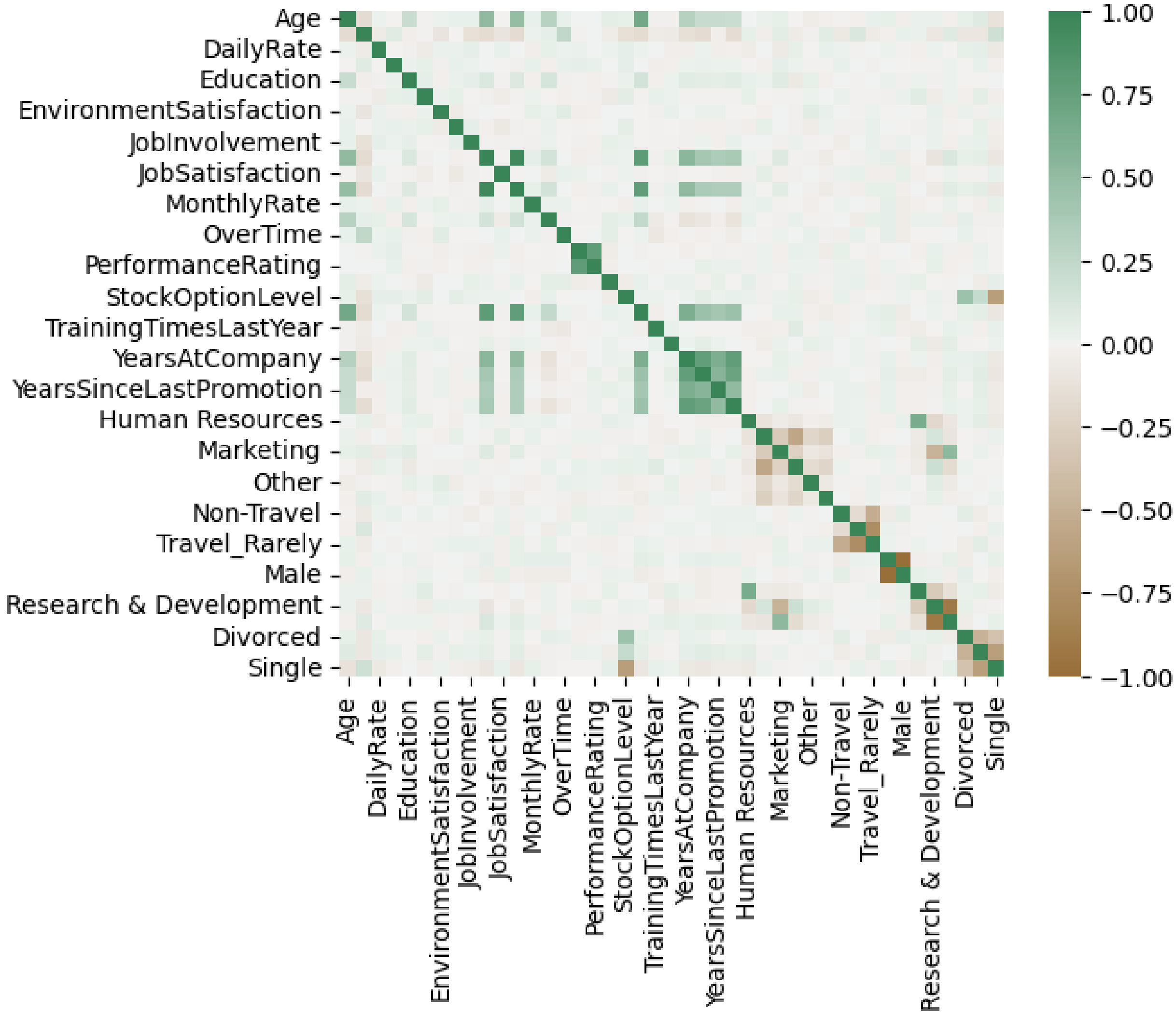
# HR PROJECT FLIGHT RISK – IBM DATASET

The dataset that we observed from the open source provided by IBM to analyze the risk of flight risk in the employee and observe the features or characteristics that contribute to attrition.

A visual inspection of the grid-based pair-plot among all the predictor variables and the heatmap of correlation values can provide a clear indication of which variables to be dropped from further analysis. Highly correlated (ideally 0.9 above) predictor variables become redundant for the decision models and hence one of such variable pairs to be dropped from the data.



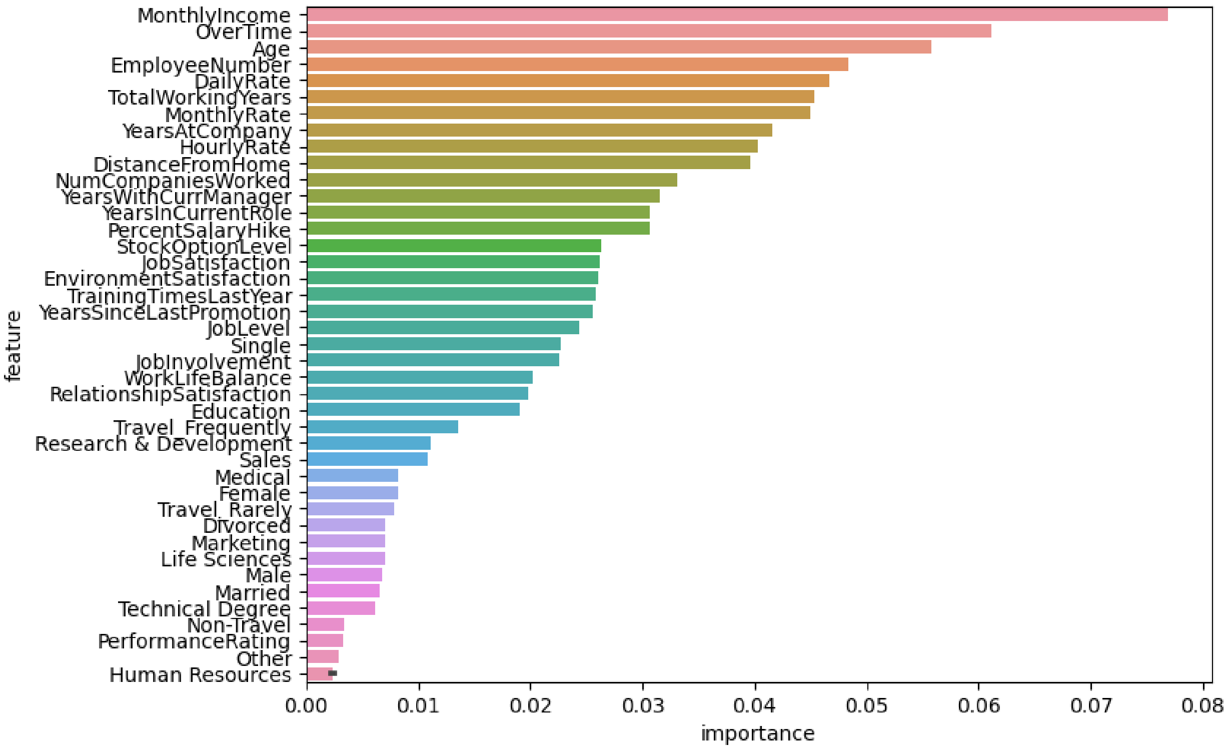
## Model Building

There are various pre-built algorithms are available in the form of decision tree structure. Most of these problems can be solved with a classification algorithm for this problem we can use random forest classifier with parameter tuning with the help of a decision tree at the end to identify the features that are important to consider to find attrition factors.

Before applying the chosen model, we must first see if we have any categorical data that can be encoded using one hot encoder to change categorical to an ordinal variable.

Features such as marital_status, Jobrole, Department, Gender and the other categorical variables are converted to ordinal variables. Once we have changed the data to numerical data and have identified attrition feature which is our target variable and other features are the independent variables.

As the new variables we have observed now is 35 columns such high columns can hinder our model accuracy and require more computational power, this problem can also be solved by finding the features that are more important than other variables with attrition variable.



The above graph shows the different features with the bars showing the importance of feature.

The features once identified can be chosen and a model developed further but in our study for a more comprehensive study, we will consider all the features present with only dropping job role as we will be considering department in place of job role as both features are correlated.

Based on the information given in the tree of the random forest , we should choose the following features:

TotalWorkingYears (entropy = 0.474)

MonthlyIncomes (entropy = 0.918)

Jobinvolvement (entropy = 0.914)

YearsAtCompany (entropy = 0.274)

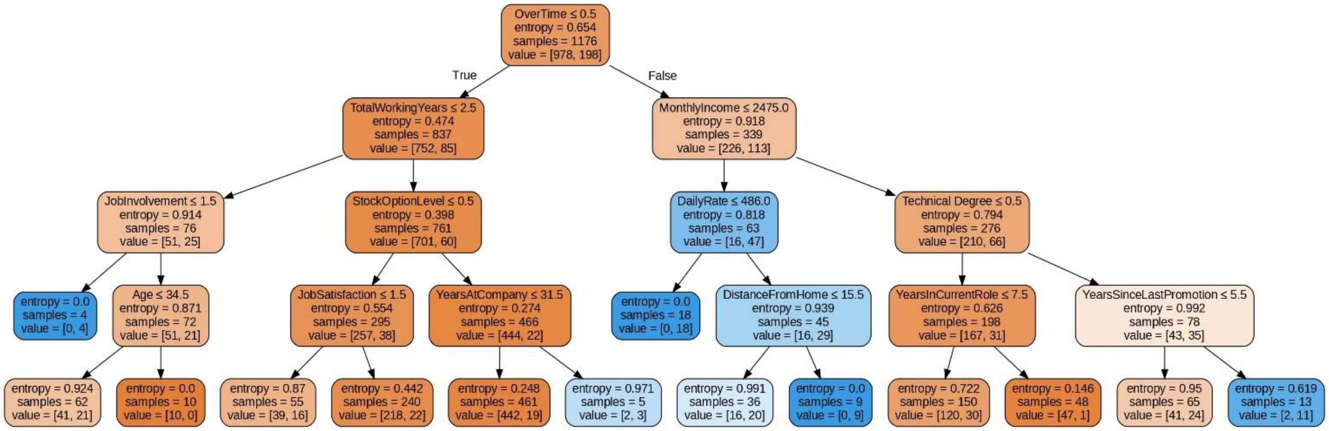
YeasinCurrentRole (entropy = 0.026)

I chose these features because they have the lowest entropy, which means that they are the most informative. Entropy is a measure of uncertainty, so a lower entropy means that there is less uncertainty about the values of the feature.

- TotalWorkingYears and Jobinvolvement are likely to be important predictors of employee performance and satisfaction. Employees with more experience and who are more involved in their jobs are more likely to be high performers and satisfied with their work.

- MonthlyIncomes is also likely to be an important predictor of employee performance and satisfaction. Employees with higher incomes are more likely to be able to afford to live comfortably and have a good quality of life, which can lead to improved job performance and satisfaction.
- YearsAtCompany is a measure of employee loyalty and commitment. Employees who have been with the company for a longer period of time are more likely to be loyal and committed to their jobs, which can lead to improved job performance and satisfaction.
- YearsinCurrentRole is a measure of employee growth and development. Employees who are constantly learning and growing in their careers are more likely to be motivated and engaged in their work, which can lead to improved job performance and satisfaction.

Overall, I believe that these five features are the most informative and predictive of employee performance and satisfaction.


OverTime ≤ 0.5
entropy = 0.654
samples = 1176
value = [978, 198]
True
False
TotalWorkingYears ≤ 2.5
entropy = 0.474
samples = 837
value = [752, 85]
MonthlyIncome ≤ 2475.0
entropy = 0.918
samples = 339
value = [226, 113]
JobInvolvement ≤ 1.5
entropy = 0.914
samples = 76
value = [51, 25]
StockOptionLevel ≤ 0.5
entropy = 0.398
samples = 761
value = [701, 60]
DailyRate ≤ 486.0
entropy = 0.818
samples = 63
value = [16, 47]
Technical Degree ≤ 0.5
entropy = 0.794
samples = 276
value = [210, 66]
entropy = 0.0
samples = 4
value = [0, 4]
Age ≤ 34.5
entropy = 0.871
samples = 72
value = [51, 21]
JobSatisfaction ≤ 1.5
entropy = 0.554
samples = 295
value = [257, 38]
YearsAtCompany ≤ 31.5
entropy = 0.274
samples = 466
value = [444, 22]
entropy = 0.0
samples = 18
value = [0, 18]
DistanceFromHome ≤ 15.5
entropy = 0.939
samples = 45
value = [16, 29]
YearsInCurrentRole ≤ 7.5
entropy = 0.626
samples = 198
value = [167, 31]
YearsSinceLastPromotion ≤ 5.5
entropy = 0.992
samples = 78
value = [43, 35]
entropy = 0.924
samples = 62
value = [41, 21]
entropy = 0.0
samples = 10
value = [10, 0]
entropy = 0.87
samples = 55
value = [39, 16]
entropy = 0.442
samples = 240
value = [218, 22]
entropy = 0.248
samples = 461
value = [442, 19]
entropy = 0.971
samples = 5
value = [2, 3]
entropy = 0.991
samples = 36
value = [16, 20]
entropy = 0.0
samples = 9
value = [0, 9]
entropy = 0.722
samples = 150
value = [120, 30]
entropy = 0.146
samples = 48
value = [47, 1]
entropy = 0.95
samples = 65
value = [41, 24]
entropy = 0.619
samples = 13
value = [2, 11]