Start coding or generate with AI.

Connect to Big Query

```
# Libraries
from google.cloud import bigquery
from google.colab import auth

# authenticate
auth.authenticate_user()

# initialize the client for BigQuery
project_id = 'my-project-hr-attrition'
client = bigquery.Client(project=project_id, location='europe-west2')


#get the dataset and table
dataset_ref = client.dataset('employeedata', project=project_id)
dataset = client.get_dataset(dataset_ref)
table_ref = dataset.table('tbl_hr_data')
table = client.get_table(table_ref)
table.schema
```

```
[SchemaField('satisfaction_level', 'FLOAT', 'NULLABLE', None, None, (), None),
 SchemaField('last_evaluation', 'FLOAT', 'NULLABLE', None, None, (), None),
 SchemaField('number_project', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('average_montly_hours', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('time_spend_company', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('Work_accident', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('Quit_the_Company', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('promotion_last_5years', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('Departments', 'STRING', 'NULLABLE', None, None, (), None),
 SchemaField('salary', 'STRING', 'NULLABLE', None, None, (), None),
 SchemaField('employee_id', 'STRING', 'NULLABLE', None, None, (), None)]
```

```
new_table_ref = dataset.table('tbl_new_employees')
new_table = client.get_table(new_table_ref)
new_table.schema
```

```
[SchemaField('satisfaction_level', 'FLOAT', 'NULLABLE', None, None, (), None),
 SchemaField('last_evaluation', 'FLOAT', 'NULLABLE', None, None, (), None),
 SchemaField('number_project', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('average_montly_hours', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('time_spend_company', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('Work_accident', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('Quit_the_Company', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('promotion_last_5years', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('Departments', 'STRING', 'NULLABLE', None, None, (), None),
 SchemaField('salary', 'STRING', 'NULLABLE', None, None, (), None),
 SchemaField('employee_id', 'STRING', 'NULLABLE', None, None, (), None)]
```

```
# convert to dataframe
df = client.list_rows(table=table).to_dataframe()
df.head()
```

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	Quit_the_Company	p
0	0.36	0.56	2	132	3	0	1	
1	0.74	0.99	2	277	3	0	1	
2	0.45	0.53	2	155	3	0	1	
3	0.40	0.53	2	151	3	0	1	
4	0.36	0.51	2	155	3	0	1	

```
# convert to dataframe
df2 = client.list_rows(table=new_table).to_dataframe()
df2.head()
```

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	Quit_the_Company	p
0	0.331690	0.847953	6	151	4	0	0	
1	0.468434	0.169659	5	303	4	0	0	
2	0.858448	0.918311	4	162	6	1	0	
3	0.056211	0.322600	2	229	5	1	0	
4	0.382648	0.434348	4	260	3	1	0	

Build Model

Install pycaret

```
## Install pycaret
!pip install pycaret
```

```
Requirement already satisfied: pycaret in /usr/local/lib/python3.10/dist-packages (3.3.2)
Requirement already satisfied: ipython>=5.5.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (7.34.0)
Requirement already satisfied: ipywidgets>=7.6.5 in /usr/local/lib/python3.10/dist-packages (from pycaret) (7.7.1)
Requirement already satisfied: tqdm>=4.62.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (4.67.1)
Requirement already satisfied: numpy<1.27,>=1.21 in /usr/local/lib/python3.10/dist-packages (from pycaret) (1.26.4)
Requirement already satisfied: pandas<2.2.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (2.1.4)
Requirement already satisfied: jinja2>=3 in /usr/local/lib/python3.10/dist-packages (from pycaret) (3.1.4)
Requirement already satisfied: scipy<=1.11.4,>=1.6.1 in /usr/local/lib/python3.10/dist-packages (from pycaret) (1.11.4)
Requirement already satisfied: joblib<1.4,>=1.2.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (1.3.2)
Requirement already satisfied: scikit-learn>1.4.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (1.4.2)
Requirement already satisfied: pyod>=1.1.3 in /usr/local/lib/python3.10/dist-packages (from pycaret) (2.0.3)
Requirement already satisfied: imbalanced-learn>=0.12.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (0.12.4)
Requirement already satisfied: category-encoders>=2.4.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (2.6.4)
Requirement already satisfied: lightgbm>=3.0.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (4.5.0)
Requirement already satisfied: numba>=0.55.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (0.60.0)
Requirement already satisfied: requests>=2.27.1 in /usr/local/lib/python3.10/dist-packages (from pycaret) (2.32.3)
Requirement already satisfied: psutil>=5.9.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (5.9.5)
Requirement already satisfied: markupsafe>=2.0.1 in /usr/local/lib/python3.10/dist-packages (from pycaret) (3.0.2)
Requirement already satisfied: importlib-metadata>=4.12.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (8.5.0)
Requirement already satisfied: nbformat>=4.2.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (5.10.4)
Requirement already satisfied: cloudpickle in /usr/local/lib/python3.10/dist-packages (from pycaret) (3.1.0)
Requirement already satisfied: deprecation>=2.1.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (2.1.0)
Requirement already satisfied: xxhash in /usr/local/lib/python3.10/dist-packages (from pycaret) (3.5.0)
Requirement already satisfied: matplotlib<3.8.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (3.7.5)
Requirement already satisfied: scikit-plot>=0.3.7 in /usr/local/lib/python3.10/dist-packages (from pycaret) (0.3.7)
Requirement already satisfied: yellowbrick>=1.4 in /usr/local/lib/python3.10/dist-packages (from pycaret) (1.5)
Requirement already satisfied: plotly>=5.14.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (5.24.1)
Requirement already satisfied: kaleido>=0.2.1 in /usr/local/lib/python3.10/dist-packages (from pycaret) (0.2.1)
Requirement already satisfied: schemdraw==0.15 in /usr/local/lib/python3.10/dist-packages (from pycaret) (0.15)
Requirement already satisfied: plotly-resampler>=0.8.3.1 in /usr/local/lib/python3.10/dist-packages (from pycaret) (0.10.0)
Requirement already satisfied: statsmodels>=0.12.1 in /usr/local/lib/python3.10/dist-packages (from pycaret) (0.14.4)
Requirement already satisfied: sktime==0.26.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (0.26.0)
Requirement already satisfied: tbats>=1.1.3 in /usr/local/lib/python3.10/dist-packages (from pycaret) (1.1.3)
Requirement already satisfied: pmdarima>=2.0.4 in /usr/local/lib/python3.10/dist-packages (from pycaret) (2.0.4)
Requirement already satisfied: wurlitzer in /usr/local/lib/python3.10/dist-packages (from pycaret) (3.1.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from sktime==0.26.0->pycaret) (24.2)
Requirement already satisfied: scikit-base<0.8.0 in /usr/local/lib/python3.10/dist-packages (from sktime==0.26.0->pycaret) (0.7.8
Requirement already satisfied: patsy>=0.5.1 in /usr/local/lib/python3.10/dist-packages (from category-encoders>=2.4.0->pycaret) (
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from imbalanced-learn>=0.12.0->py
Requirement already satisfied: zipp>=3.20 in /usr/local/lib/python3.10/dist-packages (from importlib-metadata>=4.12.0->pycaret) (
Requirement already satisfied: setuptools>=18.5 in /usr/local/lib/python3.10/dist-packages (from ipython>=5.5.0->pycaret) (75.1.0
Requirement already satisfied: jedi>=0.16 in /usr/local/lib/python3.10/dist-packages (from ipython>=5.5.0->pycaret) (0.19.2)
Requirement already satisfied: decorator in /usr/local/lib/python3.10/dist-packages (from ipython>=5.5.0->pycaret) (4.4.2)
Requirement already satisfied: pickleshare in /usr/local/lib/python3.10/dist-packages (from ipython>=5.5.0->pycaret) (0.7.5)
Requirement already satisfied: traitlets>=4.2 in /usr/local/lib/python3.10/dist-packages (from ipython>=5.5.0->pycaret) (5.7.1)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from ipyt
Requirement already satisfied: pygments in /usr/local/lib/python3.10/dist-packages (from ipython>=5.5.0->pycaret) (2.18.0)
Requirement already satisfied: backcall in /usr/local/lib/python3.10/dist-packages (from ipython>=5.5.0->pycaret) (0.2.0)
Requirement already satisfied: matplotlib-inline in /usr/local/lib/python3.10/dist-packages (from ipython>=5.5.0->pycaret) (0.1.7
Requirement already satisfied: pexpect>4.3 in /usr/local/lib/python3.10/dist-packages (from ipython>=5.5.0->pycaret) (4.9.0)
Requirement already satisfied: ipykernel>=4.5.1 in /usr/local/lib/python3.10/dist-packages (from ipywidgets>=7.6.5->pycaret) (5.5
Requirement already satisfied: ipython-genutils~=0.2.0 in /usr/local/lib/python3.10/dist-packages (from ipywidgets>=7.6.5->pycare
Requirement already satisfied: widgetsnbextension~=3.6.0 in /usr/local/lib/python3.10/dist-packages (from ipywidgets>=7.6.5->pyca
Requirement already satisfied: jupyterlab-widgets>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from ipywidgets>=7.6.5->pyca
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib<3.8.0->pycaret) (1.3.
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib<3.8.0->pycaret) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib<3.8.0->pycaret) (4.5
```

Code and train model

```
# get our model

from pycaret.classification import *


df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15004 entries, 0 to 15003
Data columns (total 11 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   satisfaction_level     15004 non-null  float64
 1   last_evaluation        15004 non-null  float64
 2   number_project         14999 non-null  Int64  
 3   average_montly_hours   15004 non-null  Int64  
 4   time_spend_company     14999 non-null  Int64  
 5   Work_accident          15000 non-null  Int64  
 6   Quit_the_Company       15004 non-null  Int64  
 7   promotion_last_5years  15004 non-null  Int64  
 8   Departments            15004 non-null  object 
 9   salary                 15004 non-null  object 
 10  employee_id            15004 non-null  object 
dtypes: Int64(6), float64(2), object(3)
memory usage: 1.3+ MB
```

```
df.columns
```

```
Index(['satisfaction_level', 'last_evaluation', 'number_project',
       'average_montly_hours', 'time_spend_company', 'Work_accident',
       'Quit_the_Company', 'promotion_last_5years', 'Departments', 'salary',
       'employee_id'],
      dtype='object')
```

```
# setup or model
setup(df, target='Quit_the_Company',
      session_id=123,
      ignore_features=['employee_id'],
      categorical_features=['salary','Departments'])
```

	Description	Value
0	Session id	123
1	Target	Quit_the_Company
2	Target type	Binary
3	Original data shape	(15004, 11)
4	Transformed data shape	(15004, 21)
5	Transformed train set shape	(10502, 21)
6	Transformed test set shape	(4502, 21)
7	Ignore features	1
8	Numeric features	7
9	Categorical features	2
10	Rows with missing values	0.0%
11	Preprocess	True
12	Imputation type	simple
13	Numeric imputation	mean
14	Categorical imputation	mode
15	Maximum one-hot encoding	25
16	Encoding method	None
17	Fold Generator	StratifiedKFold
18	Fold Number	10
19	CPU Jobs	-1
20	Use GPU	False
21	Log Experiment	False
22	Experiment Name	clf-default-name
23	USI	da69

```
<pycaret.classification.oop.ClassificationExperiment at 0x7d37e94c6800>
```

```
compare_models()
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.9886	0.9912	0.9584	0.9934	0.9756	0.9681	0.9684	0.9180
lightgbm	Light Gradient Boosting Machine	0.9856	0.9932	0.9536	0.9856	0.9693	0.9599	0.9602	0.8610
xgboost	Extreme Gradient Boosting	0.9854	0.9922	0.9584	0.9801	0.9691	0.9596	0.9597	0.4410
et	Extra Trees Classifier	0.9832	0.9901	0.9461	0.9831	0.9641	0.9532	0.9536	1.0030
gbc	Gradient Boosting Classifier	0.9766	0.9884	0.9313	0.9694	0.9499	0.9346	0.9350	1.2600
dt	Decision Tree Classifier	0.9747	0.9683	0.9560	0.9392	0.9474	0.9307	0.9309	0.2120
ada	Ada Boost Classifier	0.9572	0.9809	0.9061	0.9138	0.9097	0.8816	0.8818	0.5950
knn	K Neighbors Classifier	0.9333	0.9666	0.9185	0.8230	0.8680	0.8236	0.8259	0.2930
qda	Quadratic Discriminant Analysis	0.8583	0.9101	0.8337	0.6671	0.7387	0.6437	0.6534	0.2680
lr	Logistic Regression	0.7844	0.8167	0.3244	0.5873	0.4174	0.2985	0.3185	1.4770
ridge	Ridge Classifier	0.7752	0.8129	0.2417	0.5684	0.3386	0.2288	0.2596	0.1480
lda	Linear Discriminant Analysis	0.7719	0.8129	0.2952	0.5399	0.3807	0.2558	0.2736	0.1500
dummy	Dummy Classifier	0.7617	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2640
svm	SVM - Linear Kernel	0.6888	0.7718	0.2681	0.2022	0.1770	0.0763	0.0907	0.5460
nb	Naive Bayes	0.6691	0.8046	0.8134	0.4040	0.5397	0.3245	0.3732	0.1400

```
RandomForestClassifier
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='sqrt',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_samples_leaf=1,
                       min_samples_split=2, min_weight_fraction_leaf=0.0,
                       monotonic_cst=None, n_estimators=100, n_jobs=-1,
                       oob_score=False, random_state=123, verbose=0,
                       warm_start=False)
```

```
rf_model = create_model('rf')
```

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9914	0.9946	0.9681	0.9959	0.9818	0.9762	0.9764
1	0.9933	0.9900	0.9721	1.0000	0.9859	0.9815	0.9817
2	0.9895	0.9930	0.9680	0.9878	0.9778	0.9709	0.9710
3	0.9857	0.9923	0.9520	0.9876	0.9695	0.9601	0.9604
4	0.9895	0.9890	0.9600	0.9959	0.9776	0.9708	0.9710
5	0.9886	0.9921	0.9560	0.9958	0.9755	0.9681	0.9684
6	0.9886	0.9936	0.9560	0.9958	0.9755	0.9681	0.9684
7	0.9819	0.9863	0.9440	0.9793	0.9613	0.9495	0.9498
8	0.9857	0.9885	0.9440	0.9958	0.9692	0.9599	0.9605
9	0.9914	0.9930	0.9641	1.0000	0.9817	0.9761	0.9764
Mean	0.9886	0.9912	0.9584	0.9934	0.9756	0.9681	0.9684
Std	0.0032	0.0025	0.0094	0.0062	0.0069	0.0090	0.0089

```
final_df = predict_model(rf_model)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Random Forest Classifier	0.9904	0.9909	0.9664	0.9933	0.9797	0.9734	0.9736

```
final_df.head()
```

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5
6949	0.42	0.56	2	143	3	0	
3760	0.62	0.52	3	148	3	0	
3460	0.37	0.45	2	149	3	0	
5785	0.78	0.98	5	263	6	0	
697	0.36	0.62	4	111	6	0	

```
final_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4502 entries, 6949 to 14769
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   satisfaction_level    4502 non-null   float32
 1   last_evaluation       4502 non-null   float32
 2   number_project        4502 non-null   Int64
 3   average_montly_hours  4502 non-null   Int64
 4   time_spend_company    4502 non-null   Int64
 5   Work_accident         4502 non-null   Int64
 6   promotion_last_5years 4502 non-null   Int64
 7   Departments           4502 non-null   category
 8   salary                4502 non-null   category
 9   Quit_the_Company      4502 non-null   Int64
 10  prediction_label      4502 non-null   int64
 11  prediction_score      4502 non-null   float64
dtypes: Int64(6), category(2), float32(2), float64(1), int64(1)
memory usage: 387.4 KB
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15004 entries, 0 to 15003
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   satisfaction_level    15004 non-null  float64
 1   last_evaluation       15004 non-null  float64
 2   number_project        14999 non-null  Int64
 3   average_montly_hours  15004 non-null  Int64
 4   time_spend_company    14999 non-null  Int64
 5   Work_accident         15000 non-null  Int64
 6   Quit_the_Company      15004 non-null  Int64
 7   promotion_last_5years 15004 non-null  Int64
 8   Departments           15004 non-null  object
 9   salary                15004 non-null  object
 10  employee_id           15004 non-null  object
dtypes: Int64(6), float64(2), object(3)
memory usage: 1.3+ MB
```

```
new_predictions = predict_model(rf_model, data = df2)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Random Forest Classifier	0.9300	0	0.0000	0.0000	0.0000	0.0000	0.0000

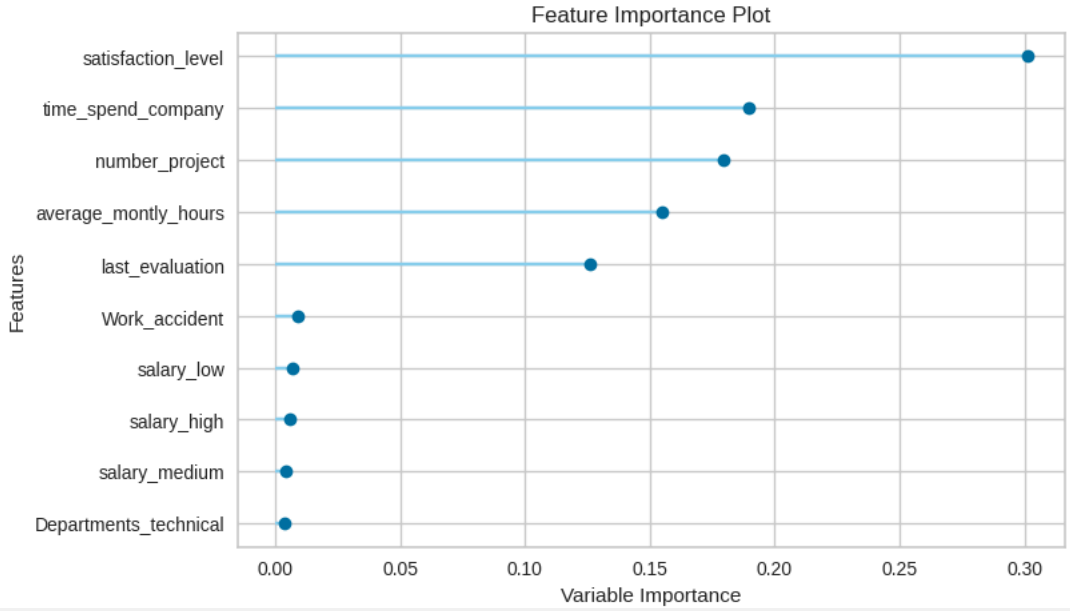
```
new_predictions.head()
```

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5yea
0	0.331690	0.847953	6	151	4	0	
1	0.468434	0.169659	5	303	4	0	
2	0.858448	0.918311	4	162	6	1	
3	0.056211	0.322600	2	229	5	1	
4	0.382648	0.434348	4	260	3	1	

```
new_predictions.to_gbq('employeedata.pilot_predictions',
                       project_id,
                       chunksize=None,
                       if_exists='replace')
```

```
100%|██████████| 1/1 [00:00<00:00, 5518.82it/s]
```

```
plot_model(rf_model,plot='feature')
```




```
# create a feature table
rf_model.feature_names_in_
```

```
array(['satisfaction_level', 'last_evaluation', 'number_project',
       'average_montly_hours', 'time_spend_company', 'Work_accident',
       'promotion_last_5years', 'Departments_accounting',
       'Departments_support', 'Departments_technical',
       'Departments_sales', 'Departments_RandD', 'Departments_IT',
       'Departments_hr', 'Departments_product_mng',
       'Departments_marketing', 'Departments_management', 'salary_medium',
       'salary_low', 'salary_high'], dtype=object)
```

```
rf_model.feature_importances_
```

```
array([0.30110658, 0.1259243 , 0.17920058, 0.15511409, 0.18978115,
       0.0088762 , 0.0015364 , 0.00194865, 0.00290416, 0.00367545,
       0.00352838, 0.00186797, 0.00168919, 0.00183491, 0.00148181,
       0.00135336, 0.00152958, 0.00396984, 0.00693197, 0.00574544])
```

```
import pandas as pd
feature_table = pd.DataFrame(zip(rf_model.feature_names_in_,rf_model.feature_importances_),
                             columns=['feature','importance'])
feature_table
```

	feature	importance
0	satisfaction_level	0.301107
1	last_evaluation	0.125924
2	number_project	0.179201
3	average_montly_hours	0.155114
4	time_spend_company	0.189781
5	Work_accident	0.008876
6	promotion_last_5years	0.001536
7	Departments_accounting	0.001949