Employee-Attrition

This is a fictional data set created by IBM data scientists
https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Jaime Suazo
Data Science enthusiast
https://github.com/suazojaime

Acerca del dataset

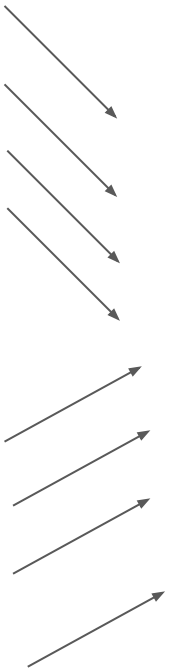
```
df = pd.read_csv('WA_Fn-UseC_-HR-Employee-Attrition.csv')
df
```

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education
0	41	Yes	Travel_Rarely	1102	Sales	1	2
1	49	No	Travel_Frequently	279	Research & Development	8	1
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2
3	33	No	Travel_Frequently	1392	Research & Development	3	4
4	27	No	Travel_Rarely	591	Research & Development	2	1
...	...	...	...	...	...	...	...
1465	36	No	Travel_Frequently	884	Research & Development	23	2
1466	39	No	Travel_Rarely	613	Research & Development	6	1
1467	27	No	Travel_Rarely	155	Research & Development	4	3
1468	49	No	Travel_Frequently	1023	Sales	2	3
1469	34	No	Travel_Rarely	628	Research & Development	8	3

1470 rows × 35 columns

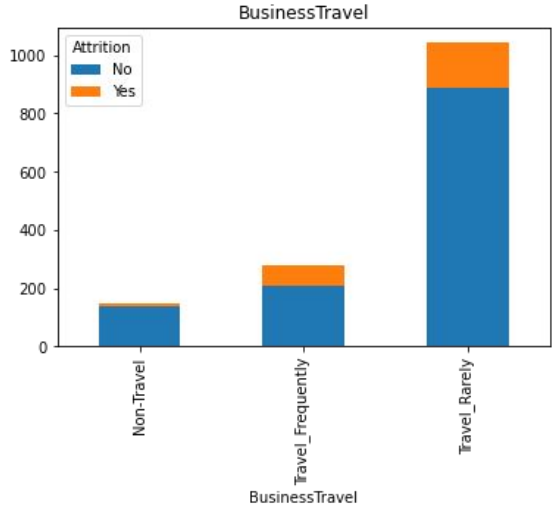
```
df.dtypes
```

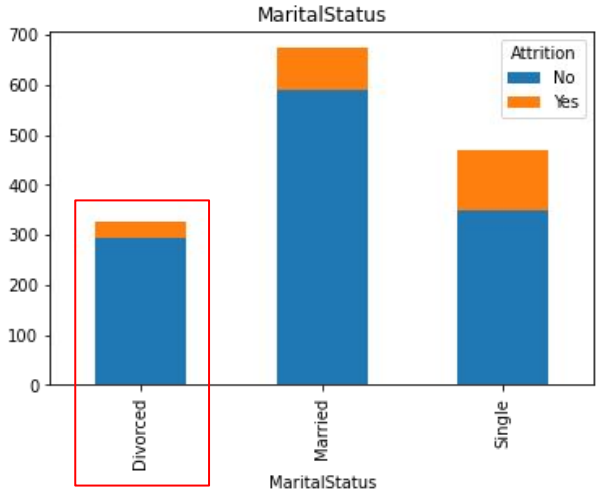
```
Age                         int64
Attrition                  object
BusinessTravel             object
DailyRate                   int64
Department                 object
DistanceFromHome            int64
Education                   int64
EducationField             object
EmployeeCount               int64
EmployeeNumber              int64
EnvironmentSatisfaction     int64
Gender                     object
HourlyRate                  int64
JobInvolvement              int64
JobLevel                    int64
JobRole                    object
JobSatisfaction             int64
MaritalStatus              object
MonthlyIncome               int64
MonthlyRate                 int64
NumCompaniesWorked          int64
Over18                     object
OverTime                   object
PercentSalaryHike           int64
PerformanceRating           int64
RelationshipSatisfaction    int64
StandardHours               int64
StockOptionLevel            int64
TotalWorkingYears           int64
TrainingTimesLastYear       int64
WorkLifeBalance             int64
YearsAtCompany              int64
YearsInCurrentRole          int64
YearsSinceLastPromotion     int64
YearsWithCurrManager        int64
dtype: object
```

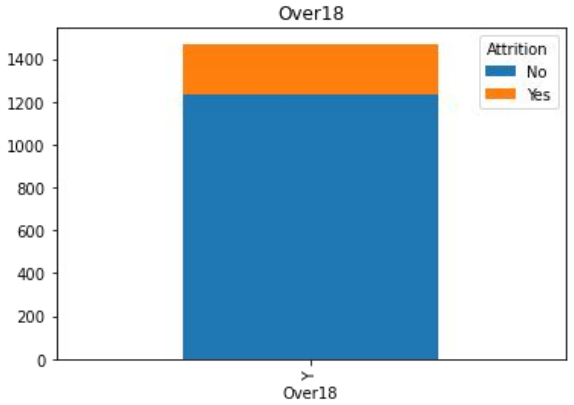


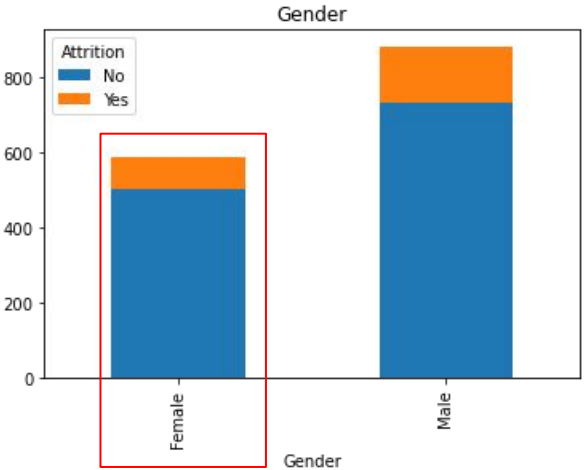
Target
Attrition

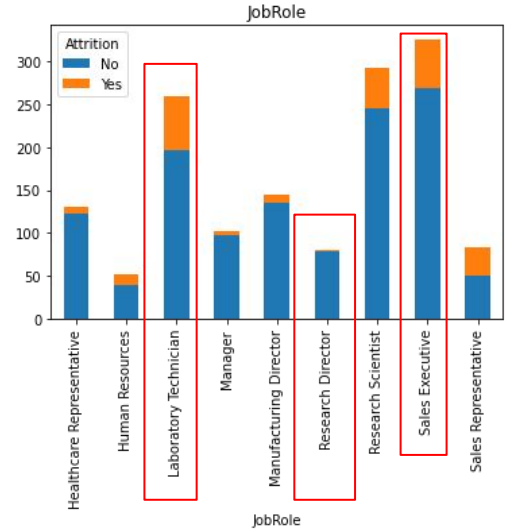
Distribución de registros categóricos por …



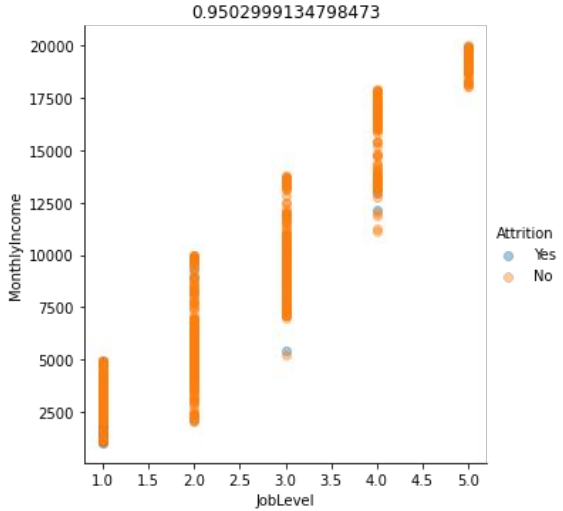


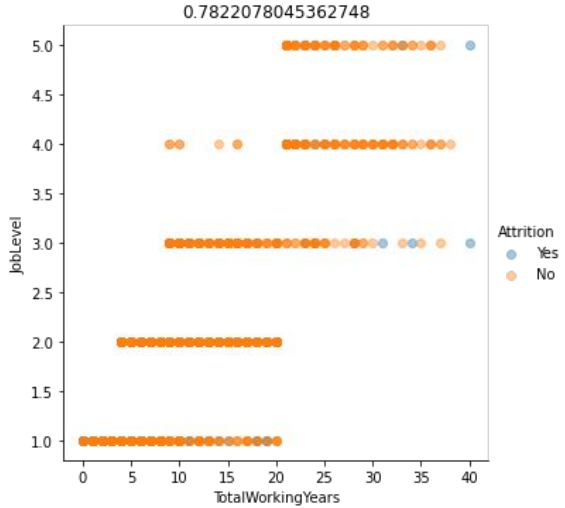


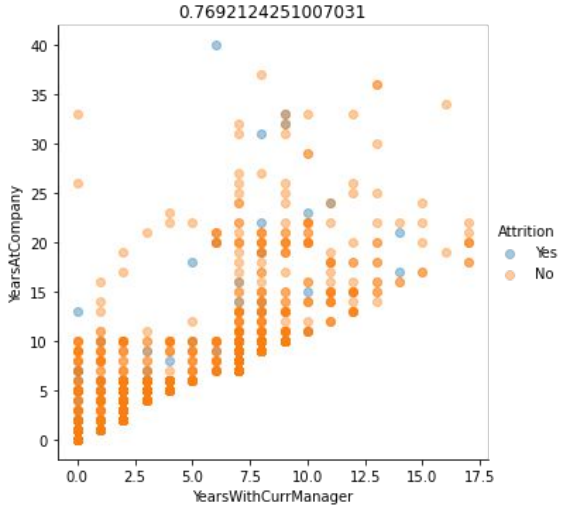




Scatter de registros numéricos por …







Datos agregados y desagregados

```
X_agregated
```

	Age	BusinessTravel	DailyRate	Departmen
0	41	Travel_Rarely	1102	Sales
1	49	Travel_Frequently	279	Research & Developmen
2	37	Travel_Rarely	1373	Research & Developmen
3	33	Travel_Frequently	1392	Research & Developmen
4	27	Travel_Rarely	591	Research & Developmen
...	...	...	...	..
1465	36	Travel_Frequently	884	Research & Developmen
1466	39	Travel_Rarely	613	Research & Developmen
1467	27	Travel_Rarely	155	Research & Developmen
1468	49	Travel_Frequently	1023	Sales
1469	34	Travel_Rarely	628	Research & Developmen

1470 rows × 31 columns

```
X_disagregated
```

	Age	DailyRate	DistanceFromHome	Education	EmployeeNumber	EnvironmentSatisfaction	H
0	41	1102	1	2	1	2	
1	49	279	8	1	2	3	
2	37	1373	2	2	4	4	
3	33	1392	3	4	5	4	
4	27	591	2	1	7	1	
...	...	...	...	...	...	...	
1465	36	884	23	2	2061	3	
1466	39	613	6	1	2062	4	
1467	27	155	4	3	2064	2	
1468	49	1023	2	3	2065	4	
1469	34	628	8	3	2068	2	

1470 rows × 52 columns

RandomForest (datos agregados)

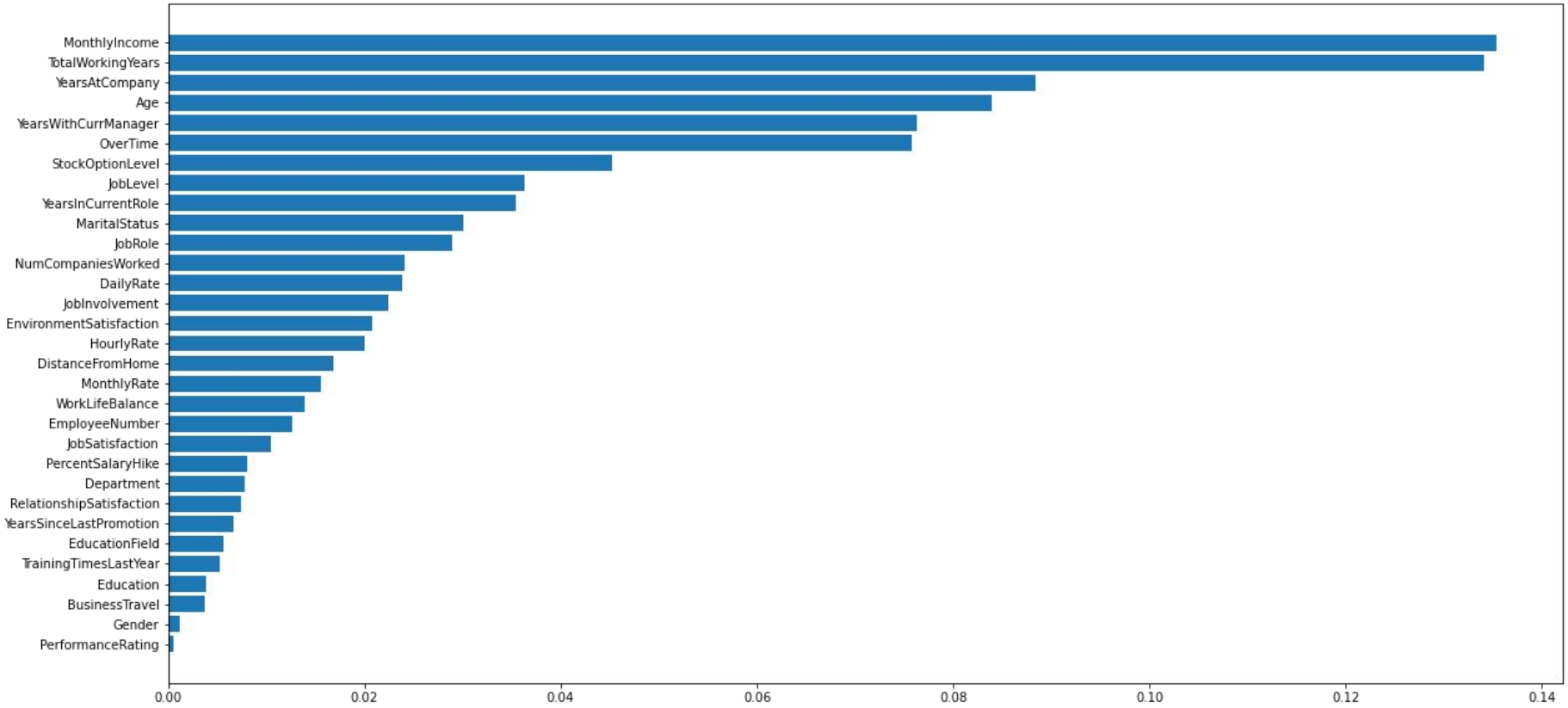
```
rfc = RandomForestClassifier(bootstrap= True,
                             max_depth= 10,
                             max_features= 'sqrt',
                             min_samples_leaf= 2,
                             min_samples_split= 5,
                             n_estimators= 500)
rfc.fit(X_agregated_train, y_train)
print(rfc.score(X_agregated_train, y_train))
print(rfc.score(X_agregated_test, y_test))
```

```
0.9514091350826045
0.8299319727891157
```

```
rfc = RandomForestClassifier(bootstrap= True,
                             max_depth= 3,
                             max_features= 'sqrt',
                             min_samples_leaf= 2,
                             min_samples_split= 5,
                             n_estimators= 500)
rfc.fit(X_agregated_train, y_train)
print(rfc.score(X_agregated_train, y_train))
print(rfc.score(X_agregated_test, y_test))
```

```
0.8639455782312925
0.8299319727891157
```

feature_importances

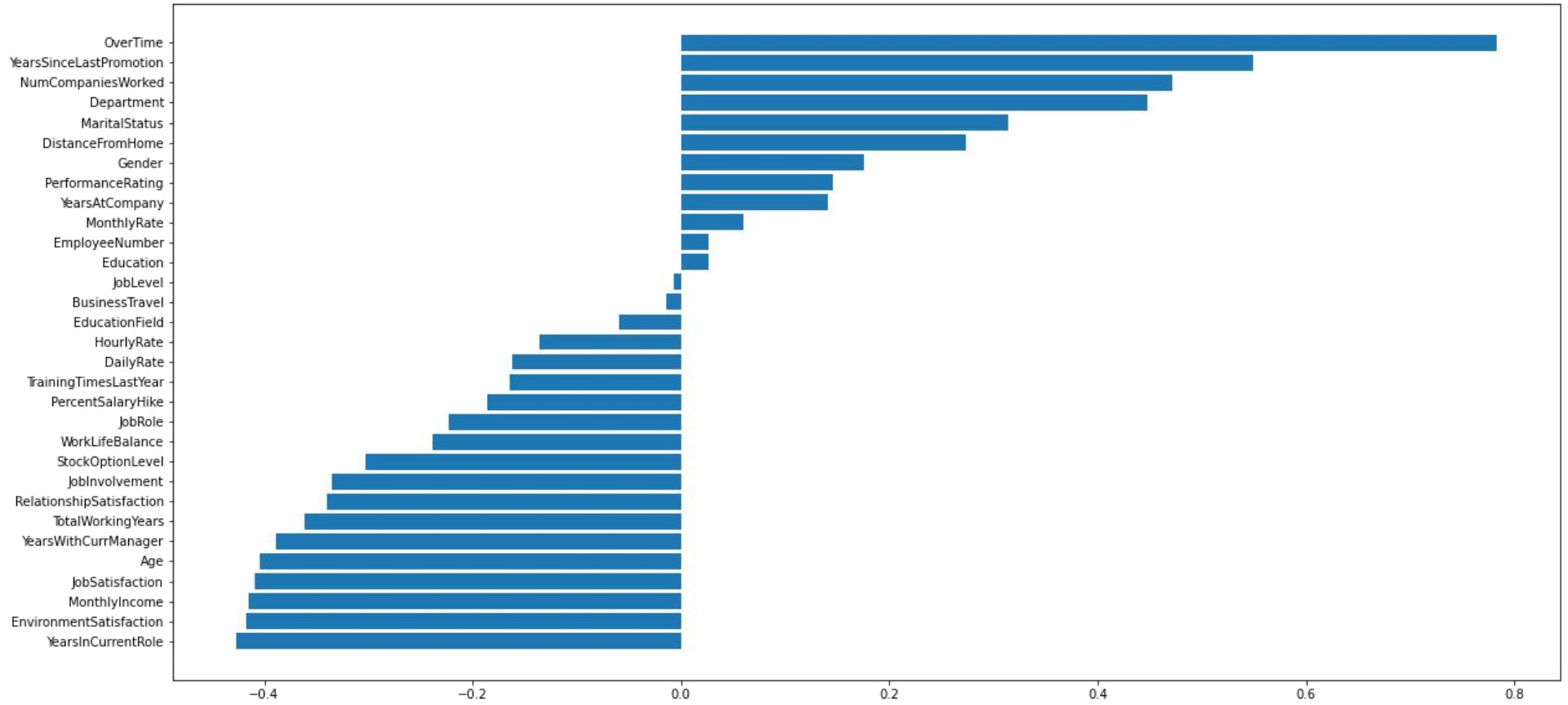


LogisticRegression (datoa agregados)

```
lr = LogisticRegression(random_state=42, multi_class='ovr').fit(X_agregated_train, y_train)
print(lr.score(X_agregated_train, y_train))
print(lr.score(X_agregated_test, y_test))
```

```
0.8794946550048591
0.8662131519274376
```

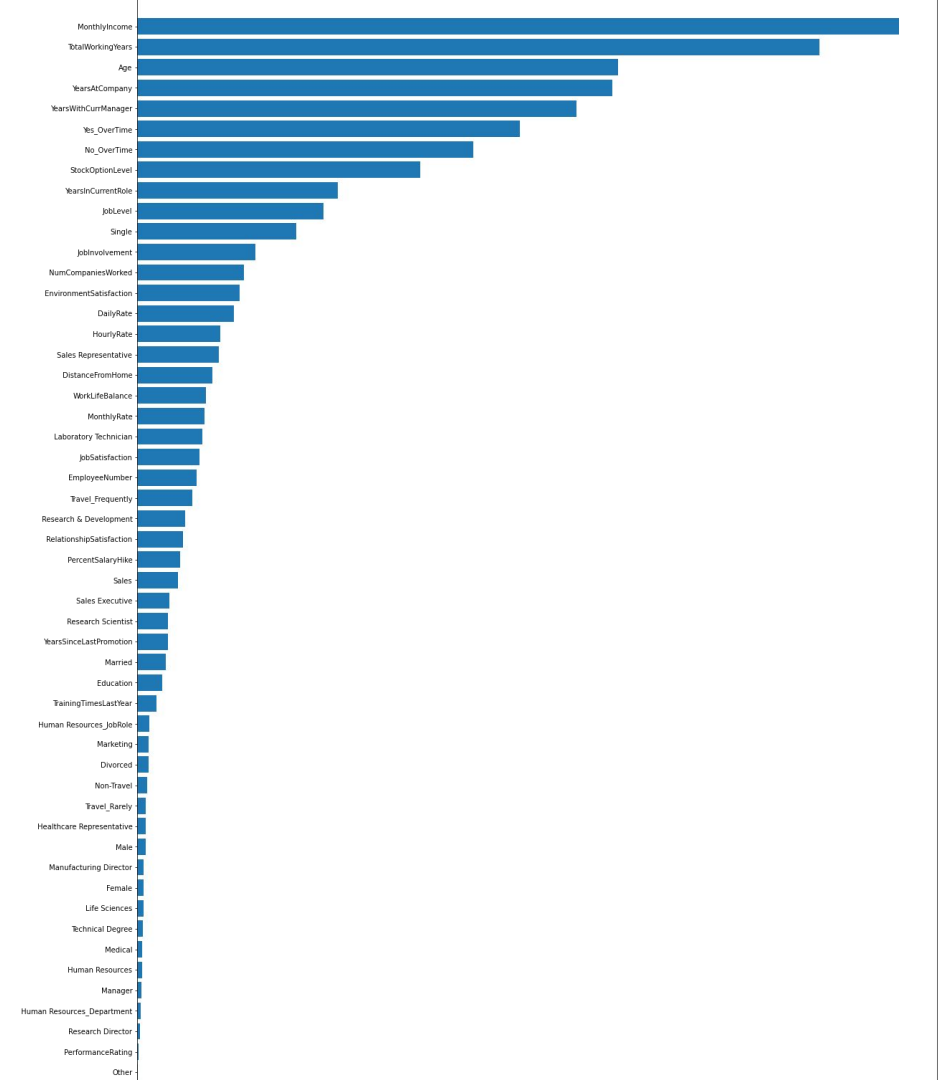
Coef_


OverTime
YearsSinceLastPromotion
NumCompaniesWorked
Department
MaritalStatus
DistanceFromHome
Gender
PerformanceRating
YearsAtCompany
MonthlyRate
EmployeeNumber
Education
JobLevel
BusinessTravel
EducationField
HourlyRate
DailyRate
TrainingTimesLastYear
PercentSalaryHike
JobRole
WorkLifeBalance
StockOptionLevel
JobInvolvement
RelationshipSatisfaction
TotalWorkingYears
YearsWithCurrManager
Age
JobSatisfaction
MonthlyIncome
EnvironmentSatisfaction
YearsInCurrentRole
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

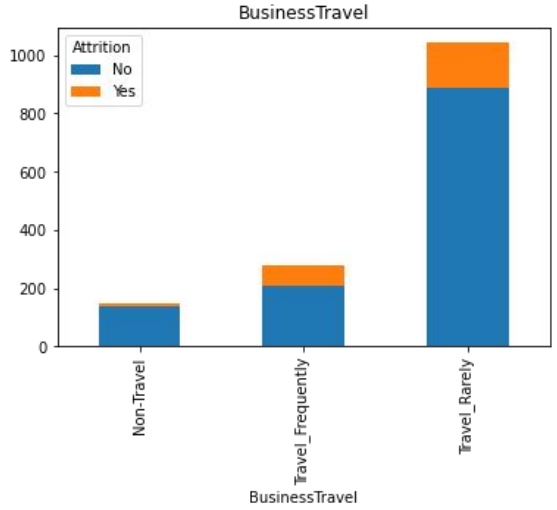
RandomForest (datos desagregados)

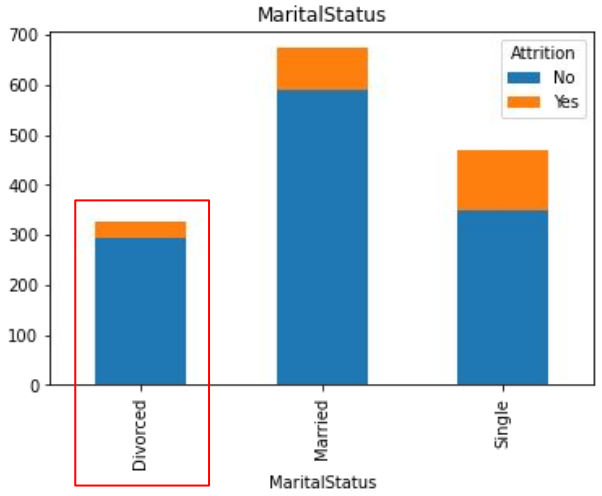
```
rfc = RandomForestClassifier(bootstrap= True,
                             max_depth= 3,
                             max_features= 'sqrt',
                             min_samples_leaf= 4,
                             min_samples_split= 10,
                             n_estimators= 1000)
rfc.fit(X_disagregated_train, y_train)
print(rfc.score(X_disagregated_train, y_train))
print(rfc.score(X_disagregated_test, y_test))
```

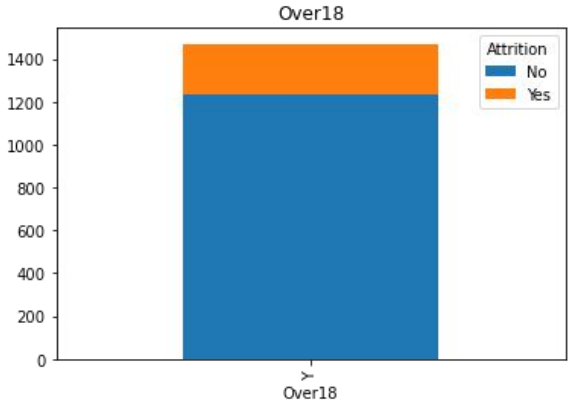
```
0.8610301263362488
0.8390022675736961
```

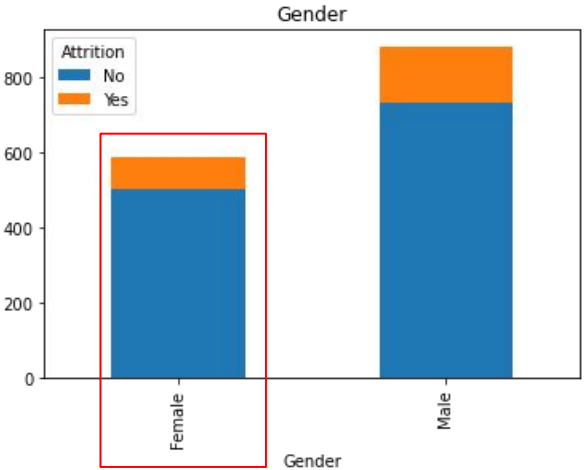

MonthlyIncome
TotalWorkingYears
Age
YearsAtCompany
YearsWithCurrManager
Yes_OverTime
No_OverTime
StockOptionLevel
YearsInCurrentRole
JobLevel
Single
JobInvolvement
NumCompaniesWorked
EnvironmentSatisfaction
DailyRate
HourlyRate
Sales Representative
DistanceFromHome
WorkLifeBalance
MonthlyRate
Laboratory Technician
JobSatisfaction
EmployeeNumber
Travel_Frequently
Research & Development
RelationshipSatisfaction
PercentSalaryHike
Sales
Sales Executive
Research Scientist
YearsSinceLastPromotion
Married
Education
TrainingTimesLastYear
Human Resources_JobRole
Marketing
Divorced
Non-Travel
Travel_Rarely
Healthcare Representative
Male
Manufacturing Director
Female
Life Sciences
Technical Degree
Medical
Human Resources
Manager
Human Resources_Department
Research Director
PerformanceRating
Other

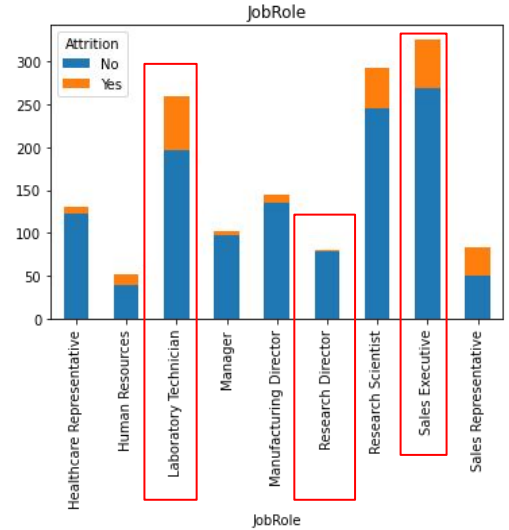
Distribución de registros categóricos por …











Logistic regresion (datos desagregados)