Introduction

Human Resource analytics is the area in the analytics field that deals with people analysis and applying analytical process to the human capital within the organization to improve employee performance and improving employee retention, by gathering related data and then using this data to make informed decisions on how to improve these processes[1].

Employee turnover is one of the biggest challenges the companies face. Several factors lead to turnover such as average number of hours spend per month by the employees, salary, promotions, job rotation, number of projects.

I work as HR assistant and salary accountant, it was interesting to know the reasons of employee's resignation especially when it is a sudden and from an expected employees. Moreover, me as an employee, during my 10 working years I didn't stay in one work more than 2 years. After conducting this analysis project, I could find that I have common features of those who is leaving their works. In this project, I am going to take the role of HR analyst to have a better understanding about the factors that are responsible for the employees leaving and to seek answers for the following questions:

- Why employees are leaving the company?
- What features affect the turnover rate most?
- Are there any common features between the leaving employees?
- Why are the most experienced employees leaving?
- Why are the high salaried employees leaving?
- Why are the satisfied employees leaving?
- Which employee will leave next?
- What the company should do to keep their high performing employees?

In order to answer these questions, I am going to analyze the wide variety of variables used to describe the employees who left, and to find relationships between them, to cluster the employees according to some criteria. Finally, I will try to make some predictions of the variables using different classifiers and to end with the conclusion and some future recommendations.

[1] https://www.questionpro.com/blog/hr-analytics-and-trends/

Data Exploration:

“Why the employees leave “is the dataset where I am going to apply data exploration analysis and modeling. I found an interesting dataset in kaggle named “HR_comma_sep". This dataset is good enough for my analysis purpose; it contains a few number of features and a large number of individuals, so we can perform solid statistics. The table below shows the first rows of our dataset.

satisfaction_level	last_evaluation	number_project	average_montly_hrs	time_at_company	work_accident	turnover	promotion_last_5years	department	salary
0.38	0.53	2	157	3	0	1	0	sales	low
0.80	0.86	5	262	6	0	1	0	sales	medium
0.11	0.88	7	272	4	0	1	0	sales	medium
0.72	0.87	5	223	5	0	1	0	sales	low
0.37	0.52	2	159	3	0	1	0	sales	low

The dataset contains 15000 rows described by 10 features (columns)

The features are divided to:

- Five numerical variables: satisfaction_level, last_evaluation, number_project, average_monthly_hours, time_at_company.
- Three Boolean variables {Yes: 1, No: 0} variables: turnover, promotion_last_5years, work_accident.
- Two categorical variables: department is nominal, Salary {low, medium, high} is ordinal.

I will explain in more detail what these variables represent:

- **satisfaction_level:** it is a numerical feature. It measures the perceived level of pleasure derived from individual performance and it is a motivating force for occupational behavior. Its values in the current dataset vary between [0.09,1].
- **last_evaluation:** it is a numerical feature. It represents the last evaluation of the employees’ performance has done by the company. Its values in the current dataset vary between [0.36,1].
- **number_project:** it is a numerical variable. It represents the number of the projects has been done by each employee. Its values in the current dataset vary between [2,7].
- **average_monthly_hrs:** it is a numerical variable. It represents the average number of working hours for each employee per month. Its values in the current dataset vary between [96,310]
- **time_at_company:** it is a numerical variable. It represents the number of working years for each employee in the company. Its values in the current dataset vary between [2, 10]
- **turnover:** it is a Boolean variable. It will be my target variable, it encoded in {0,1}; 0 if the employees stayed and 1 if the employees left.
- **promotion_last_5years:** it is a Boolean variable. It encoded in {0, 1}; 1 if the company promoted this employee and 0 if not in the last 5 years.

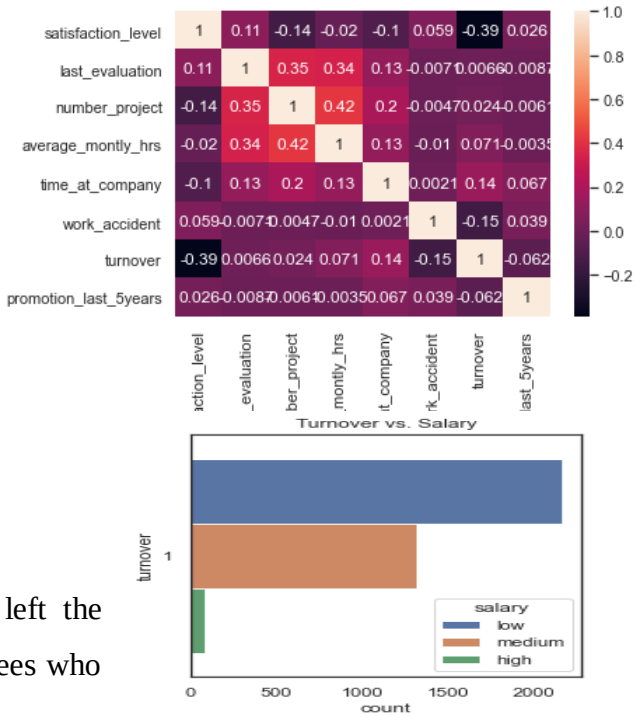
- **work_accident:** it is a Boolean variable. It encoded in {0, 1}; 1 if the employees had an accident in the company and 0 if employees didn't.
- **department**: it is a categorical variable. It represents the working department to which each employee belongs to ['sales', 'accounting', 'hr', 'technical', 'support', 'management', 'IT', 'product_mng', 'marketing', 'RandD']
- **salary:** it is a categorical ordinal variable. It represents the level of the salary for each employee in the company. It takes three ordered values {low, medium, high}.

After checking the dataset, I found no missing values. Our target variable is the "turnover" variable. After applying some statistics on the target variable, I found that there are **3571 employees left the work**. **The turnover rate =** number of the left employees /total number of employees = 3571/15000= **23.8%.** Let us have a fast look to the features of the employees who left the work

left	satisfaction_level	last_evaluation	number_project	average_montly_hrs	time_spend_company	work_accident	promotion_last_5years	salary
0	0.666810	0.715473	3.786664	199.060203	3.380032	0.175009	0.026251	1.650945
1	0.440098	0.718113	3.855503	207.419210	3.876505	0.047326	0.005321	1.414730

The table above shows that employees who left had 44% satisfaction level and they worked more than 207 hours per month, but they are not promoted in last 5 years.

Now I will go deeper in the dataset and explore the possible relations between the features and to check which of them affect our target variable. First, to check which of the numerical variables are linearly correlated. To this end, we computed the Pearson correlation coefficient between all these variables, which are graphically shown in



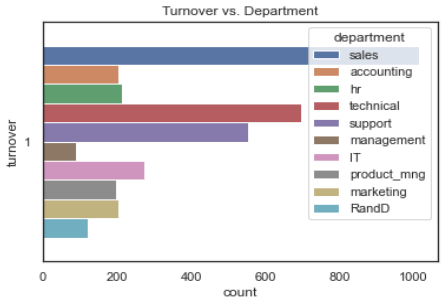
The heatmap, we can see that there is a positive correlation between the last_evaluation and the average_ monthly_ hrs **0.34** and the number of projects **0.35**; which could mean that the employees who spent more hours and did more projects were highly evaluated. In addition, there is positive correlation between the average_monthly_hrs and the number_project **0.42**; the employees that did more projects spent more working hours. As expected, there is a negative correlation **-0.39** between the turnover and the satisfaction level that could mean that less satisfied employees tend to leave their works. **Is the satisfaction level the only reason of leaving the company?**

After applying a further analysis, I got that the employees who left the company had low and medium salary unlike the high salary employees who

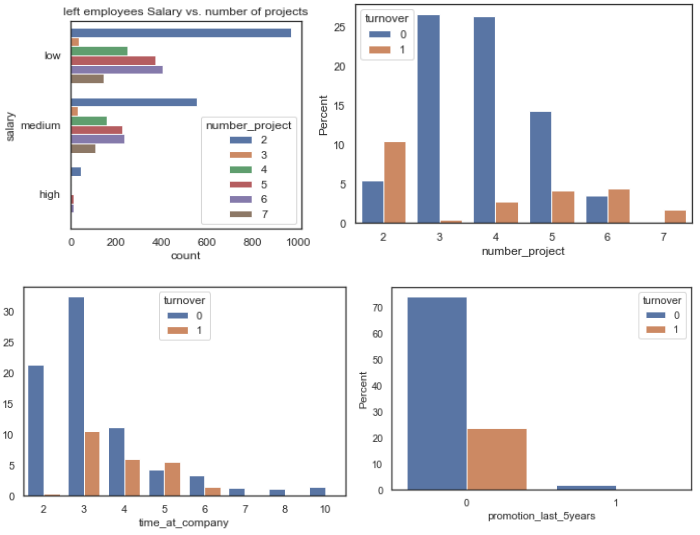
tend to stay. However, if the salary level is another reason to leave the company, **how can we explain why the high salary employees left?**

I found also that the sales, technical, and support departments had the highest turnover rate, while the management department had the lowest. I can conclude that the high salary employees who left the company belongs to the sales and technical departments where the employees there took the third and the fourth less satisfaction level departments in the company.

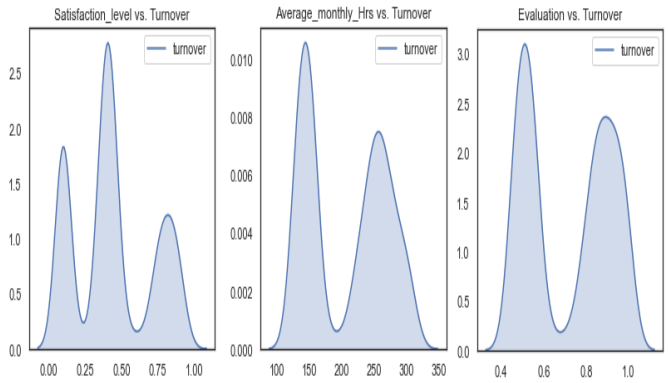
```
department
accounting     0.582151
hr             0.598809
technical      0.607897
sales          0.614447
IT             0.618142
support        0.618300
marketing      0.618601
product_mng    0.619634
RandD          0.619822
management     0.621349
Name: satisfaction_level,
```



In addition, I found that as the number of projects increases, the employee tends to leave. From the graph, we can see that all the employees who had 7 projects left. Most of who had 2 projects left the work (maybe they felt not involved in the company like my case!) .The majority of employees who stayed with 3, 4 projects. I found also that the high salaried employees who left had two projects (we can see that they are not overloaded, maybe they were waiting for a promotion or they had cultural problems). The employees also started to leave the company between the third and sixth year. I can add that the unprompted employees in the last 5 years tend to leave the company. There is no nothing to say about the work_accident feature, the employees did not leave the company because of the work accident and vice versa.



By plotting the distribution of the highly correlated features with our target variable, we can notice that there are bi/tri-mixed distributions in each graph, as if the employees are grouped in a way with these features. The graphs show us that the employees left the work in both of the extreme points: **overworked-underworked:** The employees who left either work lower than 150 hours per month, or they work more than 250 hours per month. **High performance-low performance:** the employees who left are evaluated either bad or quite good in their last performance while the employees still working in the company had an evaluation score range 0.6-0.8. **High satisfaction-low satisfaction:** Employees who left were in general less satisfied with their jobs. Alternatively, there are satisfied employees but they left the company; this means that there are other factors contributing to an employee's leaving their job other than being satisfied with their job or not.

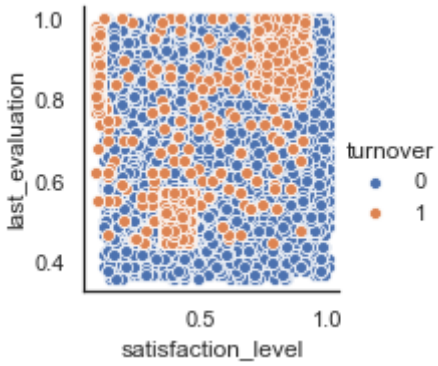


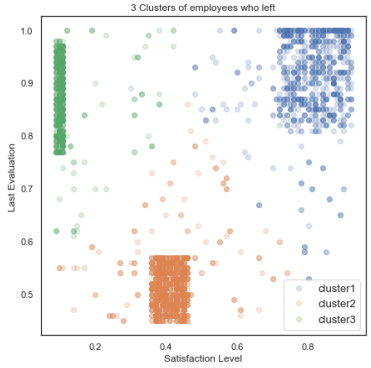
Clustering analysis:

Once we have gained an insight into the variables of the dataset and their relations, we will now try to cluster the employees. To this end, we will use all the numerical variables. We will cluster according Partitioning methods by using K-means clustering where we pick *K* points as cluster centers body, assign each data point to the closest cluster center, re-compute cluster center of each cluster until point assignment no longer change.

The satisfaction level vs. the last evaluation relationship shows an interesting grouping (clusters), which could give us another insight about the categories of the employees who left the company:

- **Cluster1:** Low evaluated -low satisfied level employees: they have low performance and felt bad in their work (maybe they are fired or they felt uninvolved), the total is 1650 employees.
- **Cluster2:** High evaluated-low satisfied level employees: they have high performance but not happy maybe they are overworked or overqualified for this work or they are not promoted. The total is 944 employees.
- **Cluster3:** High evaluated-high satisfied employees: they have high performance and felt happy but they left the company. Maybe they were waiting for a promotion but they did not get it so they found better opportunity in another workspace, the total is 977 employees.






```
clusters centers:
 [[0.41014545 0.51698182]
 [0.11115466 0.86930085]
 [0.80851586 0.91170931]]
```

Model Fitting and Prediction Techniques

- **Supervised Classification**

We will use the most popular techniques and models to make our predictions. The goal is to achieve the highest possible accuracy and test the predictive technique on the test data. If the technique is accurate then our prediction results on the test data will generate exactly the same values.

The models that will be considered are:

- Decision Tree Classifier
- Random Forest Classifier
- KNN Classifier
- Logistic Regression Classifier
- SVM

As our target variable is a discrete (binary) variable; therefore, I built a **Logistic Regression Classifier** to see the relationship between the dependent and the independent variables, then to check if this model describes the data well by calculating some classification evaluation metrics: precision, recall, F1-score, accuracy, R-squared, etc. After that to predict the test data to get the probability for the next employee to leave. As I mentioned before, the dataset has two categorical variables (salary, department). **First,** In order to build the model, I will transform these two variables to dummy variables. **Second,** standardizing the numerical variables (by applying a transformation which results are variables with mean 0 and standard deviation 1).**Third,** splitting the data to train, validation, test sets to apply fitting and predicting. **Fourth,** evaluating the classifier model, by checking the metrics in the classification report. **Fifth**, repeating the same process to build other types of classifiers: Decision Tree Classifier, Random Forest Classifier, KNN Classifier, and SVM. **Finally,** comparing them using the (yet unseen!) test set.

```
Training Set: 10499 rows
Validation Set: 2250 rows
Test Set: 2250 rows
```

The table below summarizes the models classification metrics and their running time in seconds while building and fitting the model. We can notice that when I got overfitting models I tried to fix it to get the best fit for each classifiers then comparing the statistics between testing on train and test sets.

Classifier	Classification results	Build-Fit runtime
Decision Tree	Testing on Training set: 1.0 Validation on validation set: 0.977 min_samples_leaf : 1 min_samples_split : 2 Tree Structure: depth: 21 nodes: 637 Decision Tree accuracy on testing set: 0.972	0.07497572898864746

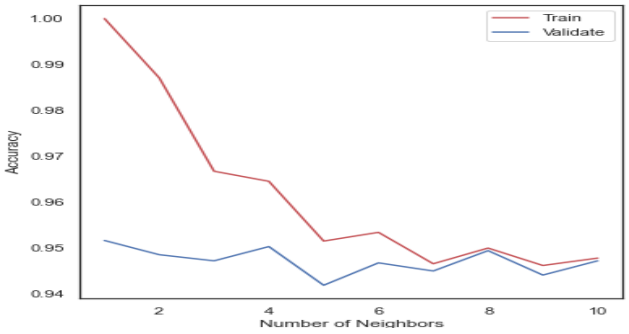
	Best fit: Max_depth:9 Testing on Training set: 0.985712925040 Validation on validation set: 0.982666666 Decision Tree accuracy on testing set: 0.980 Precision=0.98, recall=0.99,F1-score=0.99 **Classification Report:** On 1690 False Values (stayed employees) : Recall:0.99,precision:0.98,F1-score:0.99 0n 560 True Values (left employee):Recall:0.93,precision:0.98,F1-score:0.96	Best fit = 0.06947565078735352
Random Forest	Testing on Training set: 1.0 Validation on validation set 0.990666666 min_samples_leaf : 1 min_samples_split : 2 Random Forest accuracy on testing set: 0.988	1.393690586090088
	Best fit: n_estimators:1000,n_jobs:-1,max_depth=**10**,min_samples_leaf=5 Testing on TRAINING set: 0.9768549385655777 Validation on VALIDATION set: 0.9782222222222222 Random Forest accuracy on testing set: 0.972 **Classification Report:** On 1690 False Values (stayed employees) : Recall:1,precision:0.97,F1-score:0.98 0n 560 True Values (left employee):Recall:0.89,precision:0.99,F1-score:0.94	5.71064567565918
KNN Classifier	Testing on TRAINING set: 0.9514239451 Validation on validation set: 0.941777777 n_neighbors : 5 p : 2 leaf_size : 30 K-Nearest Neighbors accuracy on testing set : 0.935	0.17289996147155762
	Best fit: n_neighbors : **10** n_jobs : -1 Testing on Training set: 0.9477093056481569 Validation on validation set: 0.9471111111111111 K-Nearest Neighbors accuracy on testing set : 0.936 **Classification Report:** On 1690 False Values (stayed employees) : Recall:0.94,precision:0.97,F1-score:0.96 0n 560 True Values (left employee):Recall:0.91,precision:0.84,F1-score:0.88	0.1389176845550537
Logistic Regression	**Best fit:** Testing on Training set: 0.7928374130869607 Validation on validation: 0.7946666666666666 Logistic regression accuracy in testing set: 0.778 **Classification Report:** On 1690 False Values (stayed employees) : Recall:0.94,precision:0.97,F1-score:0.96 0n 560 True Values (left employee):Recall:0.91,precision:0.84,F1-score:0.88	0.36179089546203613
SVM	Testing on Training set: 0.7631202971711591 Validation on validation: 0.7551111111111111 SVM accuracy on testing set: 0.763	1.3496320247650146
	Best Fit: C=1.0, kernel='rbf', degree=3,coef 0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, , max_iter= -1 Testing on Training set: 0.7758834174683303 Validation on validation: 0.7706666666666667 SVM accuracy on testing set: 0.765 **Classification Report:**	6.7048020362854

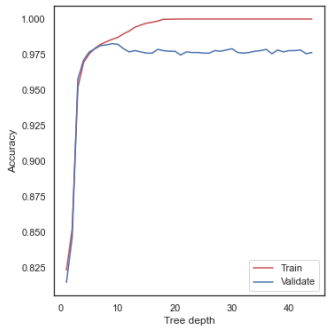
	On 1690 False Values (stayed employees) : Recall:1,precision:0.76,F1-score:0.86 0n 560 True Values (left employee):Recall:0.06,precision:1,F1-score:0.10	
Stacking: meta-model of all classifiers	**Using all the best fit models:** Accuracy on the test set: 0.9808888888888889 Testing on Training set: 0.9853319363748928 Validation on validation: 0.9853333333333333 **Classification Report:** On 1690 False Values (stayed employees) : Recall:1,precision:0.98,F1-score:0.99 0n 560 True Values (left employee):Recall:0.93,precision:0.98,F1-score:0.96	42.94318175315857

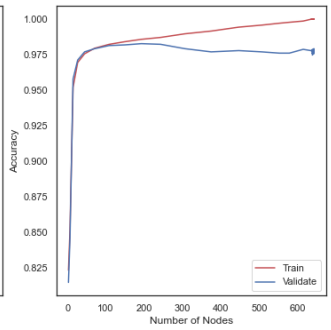
We can conclude that the Decision tree, random forest and KNN classifiers gave the highest evaluation metrics. SVM and logistic regression classifiers gave the lowest.

Overfitting and optimizing Models:

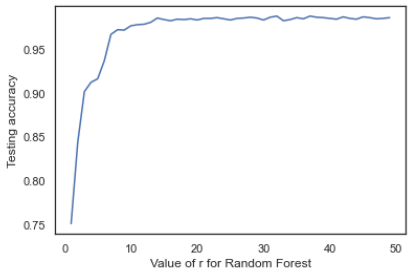
The graphs below show the overfitting between the train and validation sets in the Decision tree and KNN.
We can optimize our model and get the best fit by k (n_neighbors) =**10** in KNN models. Max depth= **9** in decision tree model.







Random forests combine many decision trees and in this way reduce the risk of overfitting by adding an additional layer of randomness to bagging. For constructing trees then classify and split each node by using the best among a subset of predictors randomly chosen at that node. The graph shows how to optimize the model by choosing a max depth "r" of around **10**, which seems to give good results to get the best fit.



Random forest classifier gave us the features by their importance:

1. satisfaction_level (0.33)
2. number_project (0.20)
3. time_at_company (0.18)
4. average_montly_hrs (0.14)
5. last_evaluation (0.12)
6. work_accident (0.01)
7. salary_low (0.01)
8. salary_high (0.01)

Conclusion:

Employee satisfaction is the highest indicator for employee turnover. However, should be taken in account that the employees with low and medium salaries constitute the highest proportion of the turnover rate. Both overworked > 250 hrs/month who had 6, 7 projects and underworked employees < 150 hrs/month who had 2 projects left the work. Moreover, both the high and low evaluated employees count as a risk of leaving the company.

The employees are grouped in 3 clusters where they have common features. The high performing employees tend to leave the work because they are overloaded; they had 6 or 7 projects which required them to work >250 hrs monthly. Due to this situation, the employees were dissatisfied, their satisfaction level rate < 0.1 .Some of them were prompted in the last 5 years, so probably because of that they kept working. While the high performing and satisfied employees (their satisfaction level >0.6) they had to work >230 hrs monthly but they were still ambitious to get a promotion. When they did not get it, they took the decision to leave their works looking for a better job chances.

Support, sales and Technical departments have the highest number of leaving employees. These departments also run a high number of projects and the employees work more than 210 hrs/month. However, the satisfaction level of these employees is above the average, people are probably leaving due to a high workload and or cultural problems within these departments.

The high salaried employees tend to stay but they show a different pattern for leaving the company. The employees who left had a satisfaction level > 0.5.They had to work >300 hrs/month. They left the company after spending 4 years in the company; they belong to the sales and technical departments. Since satisfaction had the most effect in determining employee turnover, the high salaried employee's problem can be generalized down to a personal level. Alternatively, the problem is not with the employees, but persist in a deeper level of the company.

By using decision tree or random forest classifiers, we could predict the next employee will leave. These two classifiers show very high accuracy of predicting the unseen testing set.

Analyzing and modeling the data do not change the decision of leaving the work, as it is a personal decision. Therefore, we cannot use the predictive metrics as a solution. However, we can use these tools to provide the employees with better relevant information for better decision making. In addition, to save our experienced and high performing employees in the company by understanding their conditions and by providing an incentive programs to make them stable and more satisfied. For example; by using the analysis and modeling tools, we can predict the probability of the next leaving employees, to rank them according to the highest expected loss for the company, and then allocate a limited incentive budget or promotion programs to the highest probability instances.

The dataset gave quite good insights why the employees leave the company but for deeper understanding in order to get more accurate finding, we should conduct more experiments or collect more variables about the employees that could have more impact on determining employee turnover and satisfaction such as their distance from home, gender, age, etc.