Análise de dados - Attrition

Sobre os dados: o dataset possui 35 atributos e 1470 registros.

Campo	**Tipo**	**Campo**	**Tipo**
Age	Numérica - Discreta	PerformanceRating	Categórica
Attrition	Categórica	RelationshipSatisfaction	Categórica
BusinessTravel	Categórica	StandardHours	Numérica - Discreta
DailyRate	Numérica - Discreta	StockOptionLevel	Categórica
Department	Categórica	TotalWorkingYears	Numérica - Discreta
DistanceFromHome	Numérica - Discreta	TrainingTimesLastYear	Numérica - Discreta
Education	Categórica	WorkLifeBalance	Categórica
EducationField	Categórica	YearsAtCompany	Numérica - Discreta
EmployeeCount	Numérica - Discreta	YearsInCurrentRole	Numérica - Discreta
EmployeeNumber	Numérica - Discreta	YearsSinceLastPromotion	Numérica - Discreta
EnvironmentSatisfaction	Categórica	YearsWithCurrManager	Numérica - Discreta
Gender	Categórica		
HourlyRate	Numérica - Discreta		
JobInvolvement	Categórica		
JobLevel	Categórica		
JobRole	Categórica		
JobSatisfaction	Categórica		
MaritalStatus	Categórica		
MonthlyIncome	Numérica - Discreta		
MonthlyRate	Numérica - Discreta		
NumCompaniesWorked	Numérica - Discreta		
Over18	Categórica		
OverTime	Categórica		
PercentSalaryHike	Numérica - Discreta		

Outros detalhes:	
Education	RelationshipSatisfaction
1 'Below College'	1 'Low'
2 'College'	2 'Medium'
3 'Bachelor'	3 'High'
4 'Master'	4 'Very High'
5 'Doctor'	WorkLifeBalance
EnvironmentSatisfaction	1 'Bad'
1 'Low'	2 'Good'
2 'Medium'	3 'Better'
3 'High'	4 'Best'
4 'Very High'	
JobInvolvement	
1 'Low'	
2 'Medium'	
3 'High'	
4 'Very High'	
JobSatisfaction	
1 'Low'	
2 'Medium'	
3 'High'	
4 'Very High'	
PerformanceRating	
1 'Low'	
2 'Good'	
3 'Excellent'	
4 'Outstanding'	

O primeiro passo foi descobrir qual dos atributos categóricos tem uma relação estatisticamente significativa com Attrition, que é o atributo que queremos avaliar. Para isto, foi realizado o teste Qui-quadrado que mede a relação de dependência entre duas variáveis categóricas, verificando como os valores esperados desviam dos valores observados.

Quando temos um alto valor de Qui-quadrado (nosso p-value será baixo), significa que temos evidência estatística para inferir que os valores observados e esperados não são os mesmos, portanto possuem dependência entre si. Quanto mais alto o Qui-quadrado, maior a dependência entre as variáveis.

	Attrition	p < 0.05
Over18	1.0	False
PerformanceRating	0.990075	False
Education	0.545525	False
Gender	0.290572	False
RelationshipSatisfaction	0.154972	False
EducationField	0.006774	True
Department	0.004526	True
WorkLifeBalance	0.000973	True
JobSatisfaction	0.000556	True
EnvironmentSatisfaction	0.000051	True
BusinessTravel	0.000006	True
JobInvolvement	0.000003	True
MaritalStatus	0.0	True
StockOptionLevel	0.0	True
JobLevel	0.0	True
JobRole	0.0	True
OverTime	0.0	True

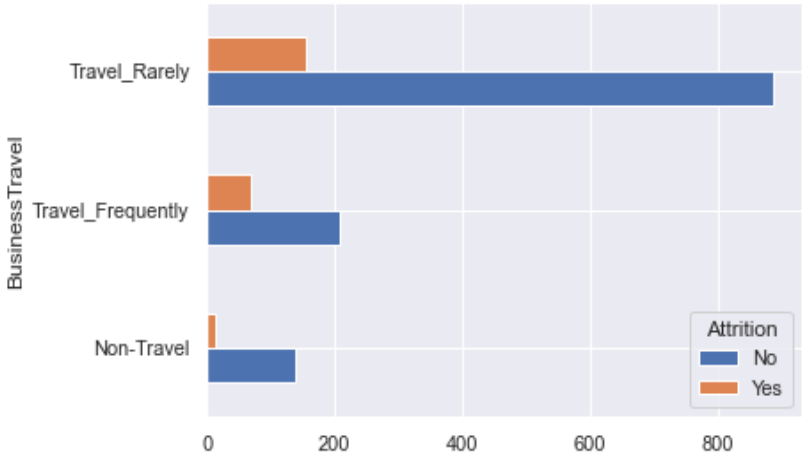
Assim, temos os atributos categóricos divididos em duas categorias:

- Sem dependência com Attrition:
    - Over18
    - PerformanceRating
    - Education
    - Gender
    - RelationshipSatisfaction
- Com dependência com Attrition:
    - EducationField
    - Department
    - WorkLifeBalance
    - JobSatisfaction

- EnvironmentSatisfaction
- BusinessTravel
- JobInvolvement
- MaritalStatus
- StockOptionLevel
- JobLevel
- JobRole
- OverTime

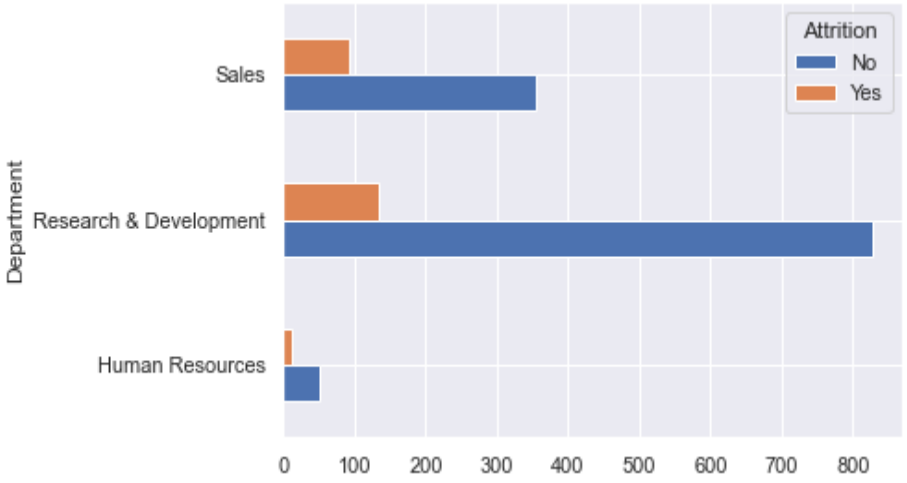
Agora podemos criar gráficos para analisar como esses atributos estão relacionados com Attrition (desgaste)

A maioria dos funcionários com desgaste (Attrition) raramente viaja



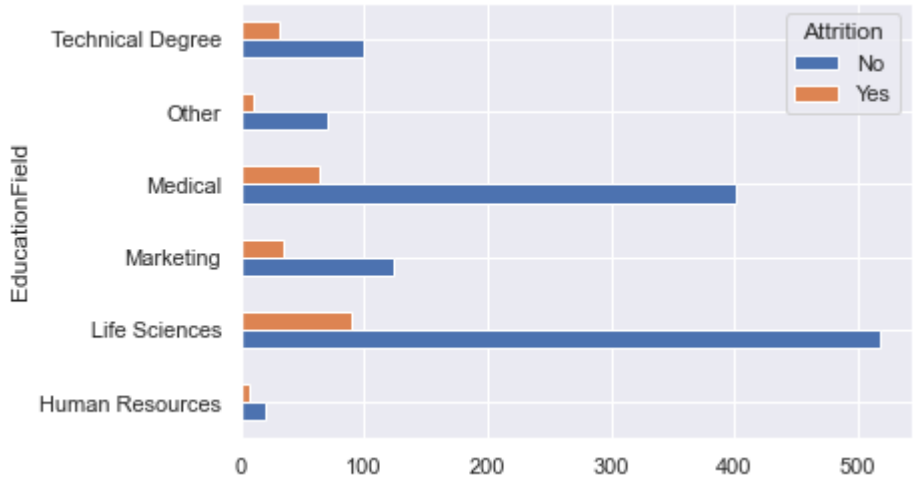
Attrition	No	Yes
BusinessTravel		
Non-Travel	11.192214	5.063291
Travel_Frequently	16.869424	29.113924
Travel_Rarely	71.938362	65.822785

Os departamentos Research & Development e Sales são responsáveis por cerca de 95% dos funcionários com desgaste



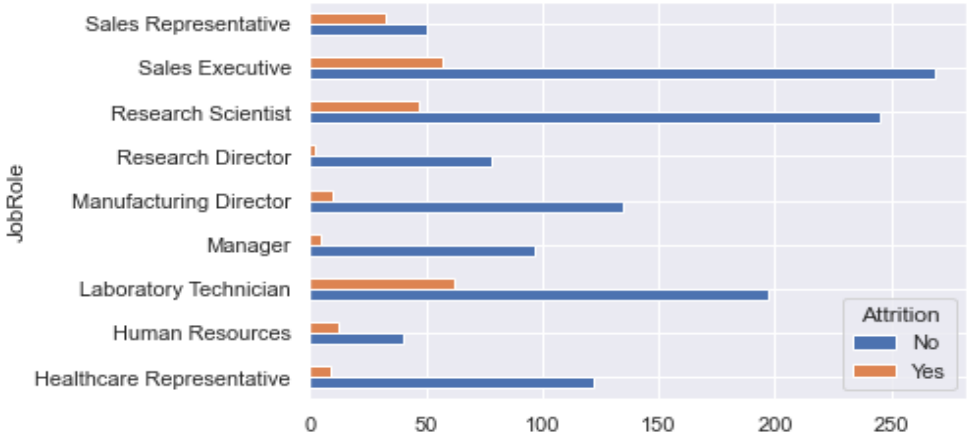
Attrition	No	Yes
Department		
Human Resources	4.136253	5.063291
Research & Development	67.153285	56.118143
Sales	28.710462	38.818565

O campo da educação de mais da metade dos funcionários com desgaste é Life Sciences e Medical



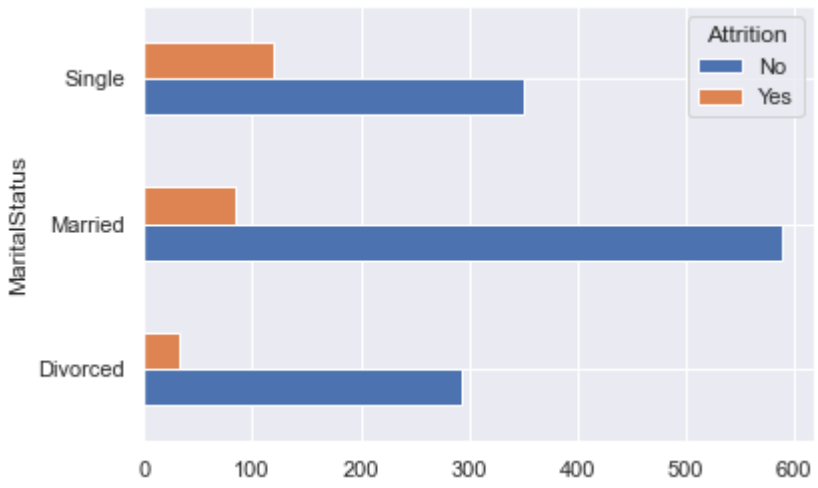
Attrition	No	Yes
EducationField		
Human Resources	1.622060	2.953586
Life Sciences	41.930251	37.552743
Marketing	10.056772	14.767932
Medical	32.522303	26.582278
Other	5.758313	4.641350
Technical Degree	8.110300	13.502110

Cerca de 70% dos funcionários com desgaste possuem os seguintes cargos laboratory technicians, sale executives, and research scientists



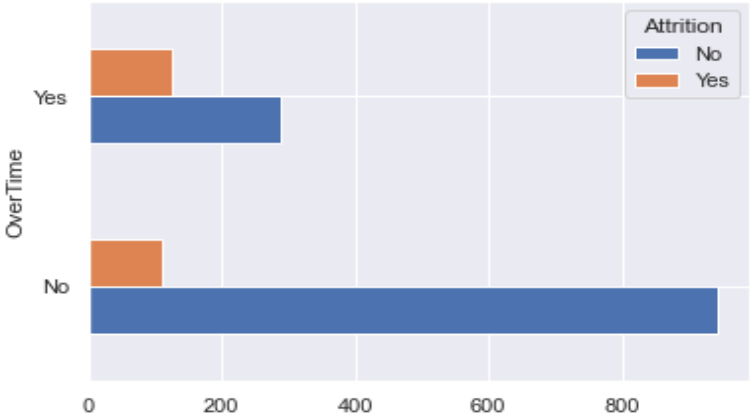
Attrition	No	Yes
JobRole		
Healthcare Representative	9.894566	3.797468
Human Resources	3.244120	5.063291
Laboratory Technician	15.977291	26.160338
Manager	7.866991	2.109705
Manufacturing Director	10.948905	4.219409
Research Director	6.326034	0.843882
Research Scientist	19.870235	19.831224
Sales Executive	21.816707	24.050633
Sales Representative	4.055150	13.924051

50 % dos funcionários com desgaste são solteiros



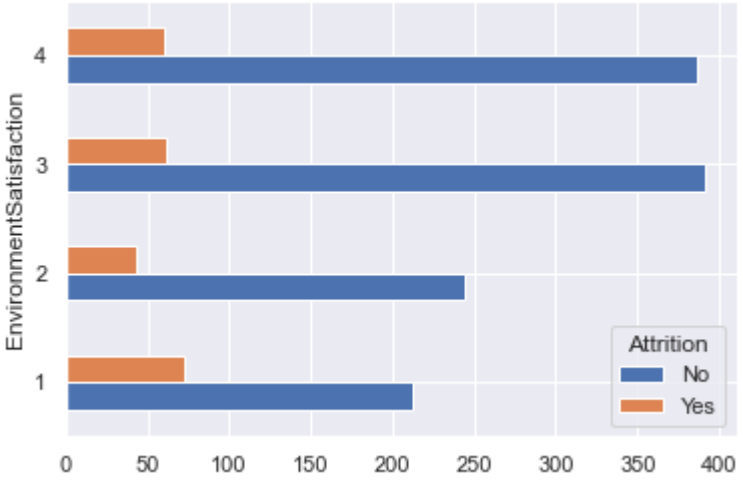
Attrition	No	Yes
MaritalStatus		
Divorced	23.844282	13.924051
Married	47.769667	35.443038
Single	28.386050	50.632911

A maioria dos funcionários com desgaste não fazem horas extras



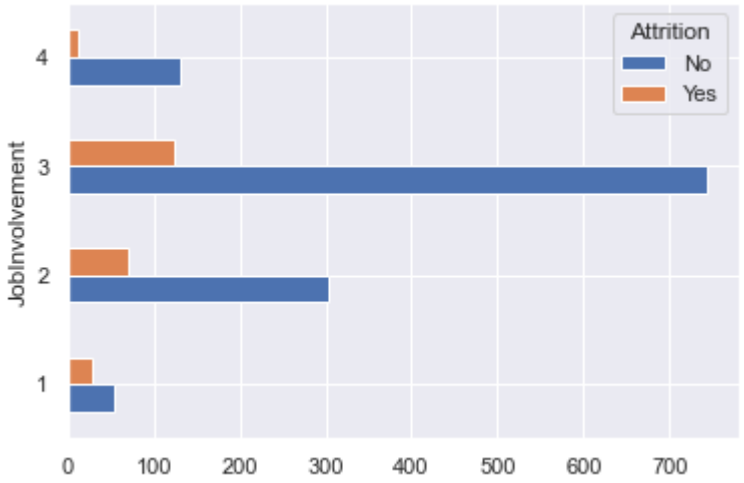
Attrition	No	Yes
OverTime		
No	76.561233	46.413502
Yes	23.438767	53.586498

Cerca de 63% das avaliações de satisfação com o ambiente, de funcionários sem desgaste, é como Alta e Muito Alta



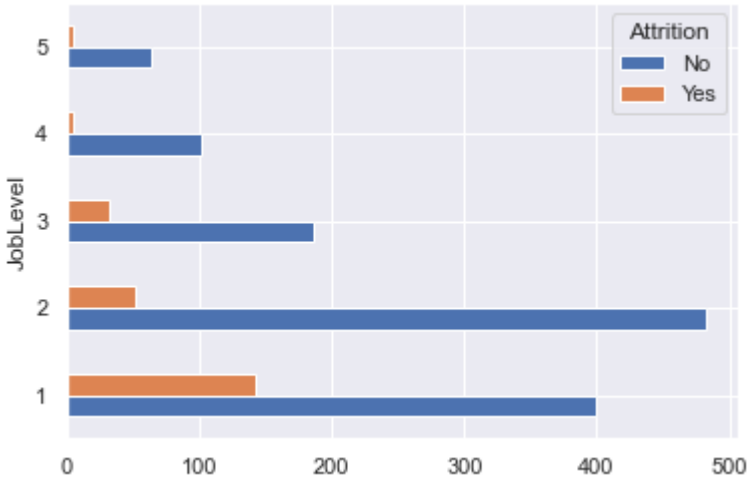
Attrition	No	Yes
EnvironmentSatisfaction		
1	17.193836	30.379747
2	19.789132	18.143460
3	31.711273	26.160338
4	31.305758	25.316456

Cerca de 80% dos funcionários com desgaste possuem envolvimento com o trabalho moderado e alto



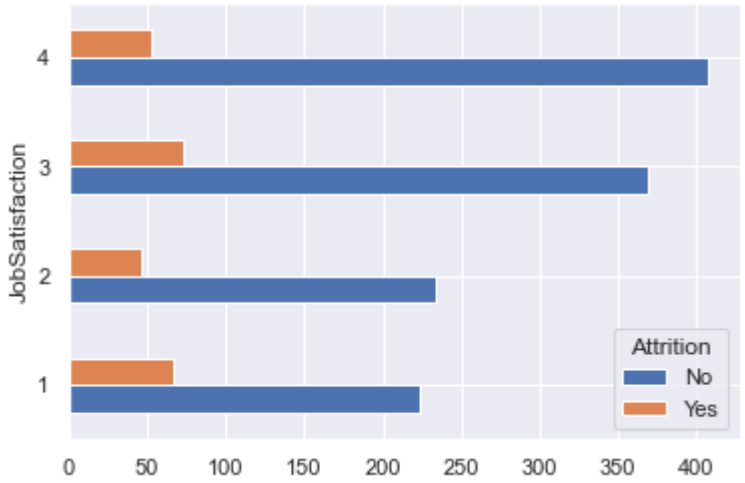
Attrition	No	Yes
JobInvolvement		
1	4.460665	11.814346
2	24.655312	29.957806
3	60.259530	52.742616
4	10.624493	5.485232

60% dos funcionários com desgaste possuem nível 1



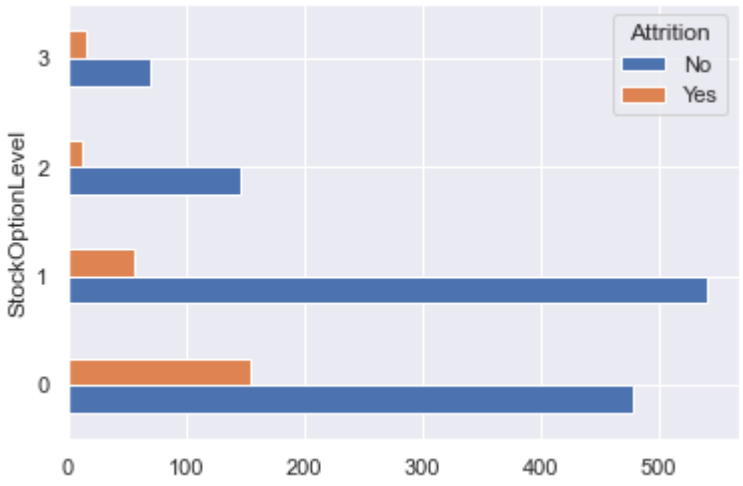
Attrition	No	Yes
JobLevel		
1	32.441200	60.337553
2	39.091646	21.940928
3	15.085158	13.502110
4	8.191403	2.109705
5	5.190592	2.109705

Cerca de 60% dos funcionários sem desgaste avaliam a satisfação com o trabalho como alta e muito alta



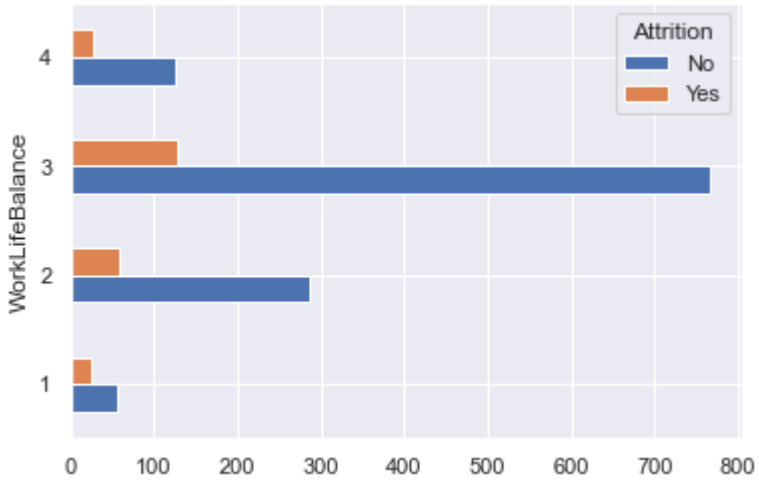
Attrition	No	Yes
JobSatisfaction		
1	18.085969	27.848101
2	18.978102	19.409283
3	29.927007	30.801688
4	33.008921	21.940928

65% dos funcionários com desgaste possuem nível de opções de ações igual a 0



Attrition	No	Yes
StockOptionLevel		
0	38.686131	64.978903
1	43.795620	23.628692
2	11.841038	5.063291
3	5.677210	6.329114

Cerca de 77% dos funcionários com desgaste classificam o equilíbrio entre a vida pessoal e profissional como boa e melhor



Attrition	No	Yes
WorkLifeBalance		
1	4.460665	10.548523
2	23.195458	24.472574
3	62.124899	53.586498
4	10.218978	11.392405

Análise dos Atributos (Colunas) numéricas em relação ao Attrition

A partir destas visualizações podemos perceber coisas como: maioria dos funcionários com desgaste então entre 25 e 40 anos, a distância de casa está entre 0 e 5, o percentual de aumento do salário é 12, a quantidade de anos desde a última promoção está entre 0 e 2,5, renda mensal inferior a 5000.


Age
DailyRate
DistanceFromHome
EmployeeCount
EmployeeNumber
HourlyRate
MonthlyIncome
MonthlyRate
NumCompaniesWorked
PercentSalaryHike
StandardHours
TotalWorkingYears
TrainingTimesLastYear
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager

Prevendo o Desgaste

Para prever o desgaste, primeiramente irei transformar os atributos do tipo object em inteiro por meio da função get_dummies do pandas. Esta função recebe uma lista, ou uma string que será convertida em lista a partir de um separador, e cria uma coluna para cada dummie (atributo categorico) com um dado número que representa aquele dummie em uma data linha.

Posteriormente foi aplicado o SMOTE (Synthetic Minority Over-sampling Technique) para gerar dados sintéticos da classe minoritária. Em seguida foram separados dos dados em 75% para treino e 25% para teste.

Random Forest

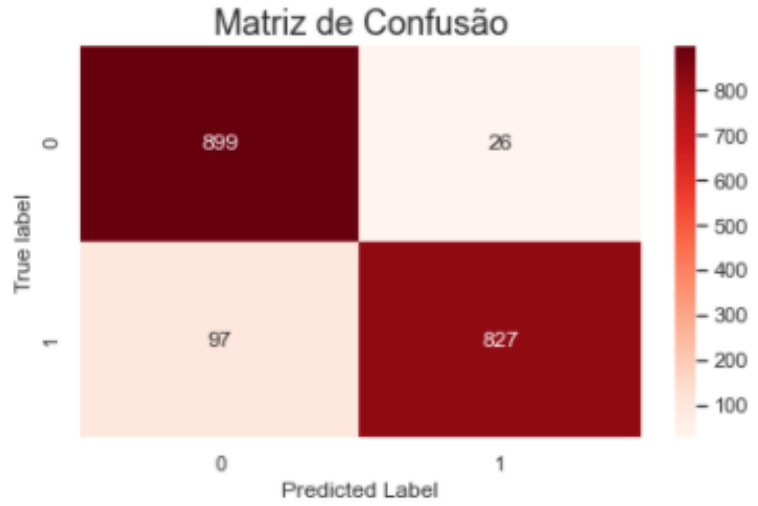
Dizendo de modo simples: o algoritmo de florestas aleatórias cria várias árvores de decisão e as combina para obter uma predição com maior acurácia e mais estável.

O algoritmo Random Forest foi setado com parâmetros padrões da biblioteca, exceto pelo n_estimetors que ficou igual a 900. No treinamento foi feita a validação cruzada k-fold com k = 10.

O Random Forest obteve os seguintes resultados:

Relatório de classificação: modelo Random Forest

	precision	recall	f1-score	support
No	0.9026	0.9719	0.9360	925
Yes	0.9695	0.8950	0.9308	924
accuracy			0.9335	1849
macro avg	0.9361	0.9335	0.9334	1849
weighted avg	0.9360	0.9335	0.9334	1849



Conclusão

Neste trabalho foram feitas algumas análises dos dados para entender o que pode afetar o desgaste de funcionários. Bem como a implementação de um algoritmos de machine learn, o Random Forest que obteve um recall de 97,5% para funcionários sem desgaste e 89,5% para funcionários com desgaste.