Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

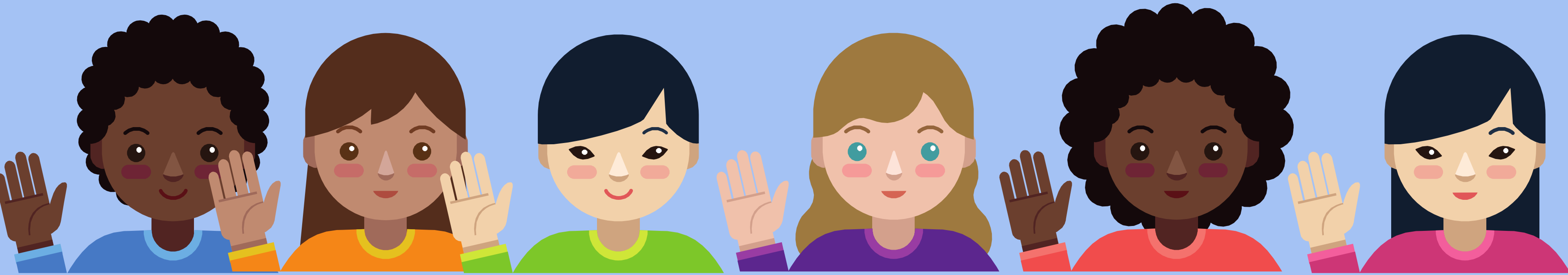
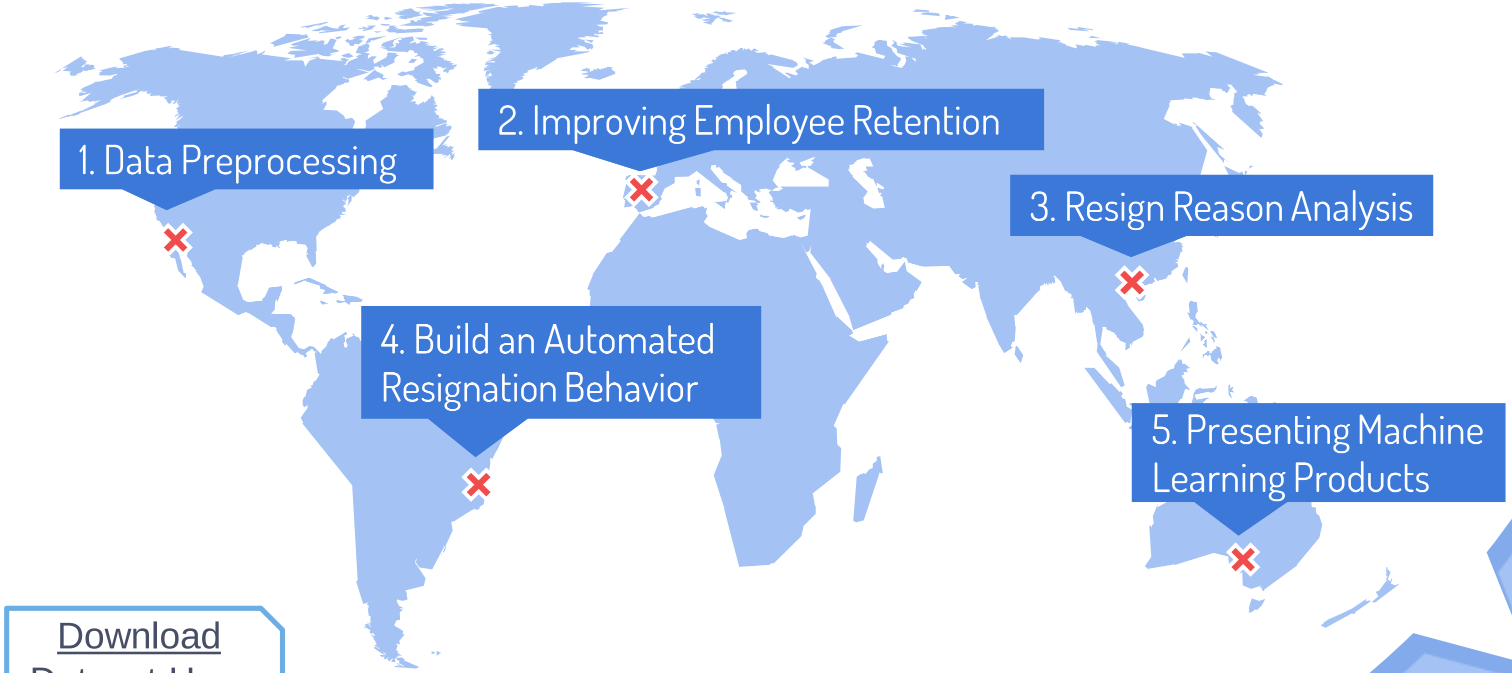


Table of Contents



1. Data Preprocessing

2. Improving Employee Retention

3. Resign Reason Analysis

4. Build an Automated Resignation Behavior

5. Presenting Machine Learning Products

Download Dataset Here

Data Preprocessing

Overcoming missing values

Use the fillna and drop functions for the following fields :

- Using the average value for the Employee satisfaction score, so that the average value obtained from the column remains
- Uses a value of 0 for Total Project Participation, Total Last Month Late, and Total Absence because a value that does not exist is considered not working so it is worth 0
- Drop the Join Program LOP column because its value has too many missing values and the Ever Worked column because the values of the columns are all 1 and yes
- Fill in the empty Reason Resign value with still_working because there is no resign date value
- Change – with 0- to make it easier to extract the year for the next process

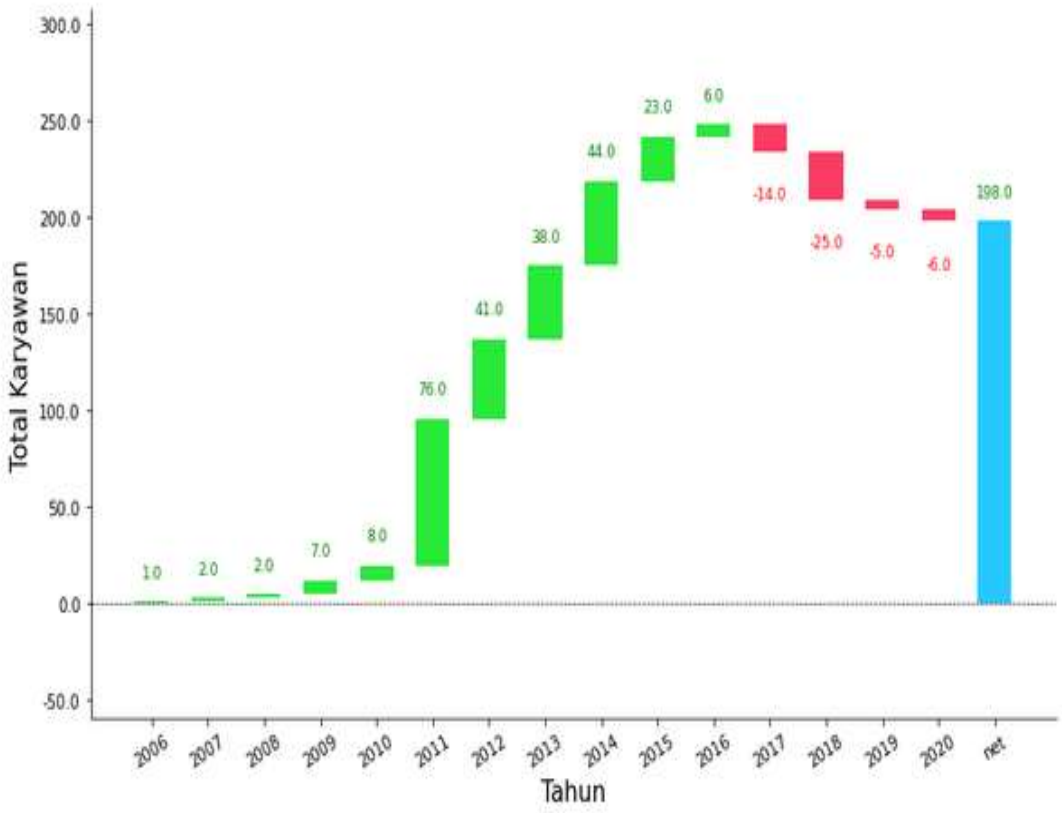
Extract year

Extract the year using the partion function - for the date column that will be used, the format has been adjusted and used to find new columns, namely age and length of work

For details, see jupyter notebook here

Annual Report on Employee Number Changes

Dinamika Total karyawan pertahun (2006-2020)

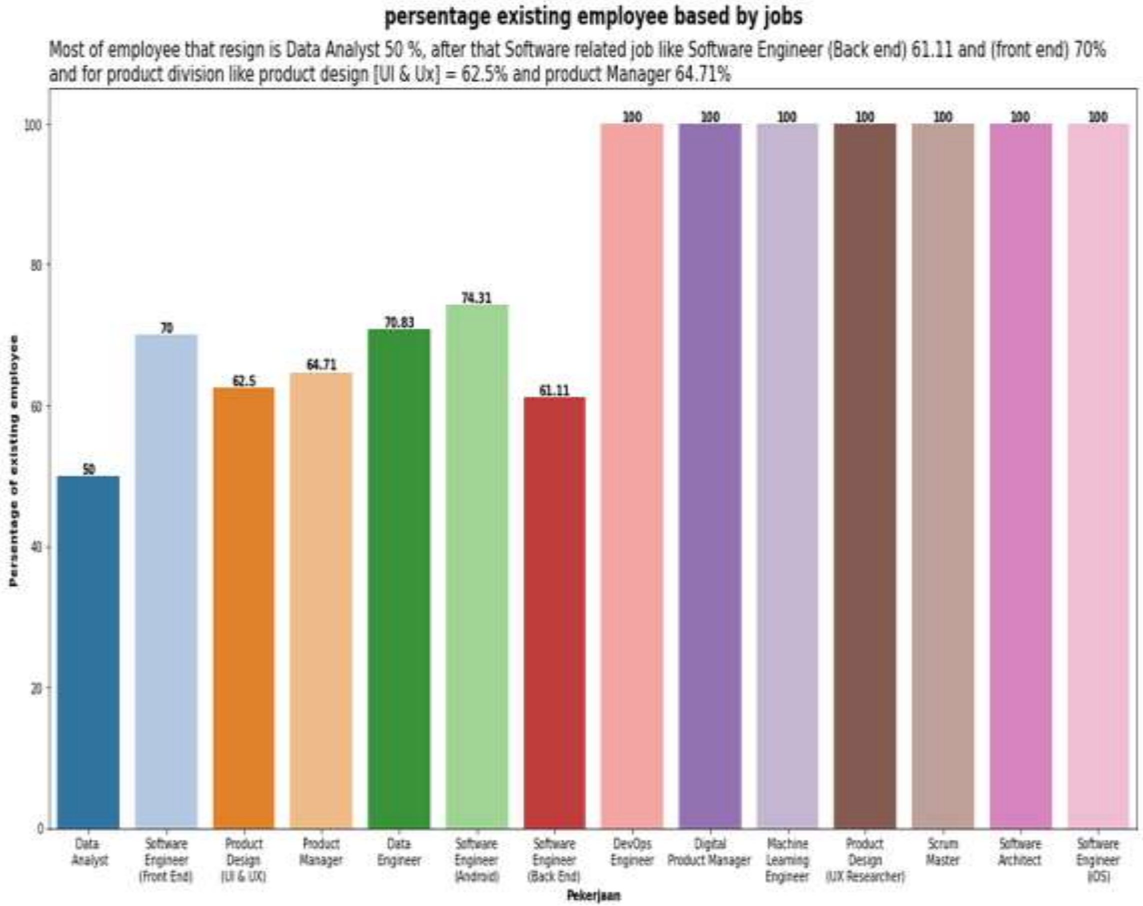


From the visualization it can be seen that :

- Initially there were many decreases compared to the increase in employees starting from 2017 to 2020
- From the existing data, the causes are career problems and the comfort of the work environment
- The total of all employees who are still working at the company is 198
- Insight: training can be given to understand the future career of the job being undertaken, and increasing the comfort of the work environment according to the reasons for resigning

For details, see jupyter notebook here

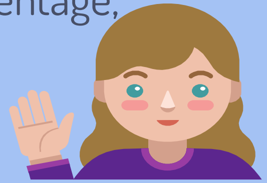
Resign Reason Analysis for Employee Attrition Management Strategy

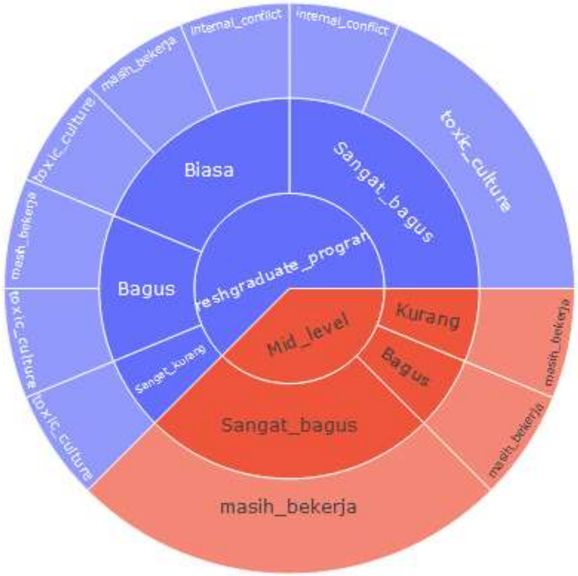


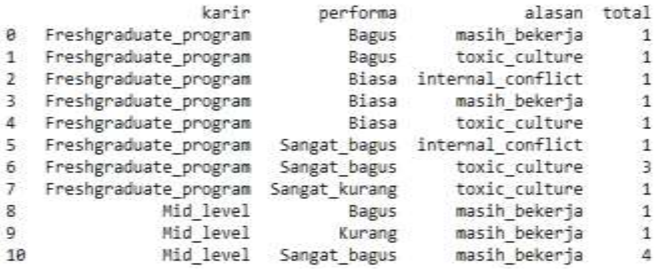
From the visualization it can be seen that :

- Most of the percentage of employees who resign are in Data Analyst jobs
- Followed by work in the Software division, where the percentage of permanent employees is Software Engineer (Back end) 61.11% and (front end) 70%.
- For the remainder, apart from data and software divisions, there are product divisions such as product design [UI & Ux] 62.5% and product manager 64.71%
- For deeper insights, sunburn visualization is carried out for jobs with the lowest percentage, namely Data Analyst

For details, see jupyter notebook here





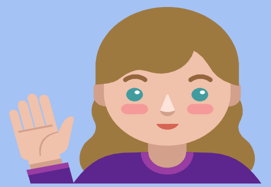


```
                   karir      performa              alasan  total
0   Freshgraduate_program         Bagus       masih_bekerja      1
1   Freshgraduate_program         Bagus       toxic_culture      1
2   Freshgraduate_program         Biasa   internal_conflict      1
3   Freshgraduate_program         Biasa       masih_bekerja      1
4   Freshgraduate_program         Biasa       toxic_culture      1
5   Freshgraduate_program  Sangat_bagus   internal_conflict      1
6   Freshgraduate_program  Sangat_bagus       toxic_culture      3
7   Freshgraduate_program Sangat_kurang       toxic_culture      1
8               Mid_level         Bagus       masih_bekerja      1
9               Mid_level        Kurang       masih_bekerja      1
10              Mid_level  Sangat_bagus       masih_bekerja      4
```

For details, see jupyter notebook here

From this visualization for Data Analyst workers it can be seen that :

- There are 2 types of careers namely Fresh graduate and Mid_level
- For Mid_level it can be seen that most of the performance given is at a very good value with 1 good and 1 less and the overall value is still working which can be said to be a Data Analyst who enters Mid_level is still comfortable being an employee
- For Fresh Graduates, employees who are still working have good and ordinary performance, while fresh graduates with Very_good performance resign due to toxic_culture and internal_conflict, fresh graduates with good and Ordinary performance each only have 1 left who is still working, for the rest resign with reasons the same thing, namely toxic_culture and internal conflict
- Insight: from these data we can say that for Fresh graduate employees who have just entered the world of work it is still difficult to adapt to a given work environment, so adjustments to the work environment are needed for fresh graduates who work as Data Analysts, who are likely to be able to enter Mid-level employees adapt

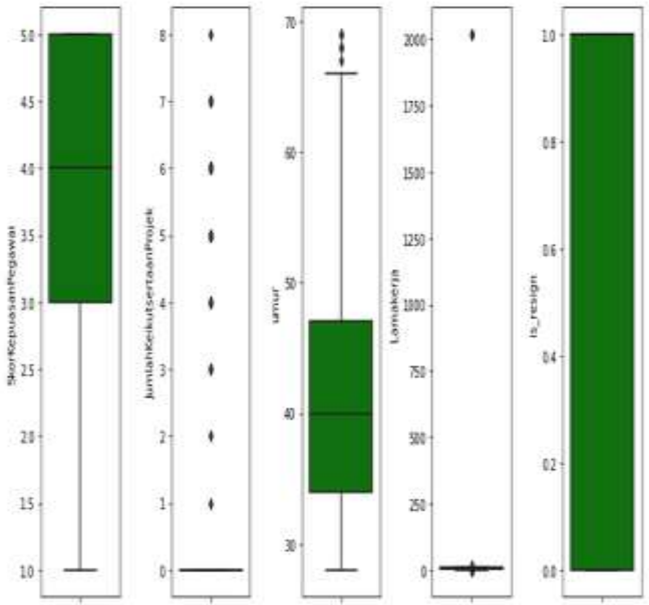


Build an Automated Resignation Behavior Prediction using Machine Learning

Data Preprocessing for Machine Learning

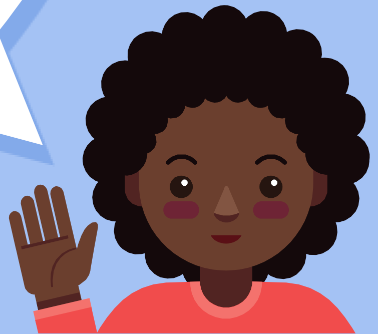
Handling missing data and duplication

Missing data in this data has been resolved in the previous preprocessing process so there is no missing data for duplicate data checked whether there is duplicate data and the result is that there is no duplicate data



Outliers Data

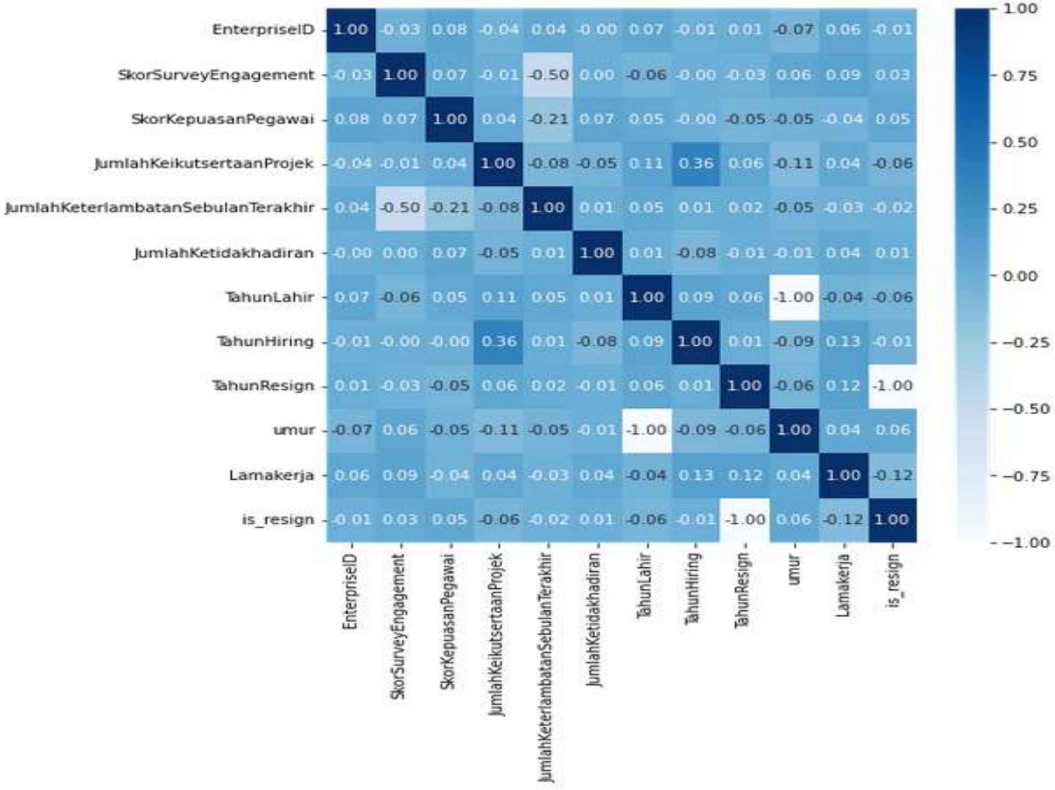
In the boxplot, it is found that there are several features with outlier data which after analysis show that the Number of Project Participation features and the age of the outliers are considered to still make logical sense so they do not need to be changed, while the length of work of outliers needs to be dropped/deleted due to inappropriate data, i.e. the year of resignation < from the year of hiring



For details, see jupyter notebook here

Data Preprocessing Feature

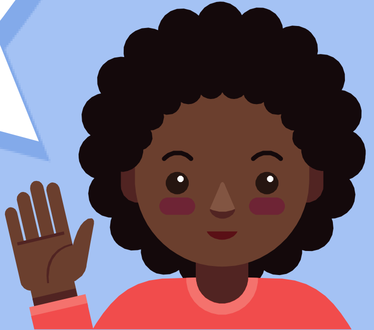
Feature selection



Determine what features will be used for ml, namely to predict the value of the Engagement survey using Classification with the target is resign which shows the employee resigned, based on the correlation value ((corr has a value of +/-) >= 0.05) and so that the results are not discriminatory, the feature that will be used are :

- Employment status
- Work
- Career path
- Performance Employee
- Hiring Platform
- Employee Satisfaction Score
- Total Project Participation
- Level of education
- Reason Resign
- age
- Length of working is_
- resign

For details, see jupyter notebook here



Data Preprocessing Feature

Feature Engineering

This process is carried out on the Job, HiringPlatform, and ReasonResign features, where each feature is changed into 3 groups to reduce the number of features to be processed in machine learning :

- Job -> Work Division: Software division, Data Division, and Product Division
- HiringPlatform -> GroupPlatform : Indeed, LinkedIn, and Others (Indeed and Linkin are the most used platforms)
- ReasonResign -> GroupReason : still_working, problem_career, problem_convenience

Feature Encoding

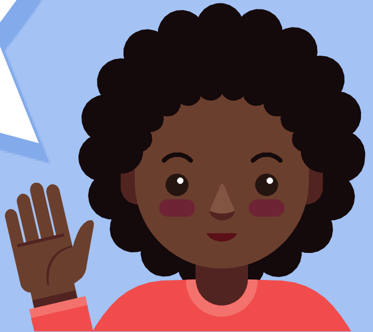
Label Encoder

This process is carried out on the Employment Status, Career Path, Employee Performance, and Education Level features in which these features are sorted in descending order from the smallest level to the largest level starting with the label 0 -> 4 (max) according to the existing level

Onehot Encoder

This process is carried out on features that have carried out the feature engineer process, namely on the Job, HiringPlatform, and ReasonResign features where the existing group value will change to a value of 1 or 0 according to the value it has

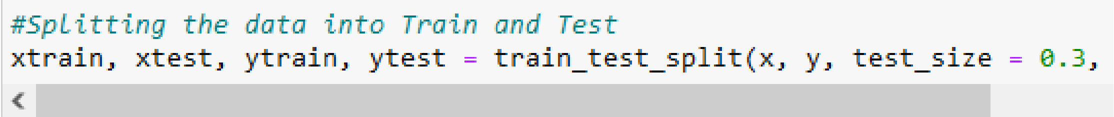
For details, see jupyter notebook here



Modelling Machine Learning

Split data train dan testing

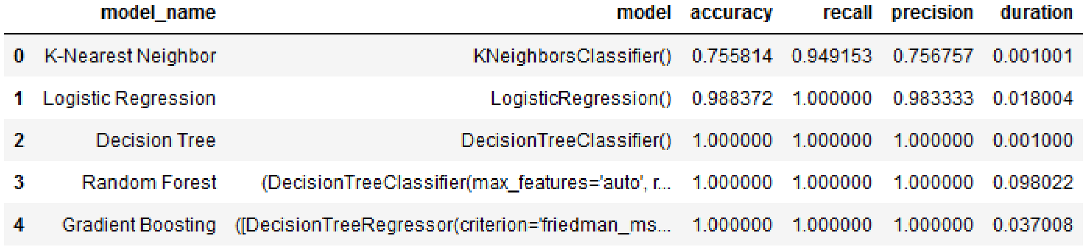
The modeling that will be carried out in this project is classification modeling with a split data test, namely



```
#Splitting the data into Train and Test
xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size = 0.3,
```

Modelling

Classification modeling carried out in this project uses the K-Nearest Neighbor algorithm, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting, the results of the best Accuracy and Precision are using the Logistic Regression algorithm, namely Acc = 98.84 % and Prec = 98 ,33 %

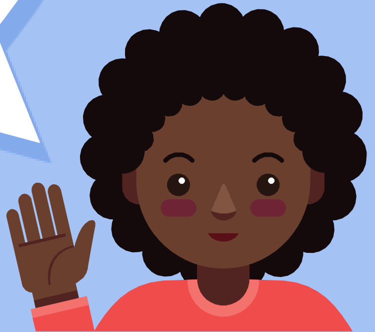


	model_name	model	accuracy	recall	precision	duration
0	K-Nearest Neighbor	KNeighborsClassifier()	0.755814	0.949153	0.756757	0.001001
1	Logistic Regression	LogisticRegression()	0.988372	1.000000	0.983333	0.018004
2	Decision Tree	DecisionTreeClassifier()	1.000000	1.000000	1.000000	0.001000
3	Random Forest	(DecisionTreeClassifier(max_features='auto', r...	1.000000	1.000000	1.000000	0.098022
4	Gradient Boosting	([DecisionTreeRegressor(criterion='friedman_ms...	1.000000	1.000000	1.000000	0.037008

Hyperparameter tuning

Tried hyperparameter tuning on the best algorithm, namely Logistic regression, but the results went down to Acc = 71% and Prec = 71%, which means the best way to model is without using Hyperparameters

For details, see jupyter notebook here



Evaluation Model

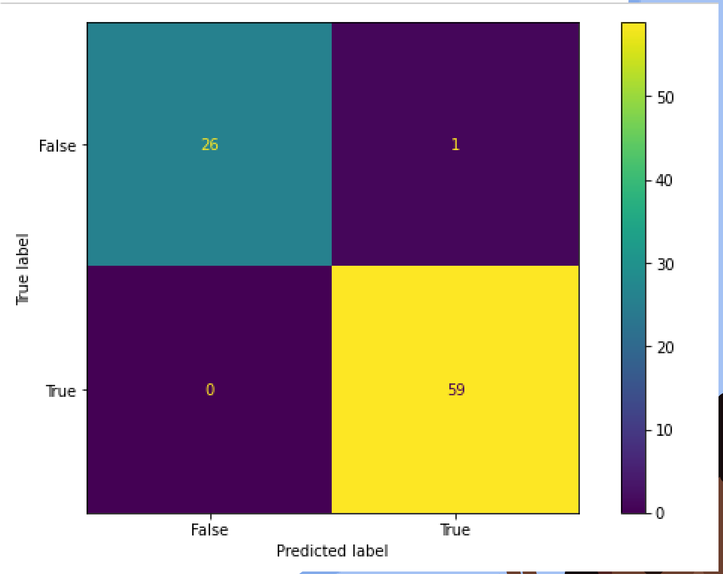
Confusion Matrix

It can be seen in the results of the is_resign Confusion Matrix, namely:(Predicted labels, True Labels)

- True True (TT) : is the correct prediction of the model that Employee Resigns (59)
- False False (FF) : is a true prediction of the model that employees do not resign (26)
- False True (FT): is a wrong prediction model that the employee does not resign but actually does (0)
- True False (TF): is a wrong prediction of the model that employees are resigning but actually are not (1)

So the results of the accuracy of the model using the Logistic Regression algorithm are:Accuracy = TT + FF / n_data = 59 + 26 /86 = 98.84%

from these results it can be concluded that the model can support compar classify employees who are likely to resign



For details, see jupyter notebook here

Evaluation Model

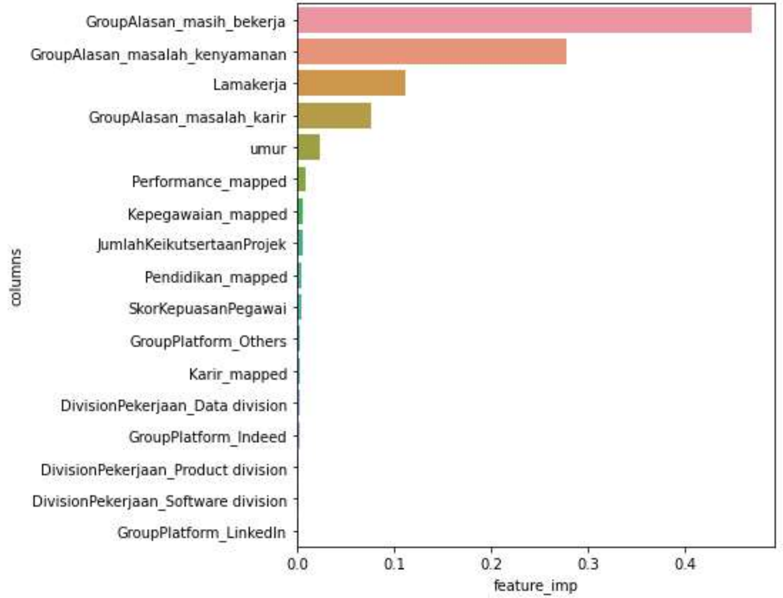
Feature Importance

It can be seen in the feature Importance graph that the features that have the most influence on the 7 biggest models are:

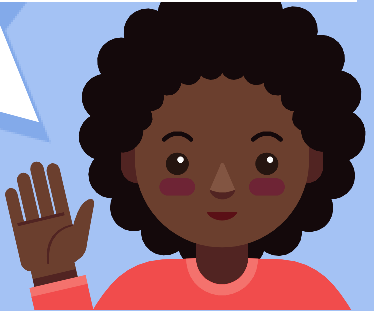
- GroupAlasan_still_working
- GroupReason_problem_convenience
- Length of work
- GroupReason_problems_career
- Age
- Performance_mapped
- Personnel_mapped

From these results it can be seen that the most important feature, namely GroupAlasan_still_working, is a feature that has the opposite effect on the target, namely is_resign because if the value is 1 it means that the value of is_resign is definitely 0

The next feature is Group Reasons_problems_convenience,which means that the cause of many employees resigning is a matter of convenience from the work environment and number 4 is a career problem



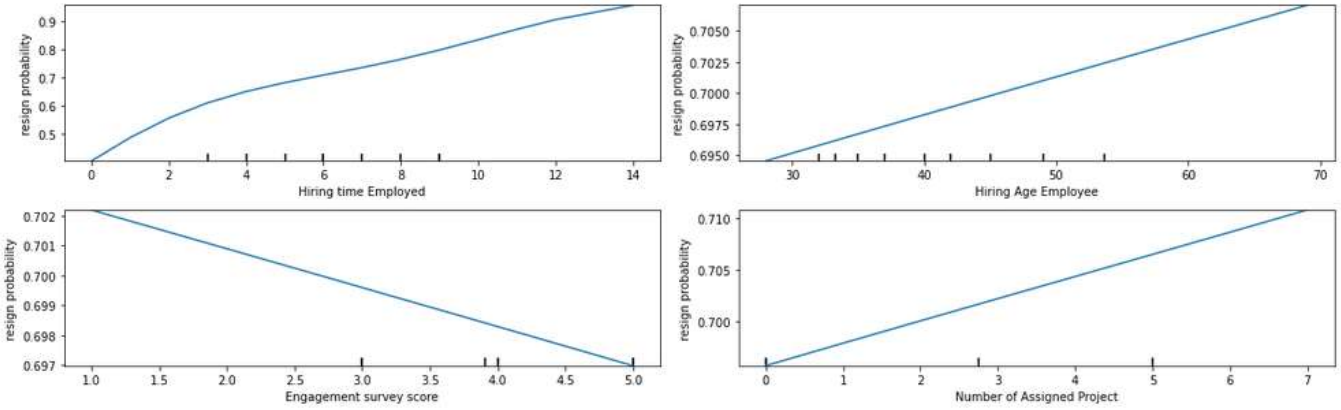
For the remaining features for the 7th largest no. 6 and no. 7, namely Performance and staffing related to the performance and level of employees in job positions



For details, see jupyter notebook here

Evaluation Model

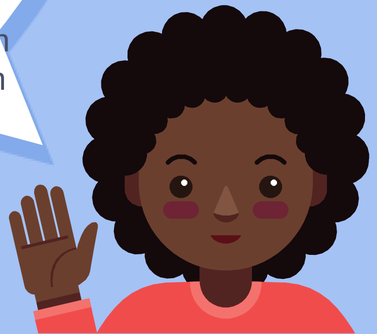
Partial Depedence Score



The partial Depedance score plot shows the relationship between the probability of an employee resigning and the following features:

- Length of Work: The longer an employee has worked at a company, the more likely he is to resign
- Age: The older the employee, the more likely it is to resign
- Employee Satisfaction Score: The greater the value of employee satisfaction, the less likely it is to resign
- Number of Project Participations: The more projects an employee works on, the more likely it is to resign

For details, see jupyter notebook here



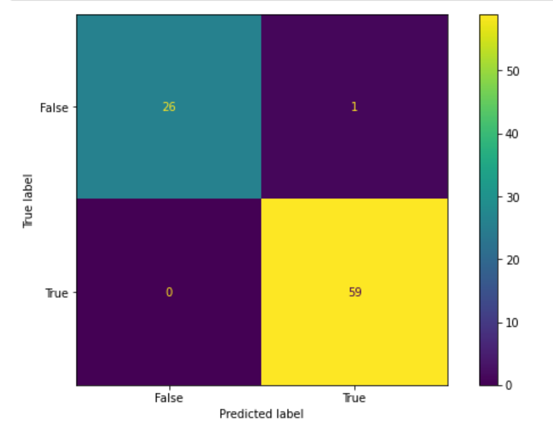
Presenting Machine Learning Products to the Business Users

Story Telling

In a company, it is very necessary to have quality Human Resources (HR) that can be managed properly so that they remain loyal and carry out their duties properly. By increasing employees or good human resources in the company, it will greatly affect the performance provided to build a company. Because of that, a machine learning was made to find out whether an employee will remain in a company or not to advance the company and reduce costs that shouldn't have to be spent to retain an employee.

The machine learning that has been created is machine learning with a Classification model to find out whether an employee will resign or be loyal to the company, this model uses a Linear Regression algorithm with the following performance:

The accuracy of the model using the Logistic Regression algorithm is:Accuracy = TT + FF / n_data = 59 + 26 /86 = 98.84%Which shows the amount of accuracy in predicting whether employees will resign or remain loyal to the company, using this model the company can determine which employees are likely to be loyal to the company

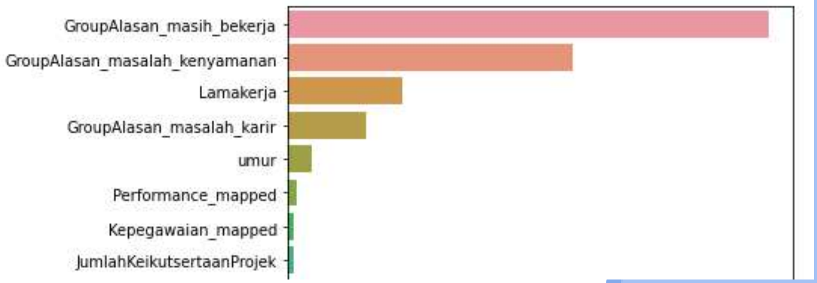


For details, see jupyter notebook here

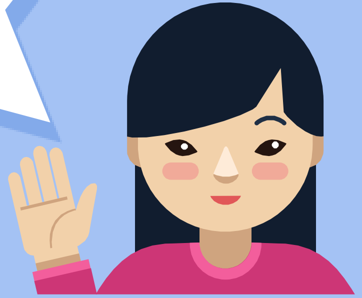
With the machine learning Classification model, employees will resign, you can also analyze what factors cause an employee to look at the plot as follows:

The plot shows the factors that influence whether an employee will resign or not :

- Still_working: shows employees who are loyal and still working for the company, which means that if the employee is still working, the employee will not resign
- Reason_problem resign : for convenience problem and career problem, which shows that convenience from work environment and clear career will decrease employee resignation rate
- Old_work and Age: shows the length of time the employee has worked which makes him older can affect the rate of employee resignation
- Performance & Number of project participation: shows the effect of employee performance and the number of projects undertaken at the level of employee resignation



From the factors using the importance feature, improvements can be made according to the factors that arise, by improving these factors, the employee resignation rate will definitely decrease and the company will continue to grow



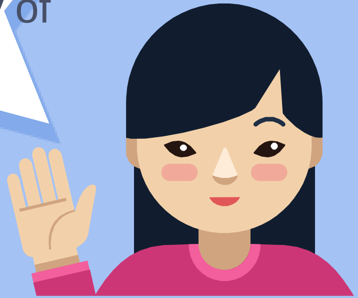
For details, see jupyter notebook here

Recomendation

Improvements can be made to the factors that affect the rate of employee resignation such as:

- Reasons for resigning: The work environment can improve the quality of the work environment to make it more comfortable for employees, such as being able to work remotely, there are psychologists to resolve internal conflicts, more suitable working hours and more, for career problems an appropriate training or workshop can be provided with the division of work in order to understand the future career
- Length of work and age: Additional salary can be given according to the length of work according to the age of the employee
- Performance and number of project participation: with the performance and results of the projects participated in, bonuses or position promotions can be given according to the performance given

The factors above are recommendations that can be given to improve the quality of Employees/SDA and reduce the probability of Employees Resigning



For details, see jupyter notebook here

Thanks!

Let's Connect

Mustiadi Zakki

Mustiadi Zakki