**Habilidades desenvolvidas: Identificação e uso de técnicas supervisionadas e Não Supervisionadas**

**Desafio Proposto:** A partir de uma crescente taxa de turnover a empresa ACME solicita a sua área de cientistas que indique através de modelos adequados quais os colaboradores ela deve fazer sua ação de retenção. Foi disponibilizada uma base de seus colaboradores para análise.

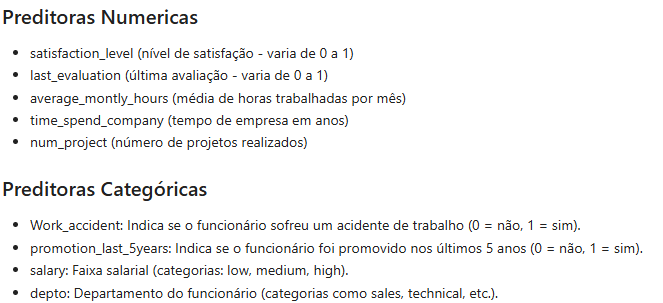
1. **Utilize a base de dados HR\_Abandono2023.**
2. **Qual o objetivo deste estudo :**

R: Este estudo tem como objetivo identificar quais variáveis podem influenciar no turnover da empresa ACME para identificar quais colaboradores têm uma propensão maior de abandonar a equipe;

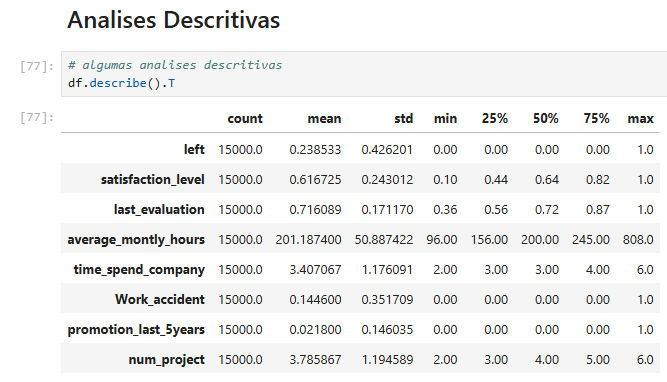
1. **Porque nesse estudo é indicado utilizar técnicas supervisionadas de Classificação :**

R: Neste caso, é recomendável utilizar a técnica de classificação pois já existe um dado (variável Target) no qual eu posso utilizar no modelo para aprender qual a relação entre a as características dos dados e os rótulos e assim entender quais são as correlações das causas de abandono;

1. **Classifique as variáveis: dependentes e independentes(preditoras)** :



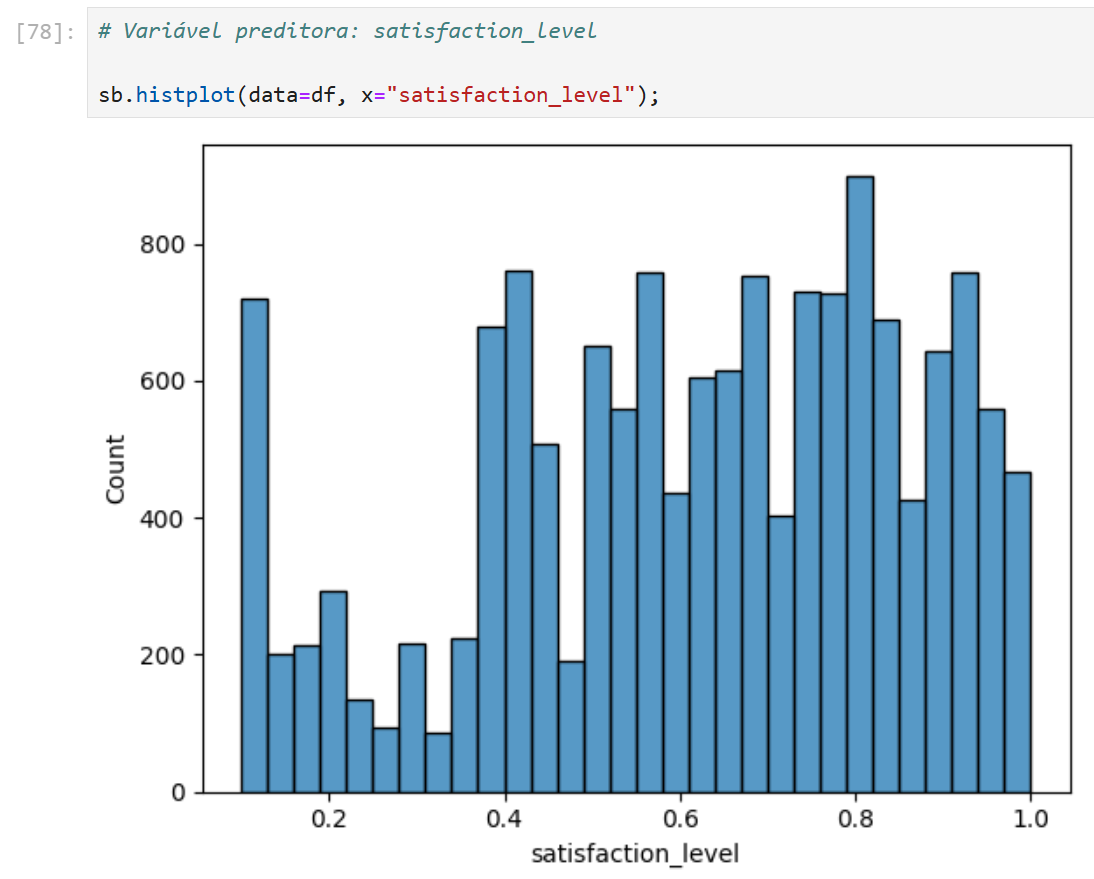
1. **Faça uma análise descritiva de todas as variáveis com a variável ‘target” separada em dois blocos:**



1. **Variáveis preditoras numéricas/quantitativas (sempre colocando uma pequena interpretação dos resultados):**

1 - satisfaction\_level (nível de satisfação - varia de 0 a 1):

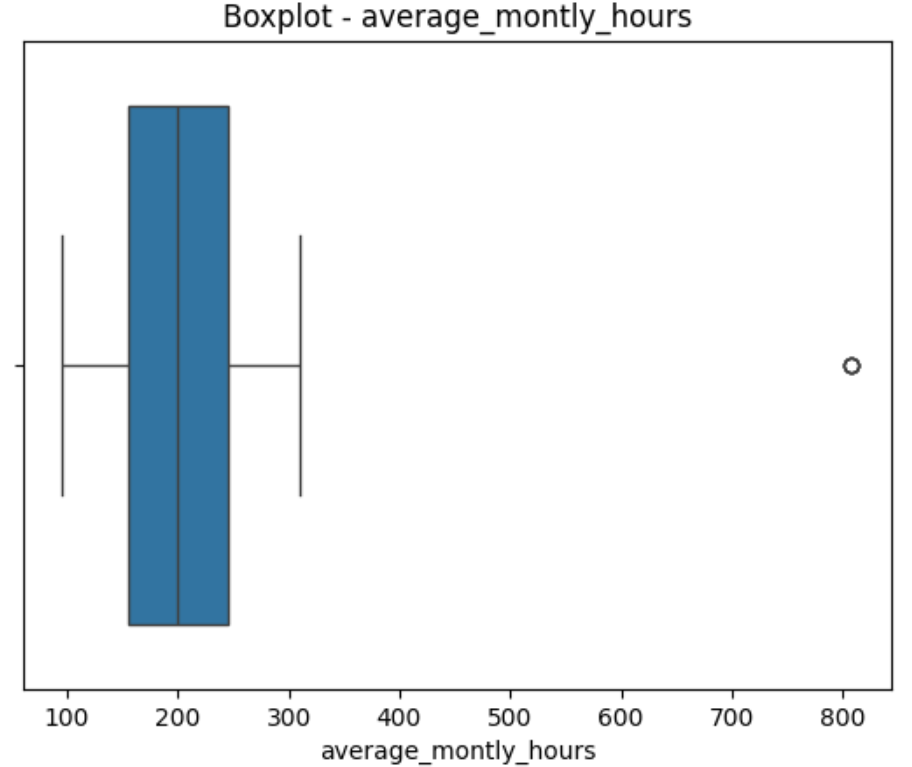
**Ao analisar o nível de satisfação, 75% da base apresenta um nível de satisfação de 82%**

****

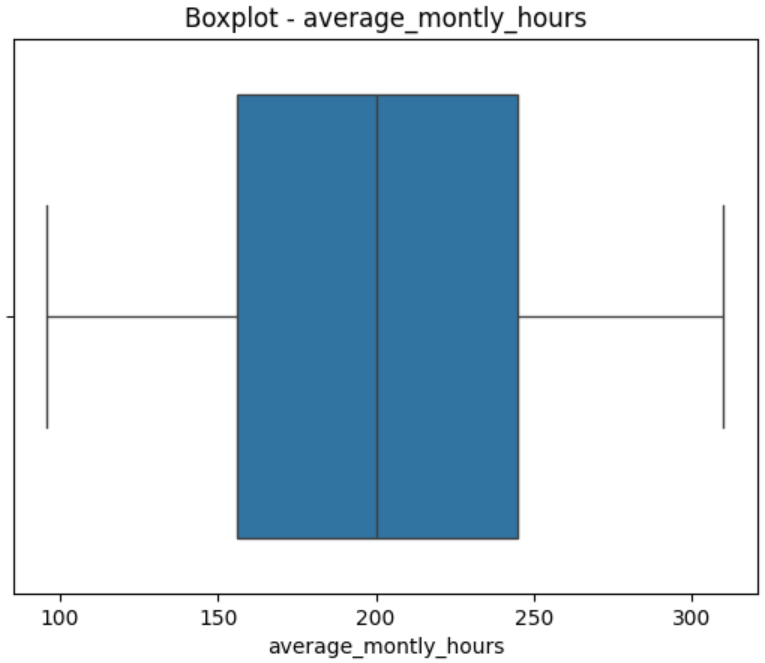
2- last\_evaluation (última avaliação - varia de 0 a 1): Na última avaliação, o valor mediano do resultado de cada funcionário foi de aproximadamente 72%;

3- average\_montly\_hours (média de horas trabalhadas por mês):

**No mês, cada funcionário trabalha em média 201 horas (é possível identificar um ponto extremo que aponta 800 horas trabalhadas no mês)**

****

**O ponto extremo foi removido levando em consideração tudo que estava maior que o último quadrante e o resultado foi o seguinte:**

****

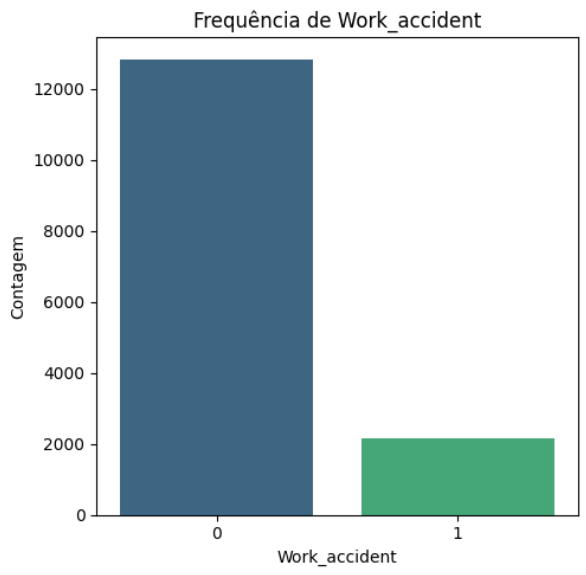
4 - time\_spend\_company (tempo de empresa em anos):

**75% dos funcionários possuem 4 anos de empresa e o tempo máximo de casa no estudo foi de 6 anos**

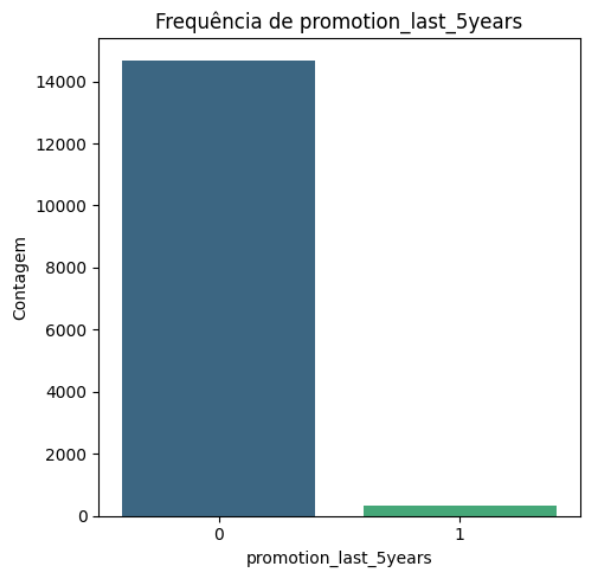
num\_project (número de projetos realizados):

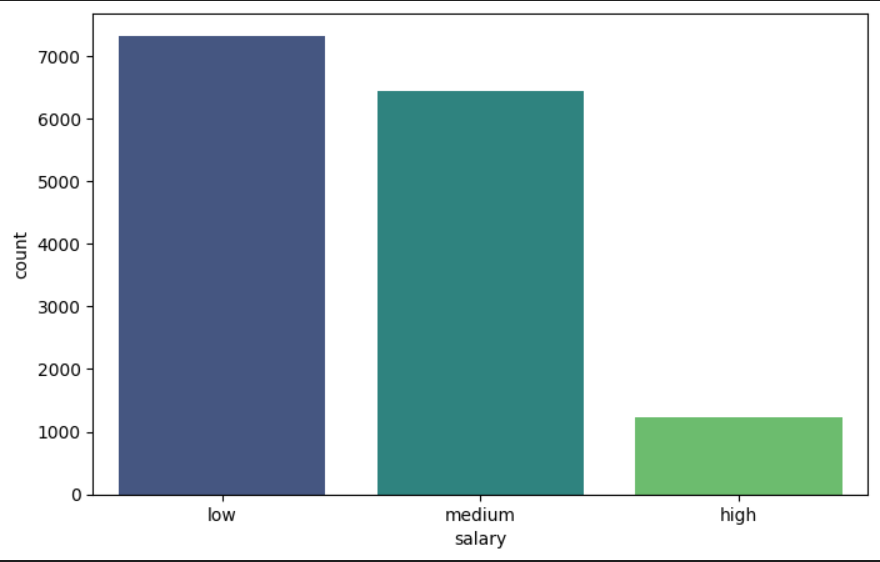
**Na média, cada funcionário participou de 3,78 projetos;**

1. **Variáveis preditoras categóricas/classes (sempre colocando uma pequena interpretação dos resultados)**

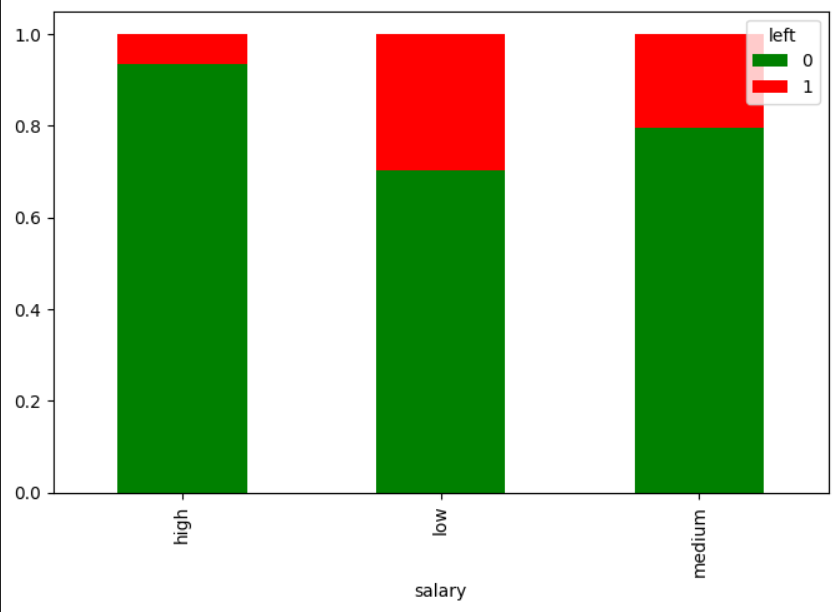
1- Work\_accident: Indica se o funcionário sofreu um acidente de trabalho (0 = não, 1 = sim)  


**Cerca de 16% dos funcionários reportaram algum tipo de acidente no trabalho.**

2- promotion\_last\_5years: Indica se o funcionário foi promovido nos últimos 5 anos (0 = não, 1 = sim)  
  
**327 funcionários tiveram alguma promoção nos últimos 5 anos.**

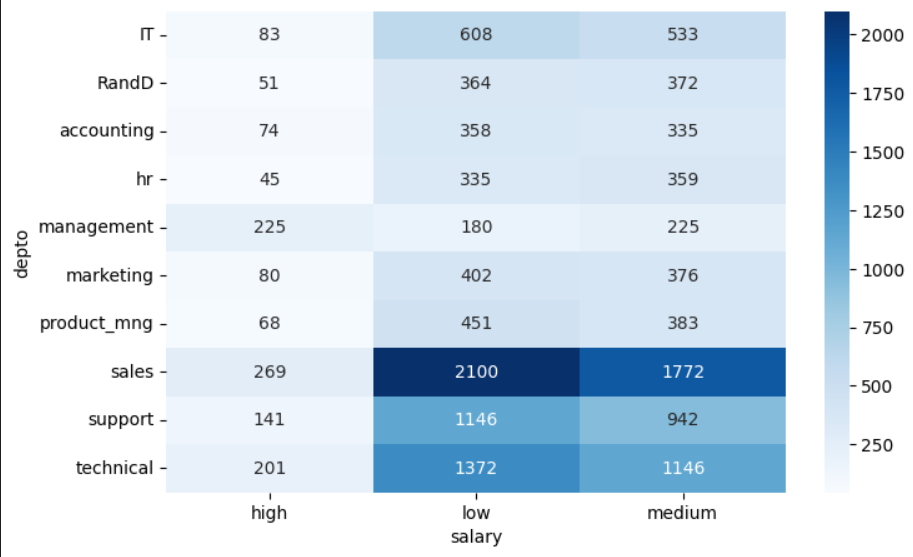
Salary: Faixa salarial (categorias: low, medium, high)  


**A maior parte dos funcionários estão concentrados na categoria de salários baixos.**



**Ao analisar a proporção de saída (turnover) por salário, é possível identificar que funcionários com salários mais baixos tem um número saída maior**

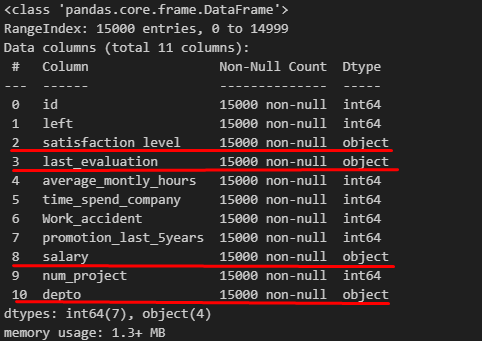
Depto: Departamento do funcionário (categorias como sales, technical, etc.).



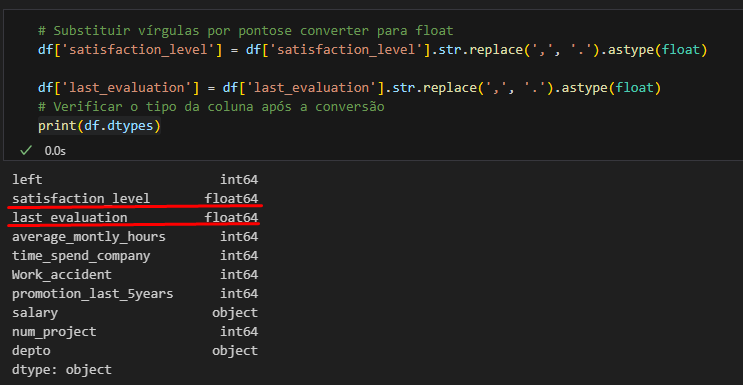
**Existe uma concentração de quantidade de funcionários dos departamentos de Sales, Support e Technical com salários baixos e médios.**

1. **Quais variáveis precisam de alguma transformação? Quais? Coloque tabelas ou gráficos dessas novas variáveis transformadas.**

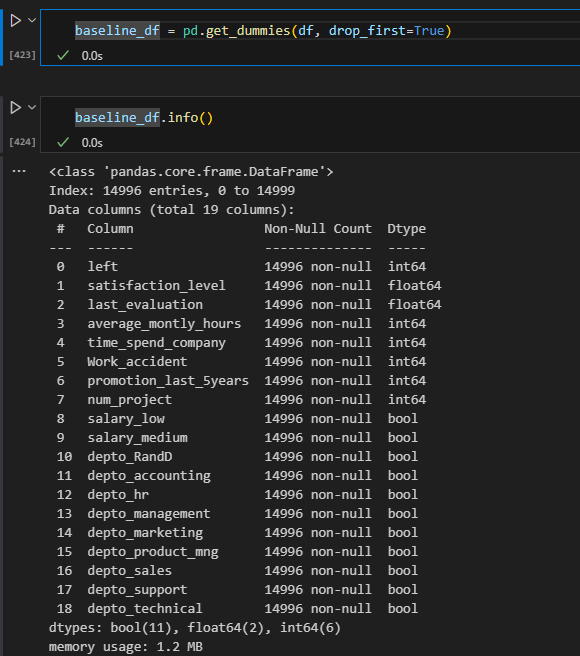
**Tipo de dados do DF sem tratamento:**



* Notamos que quando carregamos o Dataframe as variáveis:
* **satisfaction\_level** e **last\_evaluation** foram carregadas como OBJECT, por conta da “,” então transformamos as duas para FLOAT.



* Outro ponto importante é transformar as variáveis **‘salary’ e ‘depto’** em variáveis DUMMIES já que o modelo de dados não trabalha com valor de dados no formato OBJECT.



**Uma etapa para começar a usar os modelos é muito importante separar em amostras? Explique por que e como separou essas amostras.**

Separamos para os 3 modelos a base de treino e teste 70/30 porém cada um com seus Hiperparâmetros que estão descritos abaixo a fim de evitar:

* Overfitting
* Underfitting
* Avaliarmos na base de teste que utilizamos na base de treino.

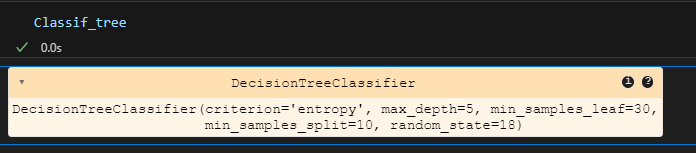
**Utilizando as técnicas Classificação: E para cada técnica**

R: Nós Utilizamos 3 Técnicas:

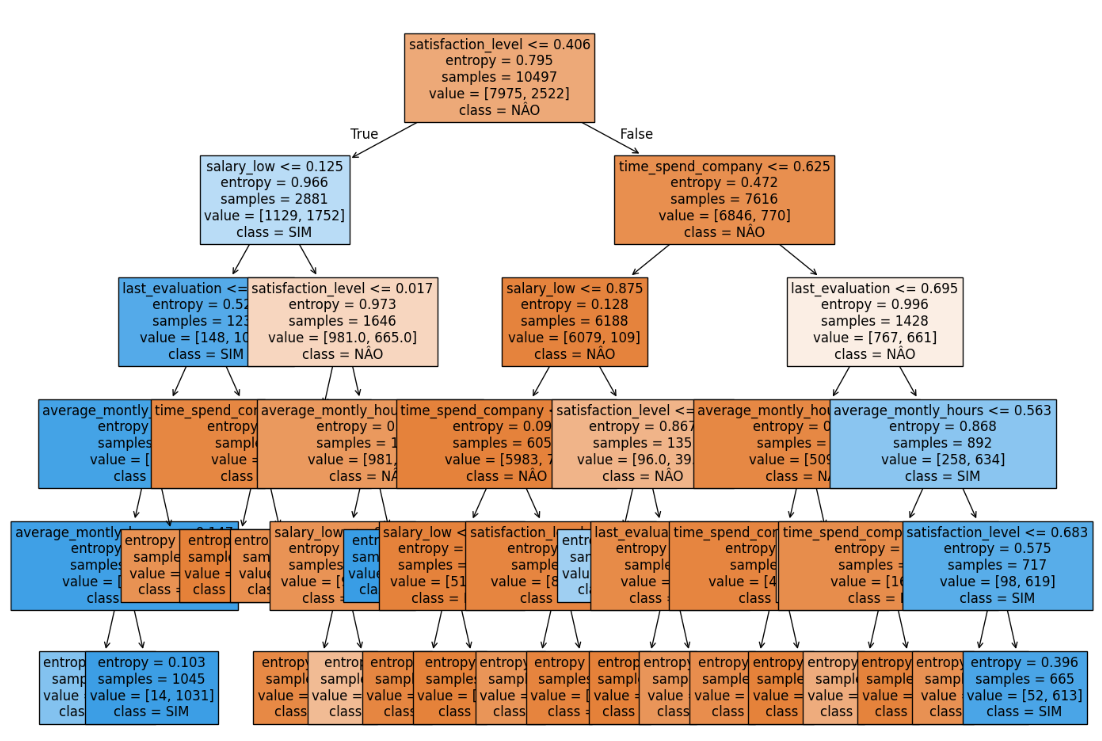
* Árvore de decisão
* Regressão Logística
* KNN K-Nearest Neighbors

1. **Coloque os parâmetros utilizados. Qual a saída do modelo.**

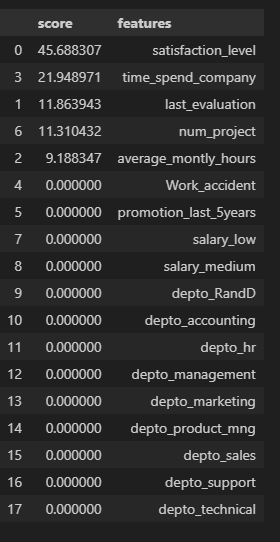
* **Árvore de decisão:** Com os parâmetros fornecidos, ele usa "entropy" para medir a qualidade das divisões, limita a árvore a 5 níveis de profundidade, exige pelo menos 30 amostras nas folhas e 10 para realizar divisões. O parâmetro random state=18 “semente” garante que os resultados sejam reprodutíveis.

****

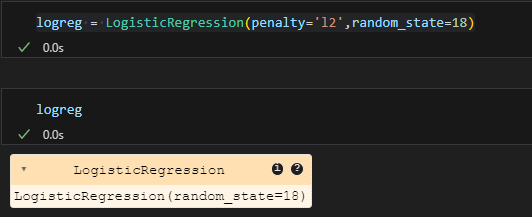
**Resultado da REGRA do modelo de árvore:**

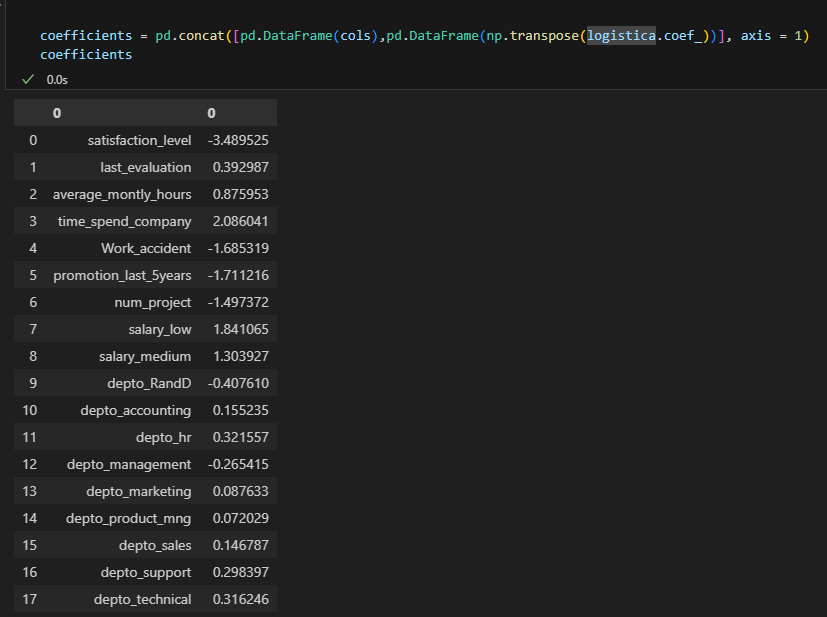
****

**Importância das variáveis no modelo:**

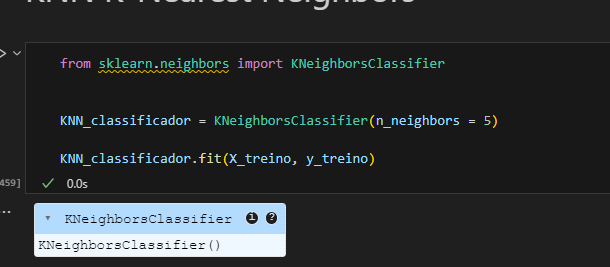
****

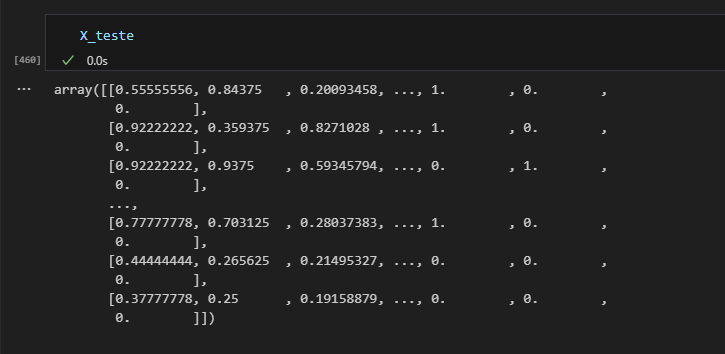
* **LogisticRegression**: Com penalty='l2', ele aplica a regularização de Ridge para evitar o overfitting. O parâmetro random\_state=18 garante resultados reprodutíveis.



**Coeficientes**:   


* KNN K-Nearest Neighbors: n\_neighbors=5, ele classifica uma amostra com base nas 5 amostras mais próximas, utilizando a votação da maioria.

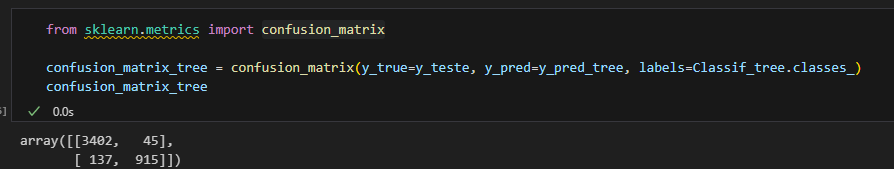


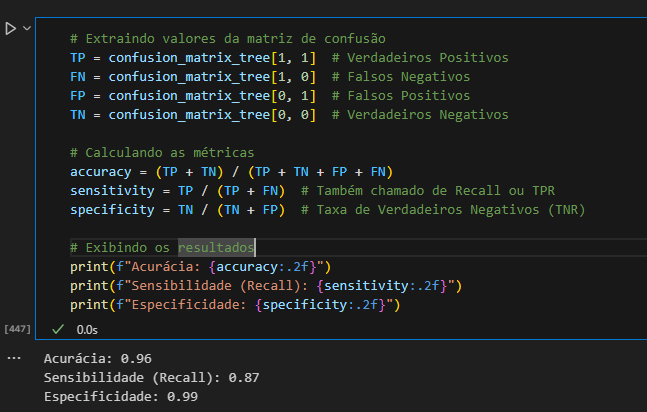


1. Avaliação na amostra teste: Qual é o Acerto do Modelo? Qual a Matriz de Confusão? Mostre o Cálculo da Acurácia? Verifique o Acerto dentro de cada grupo (Sensibilidade e Especificidade)

* **Árvore de Decisão:** Usamos a matriz de confusão para chegar nos dados de Acurácia, Sensibilidade e Especificidade:

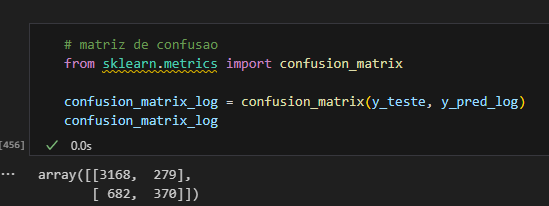
**Matriz de confusão:**

****

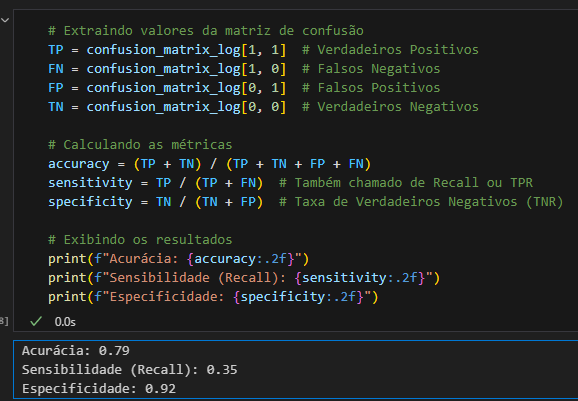
****

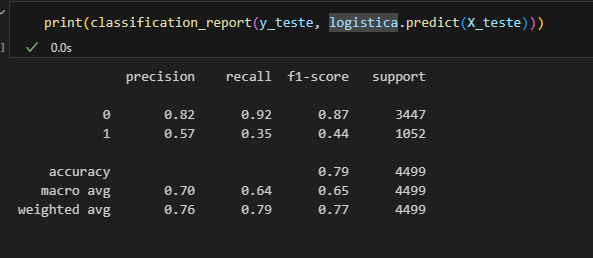
* + **Acurácia: 0.96**
  + **Sensibilidade (Recall): 0.87**
  + **Especificidade: 0.99**
* **Regressão Logística:** Usamos a matriz de confusão para chegar nos dados de Acurácia, Sensibilidade e Especificidade:

**Matriz de confusão:**



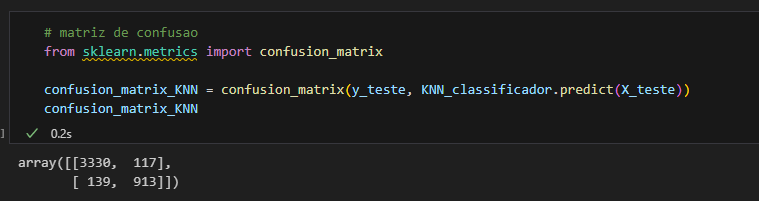
* **Acurácia: 0.79**
* **Sensibilidade (Recall): 0.35**
* **Especificidade: 0.92**



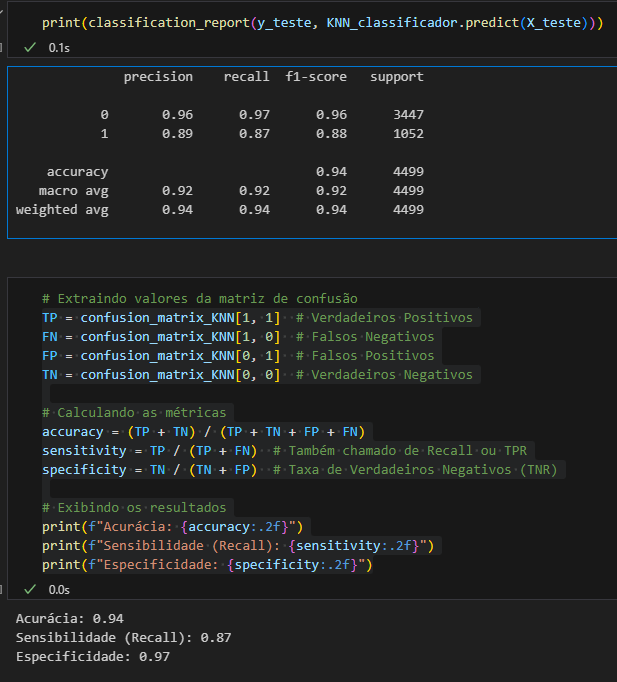
****

* **KNN K-Nearest Neighbors:** Usamos a matriz de confusão para chegar nos dados de Acurácia, Sensibilidade e Especificidade:

**Matriz de Confusão:**

****

* **Acurácia: 0.94**
* **Sensibilidade (Recall): 0.87**
* **Especificidade: 0.97**



### Comparativo Final:

* **Melhor Modelo:** **Árvore de Decisão** : Maior acurácia e equilíbrio entre sensibilidade e especificidade.
* **Modelo com pior desempenho:** **Regressão Logística**: Baixa sensibilidade, inadequada para problemas que exigem detecção de positivos.
* **Modelo competitivo:** **KNN**: Desempenho sólido, quase tão bom quanto a Árvore de Decisão.

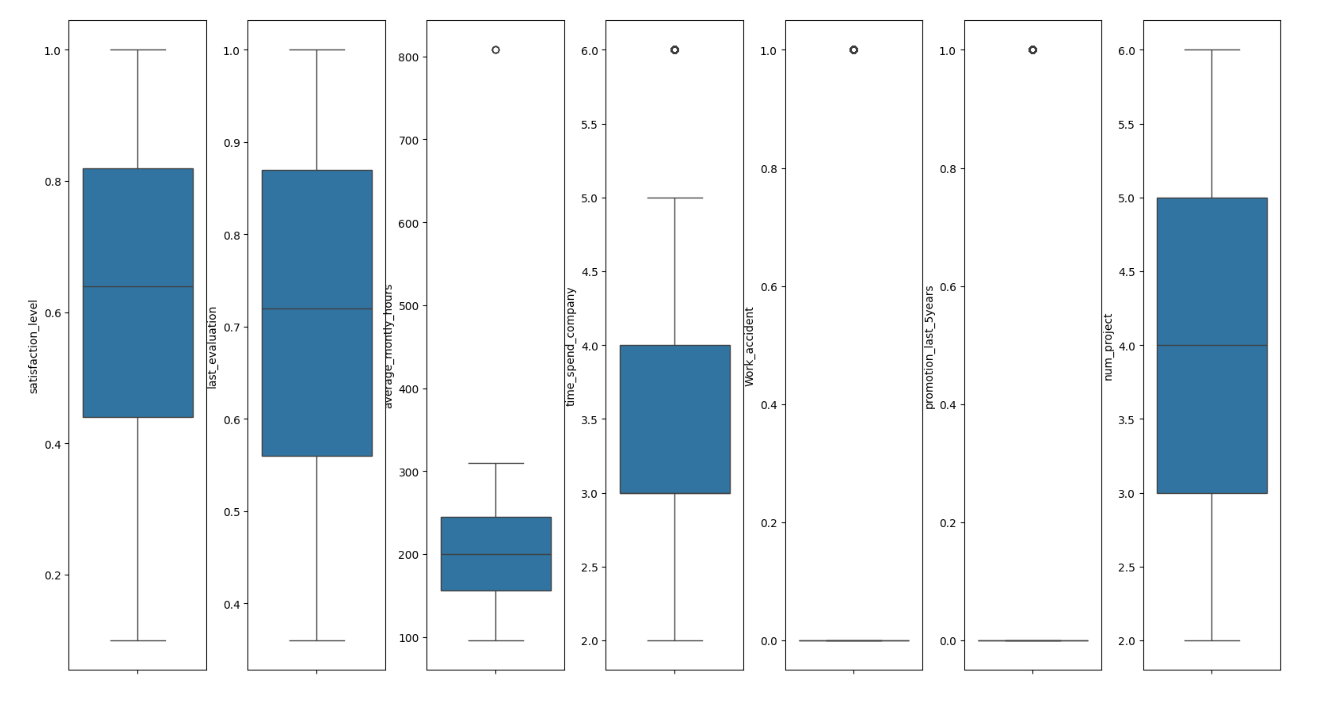
**Utilize a base de dados HR\_Abandono2023 selecionando somente as variáveis numéricas quantitativas crie segmentos de funcionários. Para cada resposta deixar a saída do programa.**

1. **Análise descritiva dessa base com foco no objetivo. Lembre-se de verificar se tem outliers. Indique os critérios de eliminação de outliers, se necessário**.

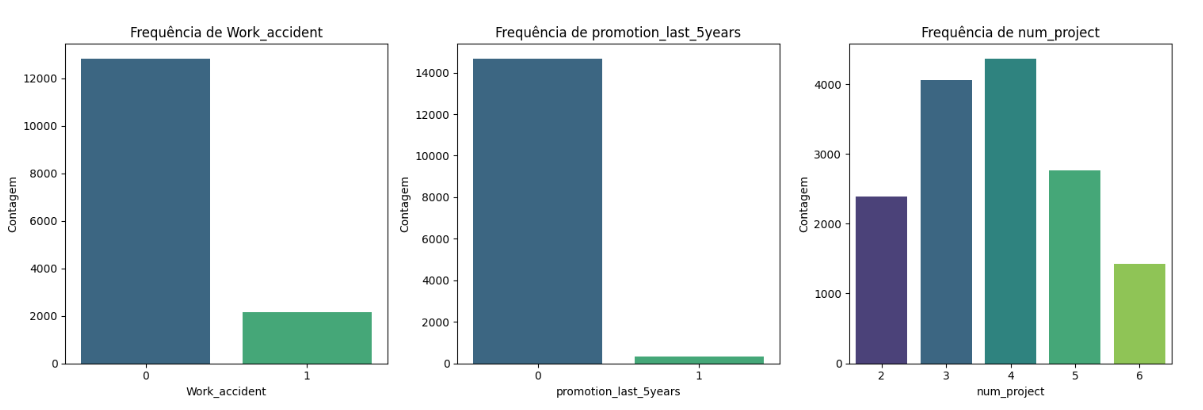
R: Para o modelo de Clusterização utilizamos apenas variáveis Quantitativas que são:

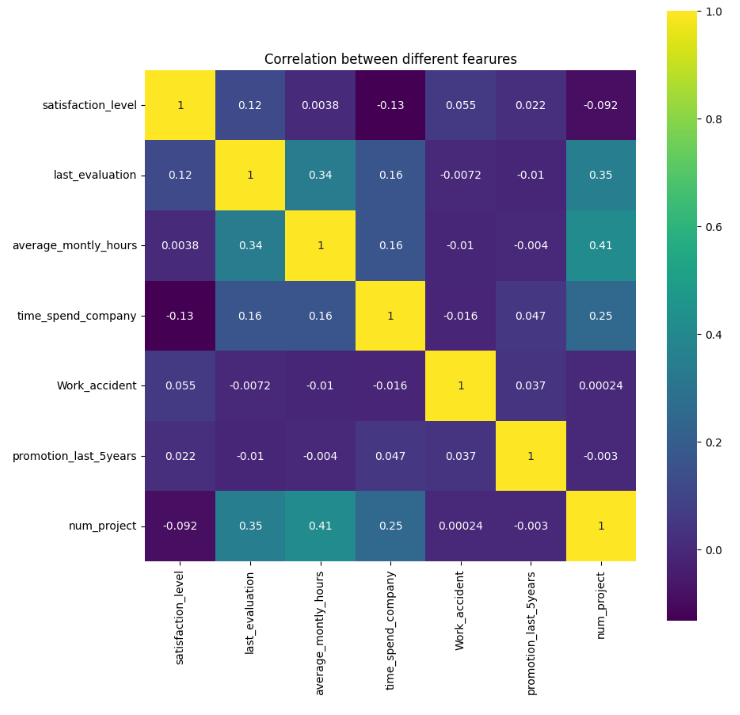
* satisfaction\_level
* last\_evaluation
* average\_montly\_hours
* time\_spend\_company
* Work\_accident
* 'promotion\_last\_5years
* 'num\_project

**Análise Exploratória das variáveis:**



Algumas variáveis a apresentação não ficou interessante no boxplot então utilizamos gráfico de Barras:



**Matriz de Correlação:**

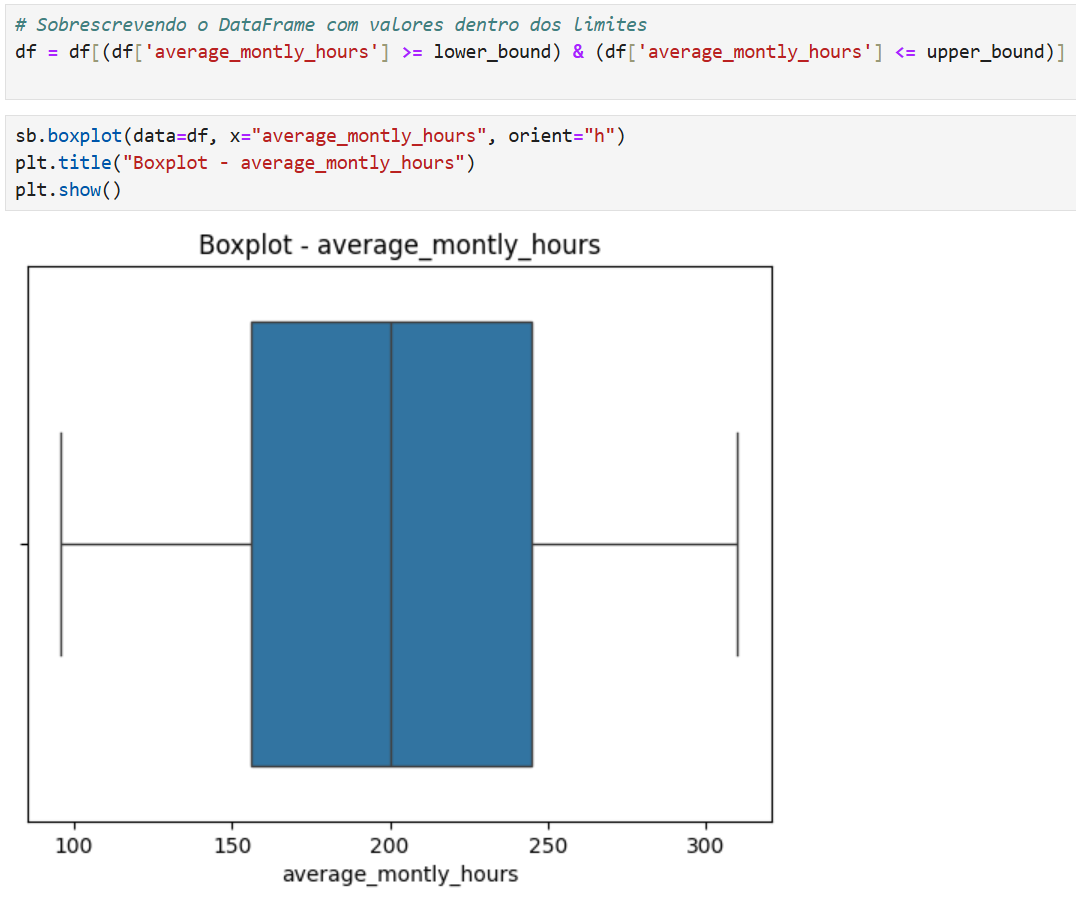
1. **Correlação forte**:  
   "Número de projetos" e "última avaliação" (0.350.35).
   * "Número de projetos" e "média de horas trabalhadas" (0.410.41), o que faz sentido, pois mais projetos exigem mais tempo.
2. **Correlação moderada**:  
   * "Tempo de casa" e "número de projetos" (0.250.25), indicando que funcionários mais experientes tendem a lidar com mais projetos.
   * "Última avaliação" e "média de horas trabalhadas" (0.340.34), sugerindo que mais esforço é recompensado com boas avaliações.
3. **Correlação negativa**:
   * "Nível de satisfação" e "número de projetos" (−0.092-0.092), possivelmente indicando que muitos projetos podem reduzir a satisfação.
4. **Pouca influência**:
   * "Acidentes de trabalho" e "promoções nos últimos 5 anos" têm correlações baixas com outras variáveis, sugerindo menor impacto direto no comportamento analisado.

Conforme citado anteriormente, foram encontrados alguns outliers no número de horas mensais.

Sendo assim aplicamos os seguintes critérios para eliminá-los:

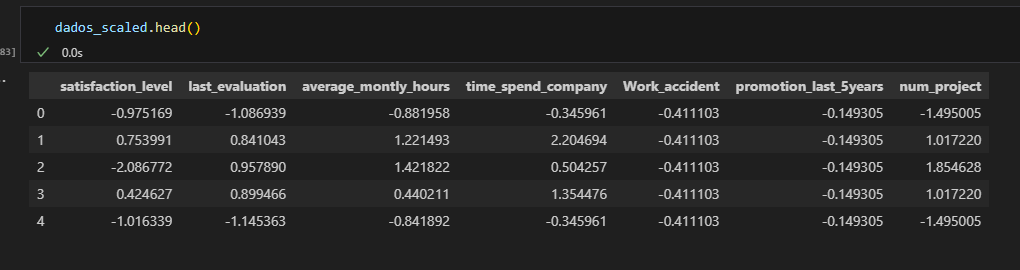


Agora sim, com om outliers removidos, chegamos no seguinte resultado:



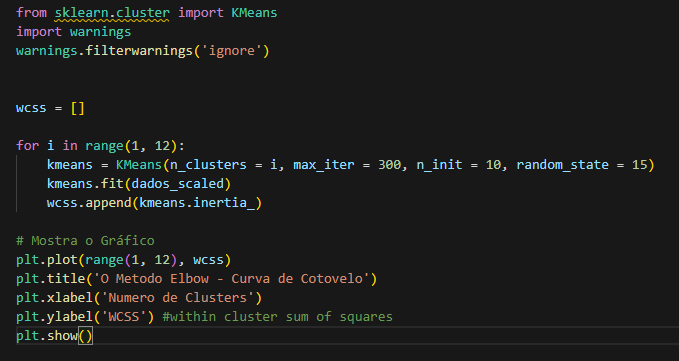
1. **Porque é necessário fazer a padronização das variáveis segmentadoras. Coloque a sumarização das variáveis padronizadas.**

R: Fizemos a padronização das variáveis segmentadoras para garantir que todas tenham a mesma escala, com média zero e desvio padrão igual a um. Isso evita que variáveis com valores maiores influenciem mais o modelo, permitindo que ele performe melhor e trate todas as variáveis de forma equilibrada



1. **Qual a quantidade de clusters indicado pelo método de Elbow e com o Método Hierárquico**

Parâmetros utilizados:

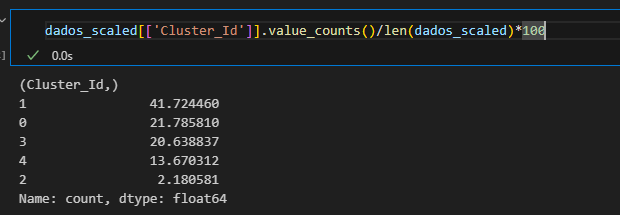


R: Notamos que a curva do cotovelo varia entre 4 e 6, porém escolhemos 5 Clusters.

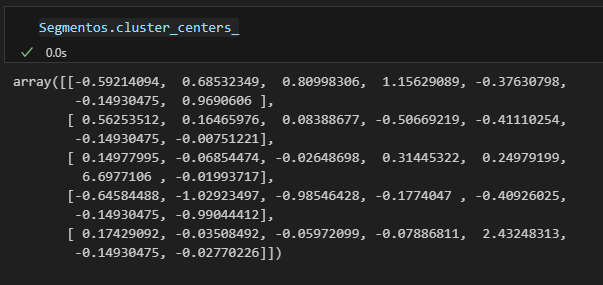


1. **Utilizando o Método K-Means, qual a quantidade e percentual de funcionários em cada cluster.**

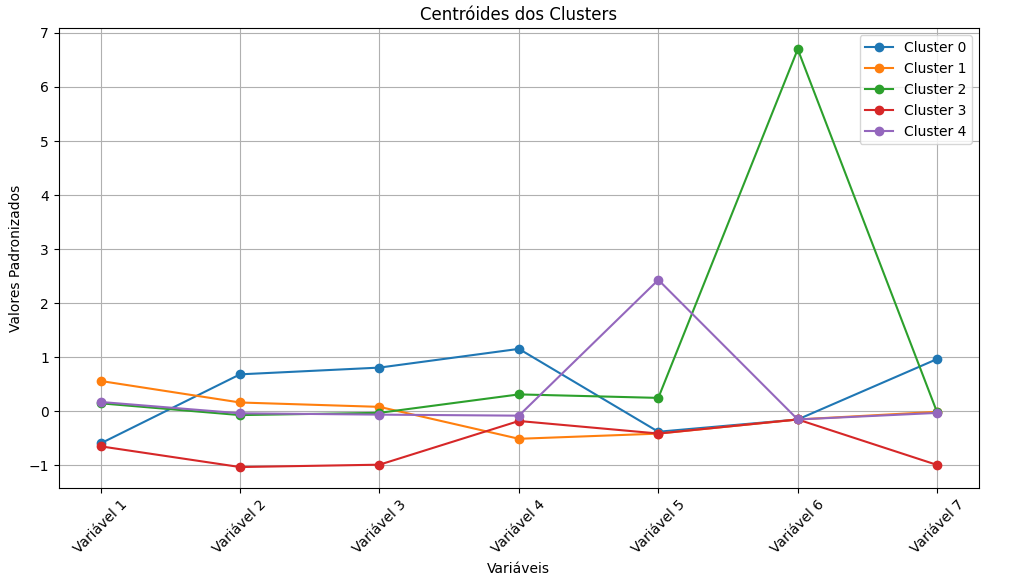
* Cluster 1: 41.724460 (6257)
* Cluster 0: 21.785810 (3267)
* Cluster 3: 20.638837 (3095)
* Cluster 4: 13.670312 (2050)
* Cluster 2: 2.180581 (327)



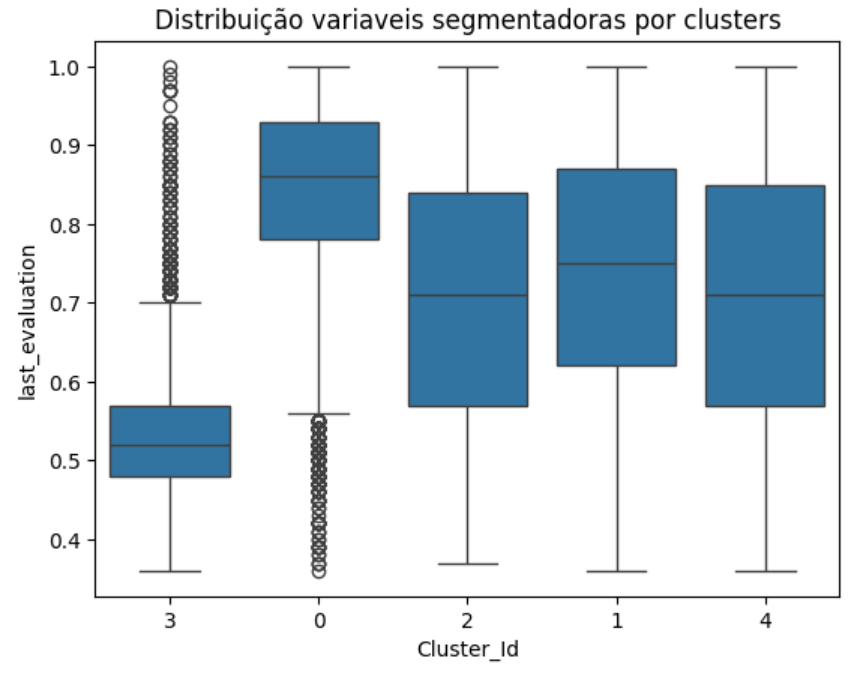
1. Qual o centróide dos clusters.



Centróides dos Clusters.



1. **Como vocês interpretam cada cluster. Descrever as características de cada cluster**

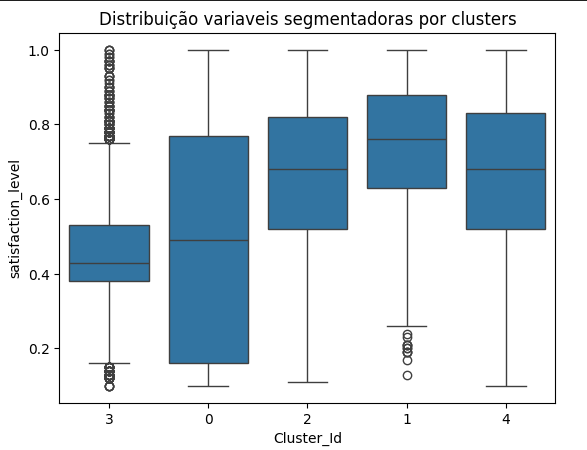


**Last Evaluation:**

Ao analisar a última avaliação dos funcionários, é possível separar os cluster por grupos e identificar um perfil.

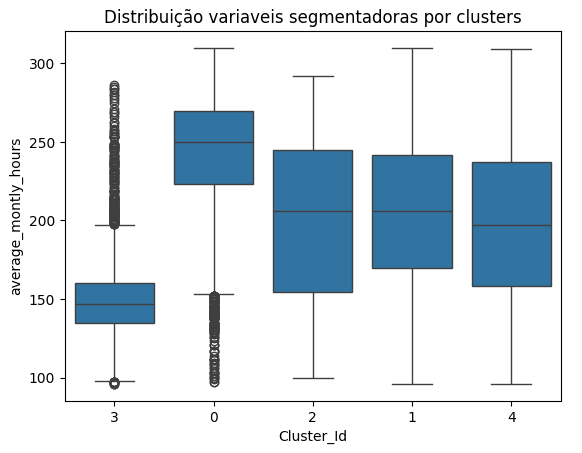
* **Cluster 3:** Apresenta uma mediana relativamente baixa se comparado aos outros clusters, indicando assim um número de desempenho bem abaixo.
* **Cluster 0:** Os funcionários agrupados neste cluster demonstram um alto desempenho na última avaliação. Possivelmente são pessoas que mantiveram um alto nível nas entregas e por consequência tiveram um reconhecimento/avaliação acima da média.
* **Cluster 1,2 e 4:** O conjunto desses clusters apresentou medianas bem próximas, o que leva a crer que os funcionários desse grupo apresentam um desempenho moderado

**Satisfaction Level**

****

* **Cluster 3**: Apresenta os menores níveis de satisfação geral.Muitos outliers estão presentes no topo, indicando uma pequena quantidade de funcionários com satisfação acima da média."Funcionários insatisfeitos".
* **Cluster 0**: Mostra uma grande variação nos níveis de satisfação, abrangendo desde os mais baixos até níveis médios e altos. "Grupo heterogêneo".
* **Cluster 2 e 4**: Apresentam níveis de satisfação predominantemente altos, com mediana próxima de 0.8."Funcionários satisfeitos".
* **Cluster 1**: Tem uma mediana de satisfação moderada, com menor variação e poucos outliers no nível baixo. "Funcionários moderadamente satisfeitos".

**Average Montly Hours**

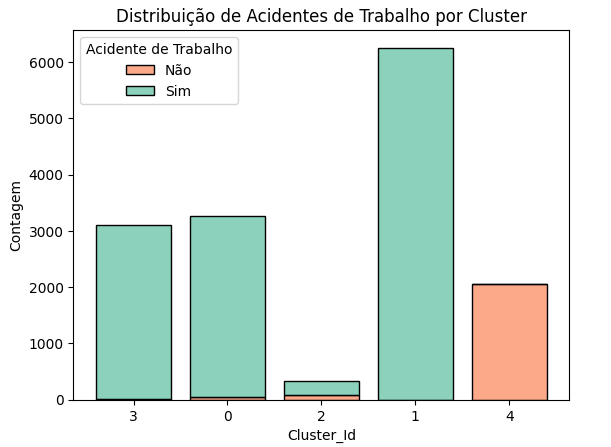


* **Cluster 3**: Apresenta a quantidade média de horas trabalhadas.Muitos outliers estão presentes no topo, indicando uma pequena quantidade de funcionários quantidade de horas acima da média."Funcionários improdutivos".
* **Cluster 0**: Mostra uma pequena quantidade em comparação a outros grupos média de horas trabalhadas acima de 250 horas, abrangendo desde os mais baixos até níveis médios e altos, apesar de ter uma boa quantidade de outlier com a média baixa. "Funcionários super produtivos".
* **Cluster 1, 2 e 4**: Quantidade de horas trabalhadas na média esse 3 grupos contêm a maior quantidade de Funcionários com mediana próxima de 200 horas."Funcionários Medianos".

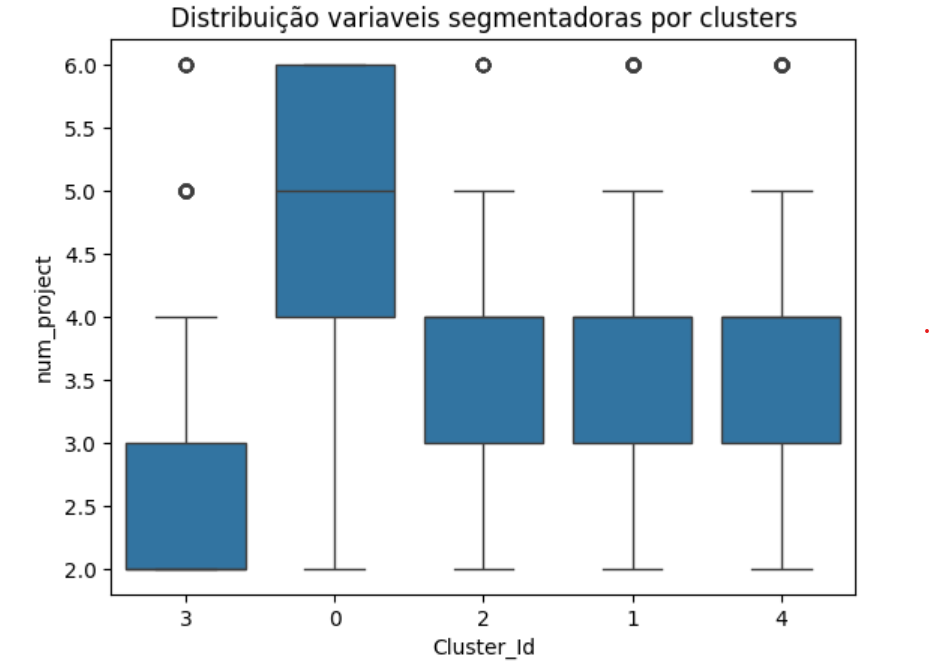


* **Cluster 3**: Este grupo possui os funcionários com o menor tempo de empresa, com valores concentrados entre 2 e 3 anos. São provavelmente os “novatos” na organização, ainda em fase inicial de adaptação.
* **Cluster 0**: Apresenta uma distribuição mais concentrada entre 4 e 6 anos de tempo de empresa, com outliers que podem ultrapassar essa faixa. Esse cluster parece agrupar colaboradores com experiência intermediária.
* **Cluster 2**: Tem uma amplitude maior no tempo de empresa, variando de 2 a 6 anos, com alguns outliers abaixo de 2 anos. Indica uma possível mistura de novos e experientes no mesmo grupo.
* **Cluster 1**: É caracterizado por uma distribuição mais centralizada, com valores mediana abaixo de 4 anos. Esse grupo apresenta uma faixa mais homogênea de tempo na empresa.
* **Cluster 4**: Similar ao cluster 0, mas com uma maior concentração na mediana, em torno de 4,5 anos. Os colaboradores desse grupo provavelmente têm uma experiência consolidada, mas sem outliers extremos.

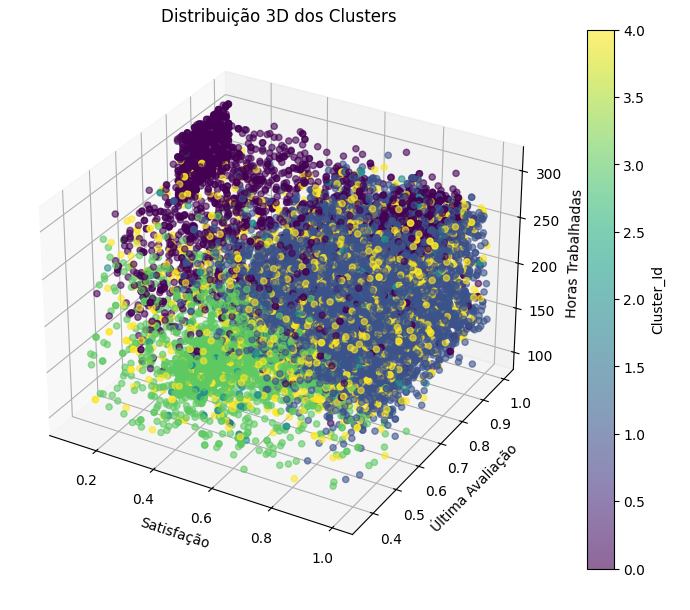
**Work Accident**



* **Cluster 1:** São funcionários que não sofrem acidentes. “Totalmente Seguros”.
* **Cluster 0 e 3**: Grupo de funcionários que quase nunca sofrem. Representa um número considerável de funcionários, porém a quantidade de acidentes é bem pequena. “Quase seguros”.
* **Cluster 2:** São funcionários que eventualmente sofrem acidentes de trabalho, porque o total de acidentes em comparação a quantidade de funcionários é relevante.“Possivelmente Seguros.
* **Cluster 4:** Concentra o grupo de funcionários que sempre sofrem acidente de trabalho. “Totalmente inseguros.

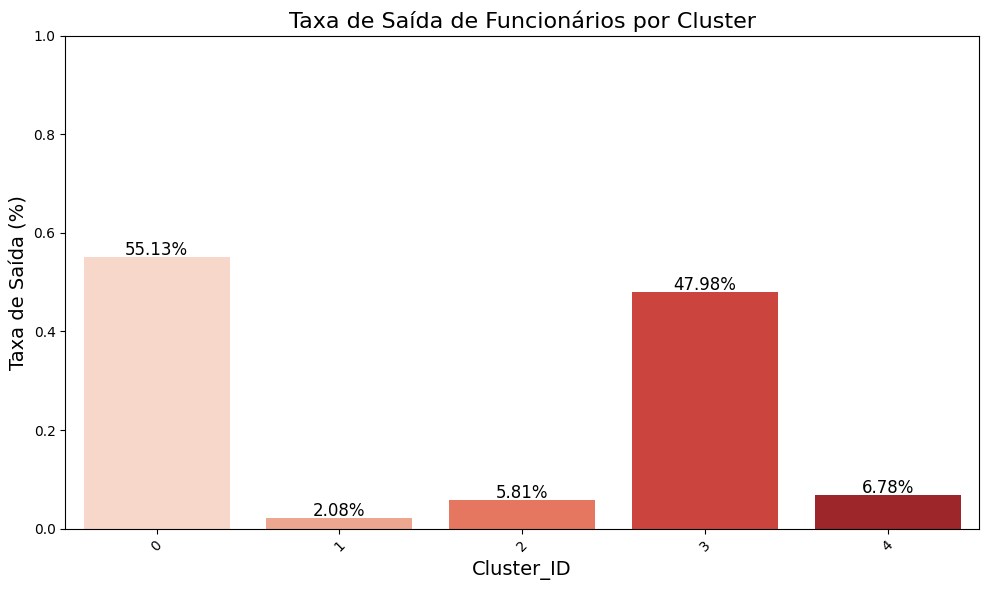


* **Cluster 3:** São funcionários que não possuem uma boa quantidade de projetos, sua Mediana de projeto está em 2. “Funcionários quase ociosos”
* **Cluster 0:** São funcionários que estão com alta quantidade de projetos, possuem Mediana com 5 projetos. “Funcionários Sobrecarregados”
* **Clusters 2,1,4:** São clusters que apresentam uma distribuição estável entre eles, o que leva a crer que existe um equilíbrio no número de projetos para funcionários desse grupo.



1. **A partir dos clusters criados volte na base inicial de funcionários e traga a informação se ele está ativo ou está desligado. Faça um cruzamento do cluster com a penetração de funcionários que já saíram da empresa (left=1). Quais análises ou considerações vocês podem tirar daqui.**

Fizemos uma análise para entender a Taxa de saída de acordo em relação aos Clusters:



* **Cluster 0 e 3**: Alta taxa de saída. Neste grupo de clusters, é possível identificar que existe um grande número de funcionários deixando a empresa
* **Cluster 2 e 4:** Saída moderada. Neste grupo, podemos observar um número de saída moderado, levando a crer que são funcionários que estão em áreas mais estáveis
* **Cluster 1:** Para esse cluster, notamos poucas saídas, o que garante a ele o título de grupo mais estável. É bem provável que seja composto por funcionário de áreas mais estratégicas com baixíssima rotatividade.

LINK para o Notebook:

<<https://github.com/Uill-Iam/Estudo/blob/main/Python/ML/ACEM/ACEM.ipynb>>