

ANÁLISE DE RECURSOS HUMANOS

João Victor Soares Saraiva

2022



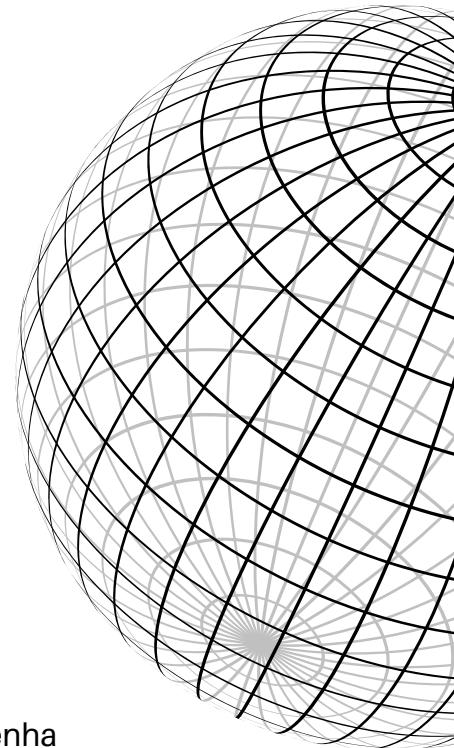
INTRODUÇÃO

Objetivos do Desenvolvimento

Neste projeto cobri todas as etapas de um projeto real de Data Science, pude resolver o problema de como utilizar dados para responder a questões importantes para permitir que uma empresa tenha conhecimento sobre:

- Quais são os fatores que influenciam para um colaborador deixar a empresa?
- Como reter pessoas?
- Como antecipar e saber se um determinado colaborador vai sair da empresa?

E por fim disponibilizar recursos para que a empresa consiga realizar a predição para verificar se um colaborador vai ou não deixar a empresa com base em atributos como comportamento e carga de trabalho, nível de satisfação com a empresa e resultados de performance.



SOLUÇÃO PROPOSTA



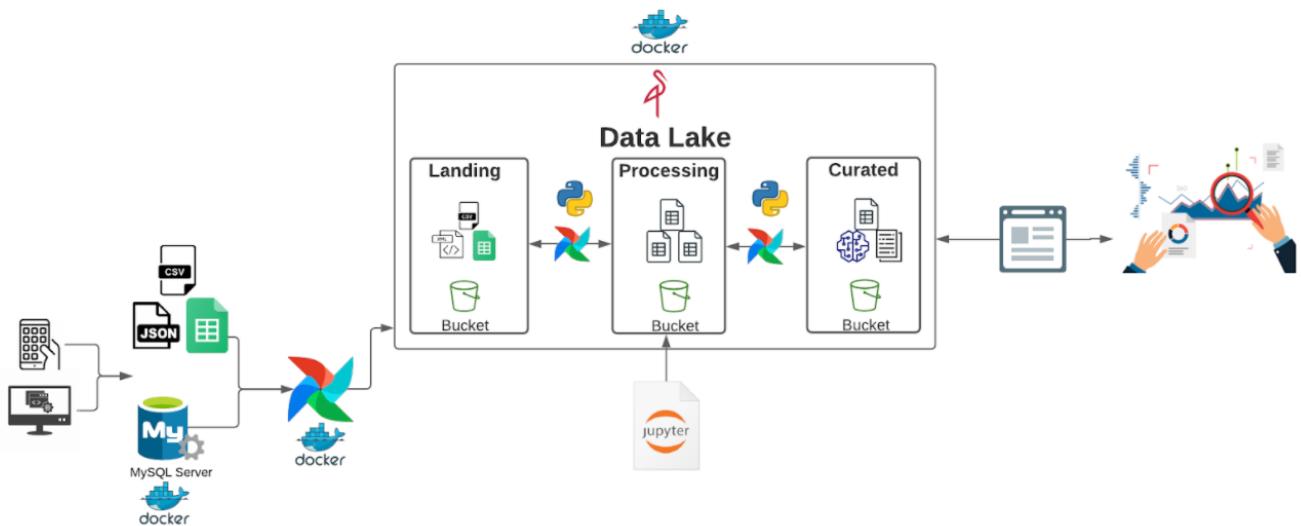
Para resolver esse problema foi construído uma solução completa para armazenamento, gestão e automatização de fluxos de dados, utilizando tecnologias como:

- **Docker**: Foi utilizado para virtualizar as tecnologias utilizadas.
- **Apache Airflow**: Foi utilizado para gerenciar o fluxo de dados.
- **MySQL**: Foi o banco de dados e uma das fontes de dados utilizadas no projeto.
- **Minio**: É a tecnologia utilizada como Data Lake.

Além das tecnologias, foi explorada uma suíte poderosa de tecnologias e bibliotecas para trabalhar com Análise de Dados e Machine Learning que são:

- **Pandas**: Para manipulação de dados;
- **Seaborn e Matplotlib**: Para visualização de dados;
- **Scikit-learn**: Para implementar algoritmos de aprendizado de máquina;
- **Pycaret**: Para automatizar fluxo de criação e avaliação de algoritmos de Machine Learning;
- **SweetViz**: Para gerar relatório automático da base de dados;
- **Streamlit**: Para criação de uma interface gráfica interativa para o usuário final;
- **GitHub**: Para armazenar todos os códigos desenvolvidos.

SOLUÇÃO PROPOSTA



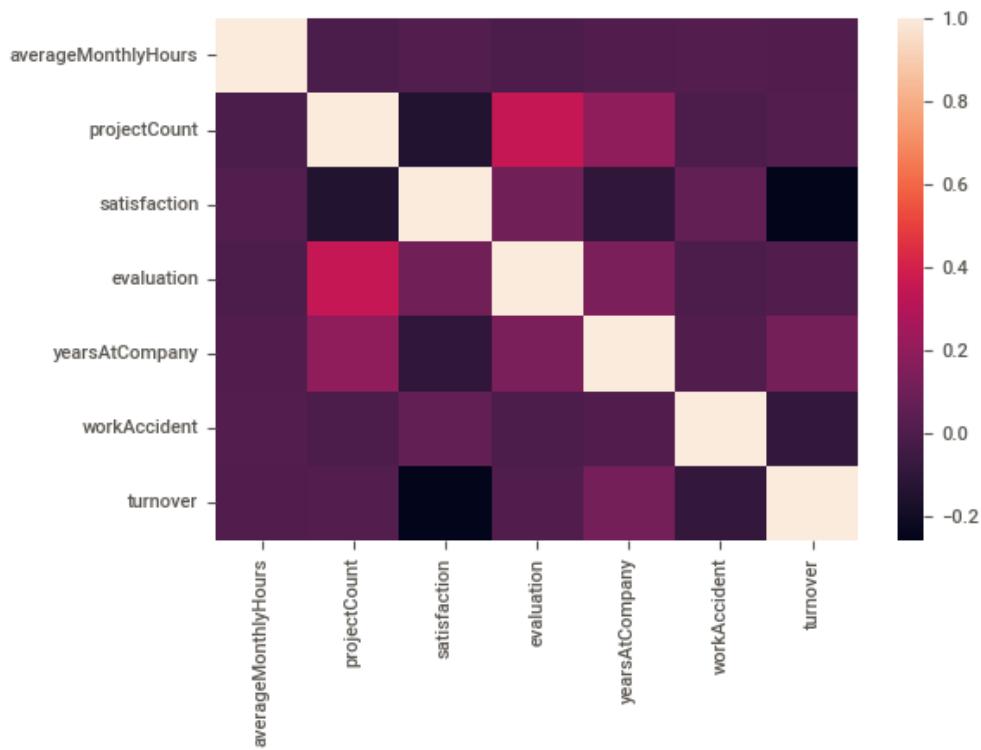
A imagem acima representa todo o fluxo de dados do projeto, no início do fluxo temos as fontes de dados, logo em seguida temos o processo de ingestão dos dados para a zona *Landing*, no Data Lake. Após os dados estiverem no Data Lake, é iniciado o processo de transformação e carga dos dados nas zonas *Processing* e *Curated*, respectivamente. Logo após os processos de ingestão, tratamento e carga dos dados, desenvolvi um Data App para consumir o modelo de Machine Learning.

Depois da infraestrutura devidamente criada e configurada, e levando em consideração o desafio proposto, foram criados e modelados os atributos relevantes para análise utilizando fontes de dados diversas como arquivos em formato xlsx, json e dados no Sistemas de Gerenciamento de Banco de Dados MySQL

RESULTADOS

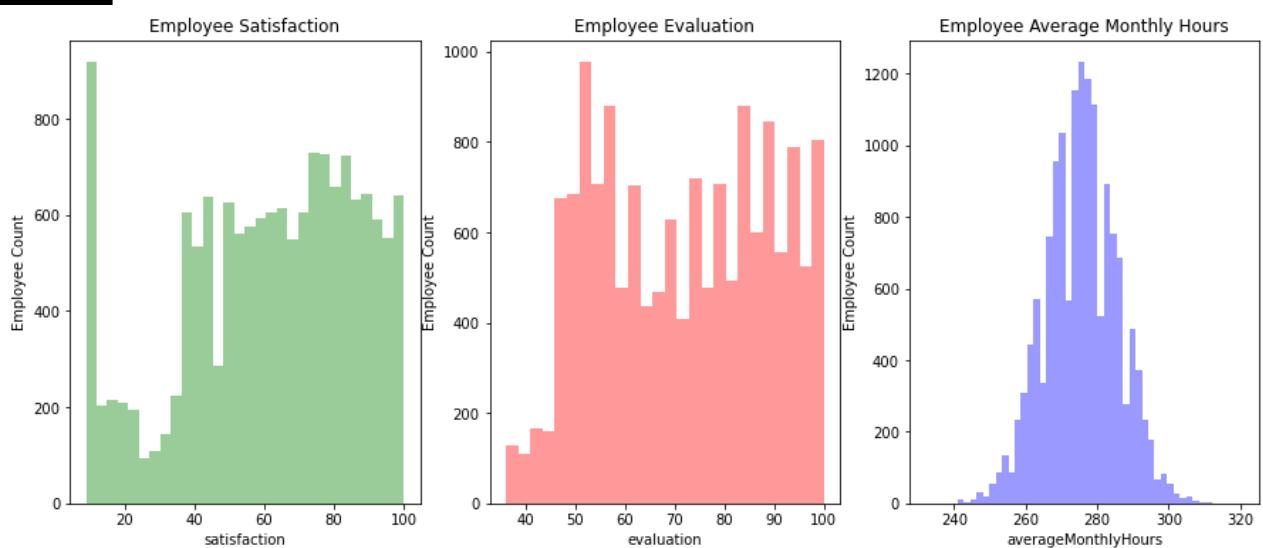
NA ETAPA DE ANÁLISE EXPLORATÓRIA DE DADOS FORAM DESCOBERTOS VÁRIOS INSIGHTS IMPORTANTES ABAIXO:

- A empresa tem uma rotatividade de 24%.
- A satisfação média dos empregados é de 61.
- A satisfação média dos empregados que deixaram a empresa é 49
- Podemos assumir que os empregados que mais deixam a empresa estão menos satisfeitos



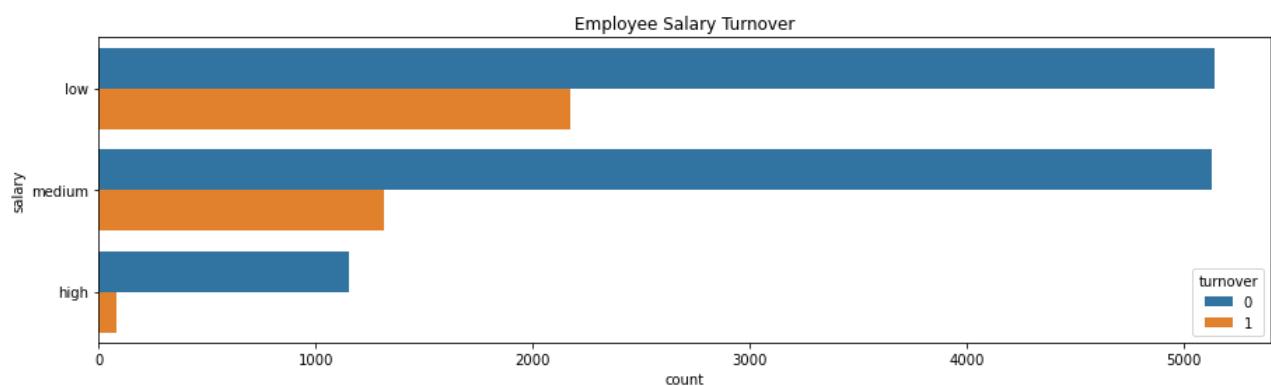
- Existe uma correlação positiva entre os atributos projectCount e Evaluation.
- Faz sentido que empregados que estão envolvidos em mais projetos, trabalham mais e tem melhor avaliação.
- Existe uma correlação negativa entre os atributos satisfaction e turnover.
- Podemos assumir que empregados que mais deixam a empresa estão menos satisfeitos.

RESULTADOS



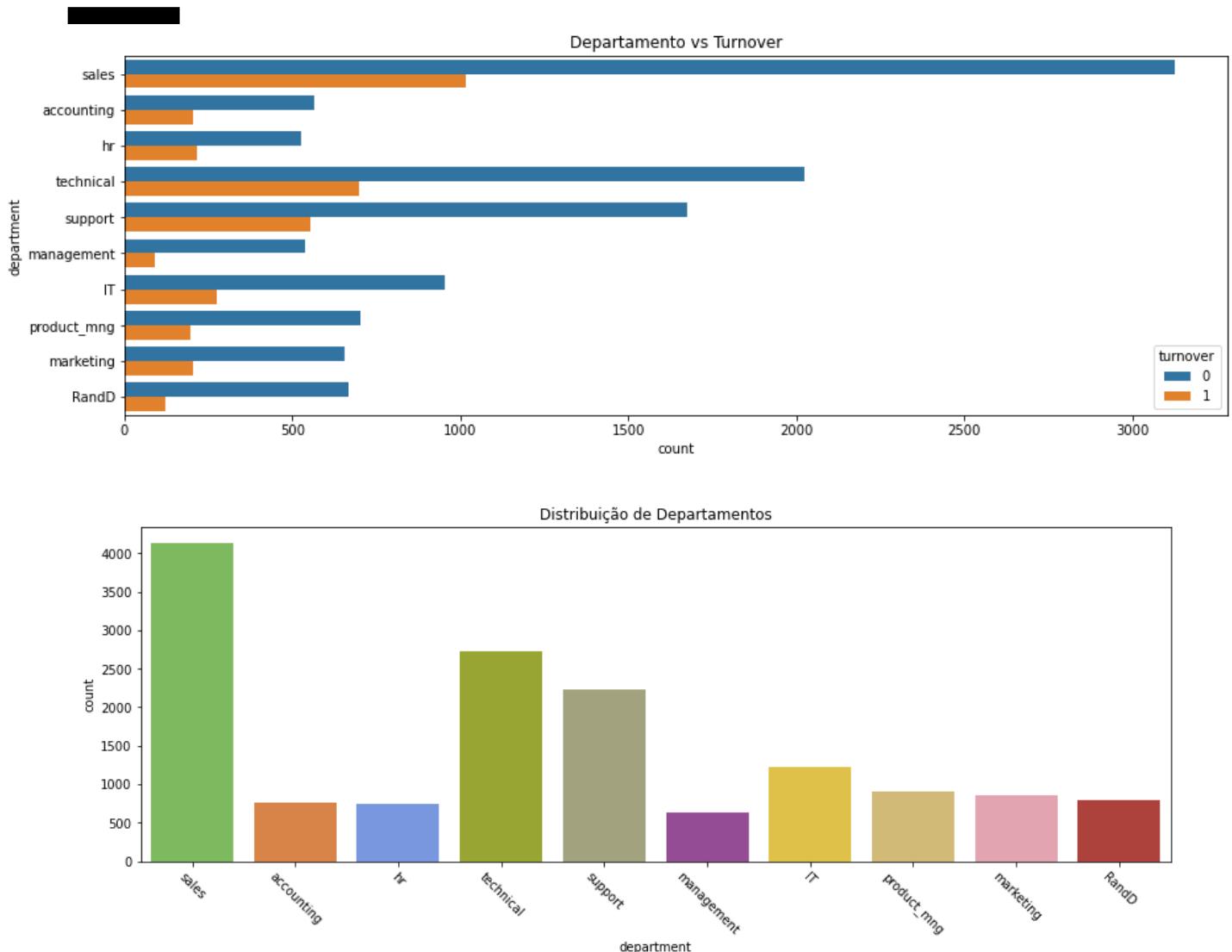
Examinando a distribuição de alguns atributos do conjunto de dados:

- **Satisfaction** - Existe um pico de empregados com baixa satisfação, mas a maior concentração está em 60 a 80.
- **Evaluation** - Temos uma distribuição bimodal de empregados com avaliações baixas, menor que de 60, e altas maior que 80.
- **AverageMonthlyHours** - A concentração da quantidade de horas trabalhadas nos últimos 3 meses está ao redor da média em 275 horas.



- A maioria dos empregados que saíram tinha salário baixo ou médio.
- Quase nenhum empregado com alto salário deixou a empresa.

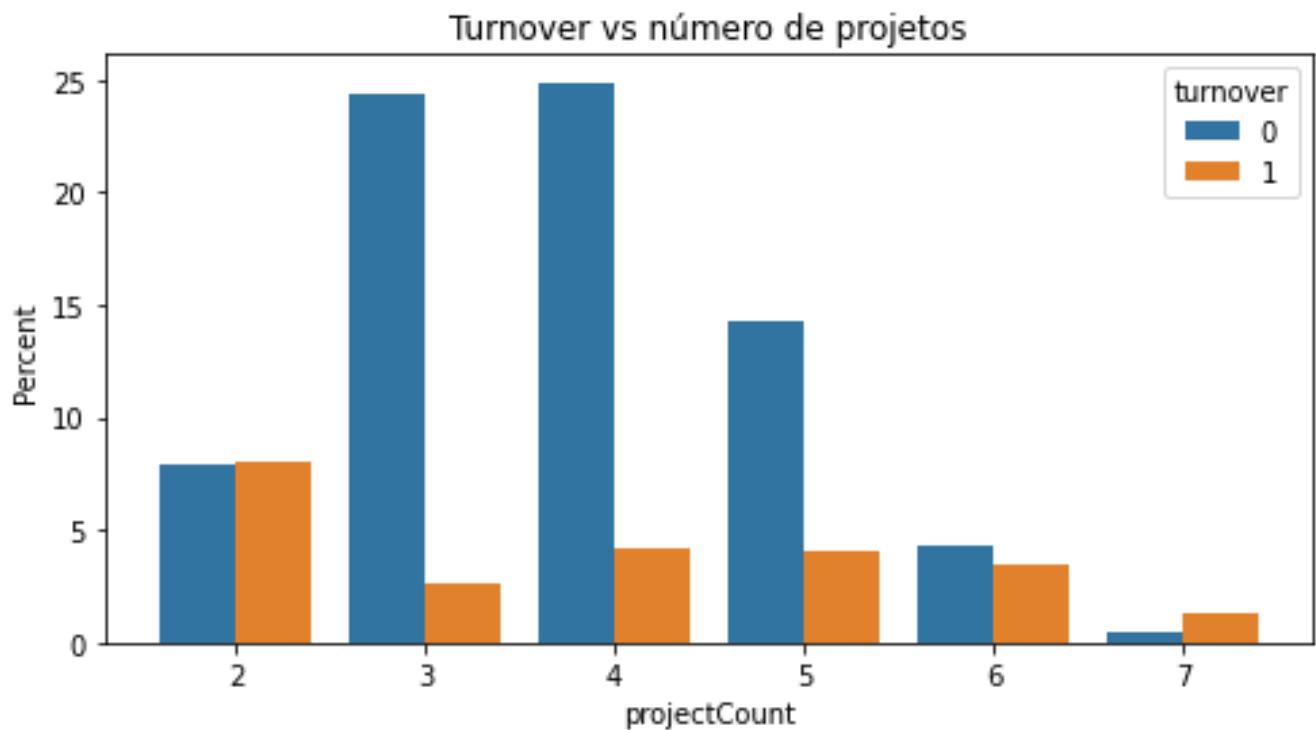
RESULTADOS



Informações sobre os departamentos da empresa.

- Os departamentos de vendas, técnico e suporte são top 3 departamentos com maior índice de turnover.
- O departamento management tem o menor volume de turnover.

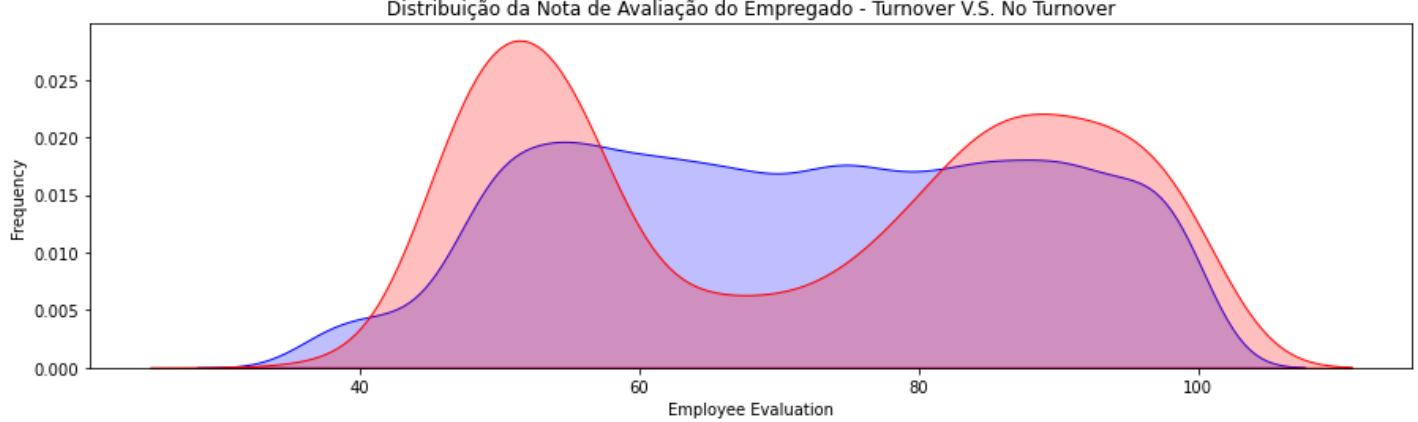
RESULTADOS



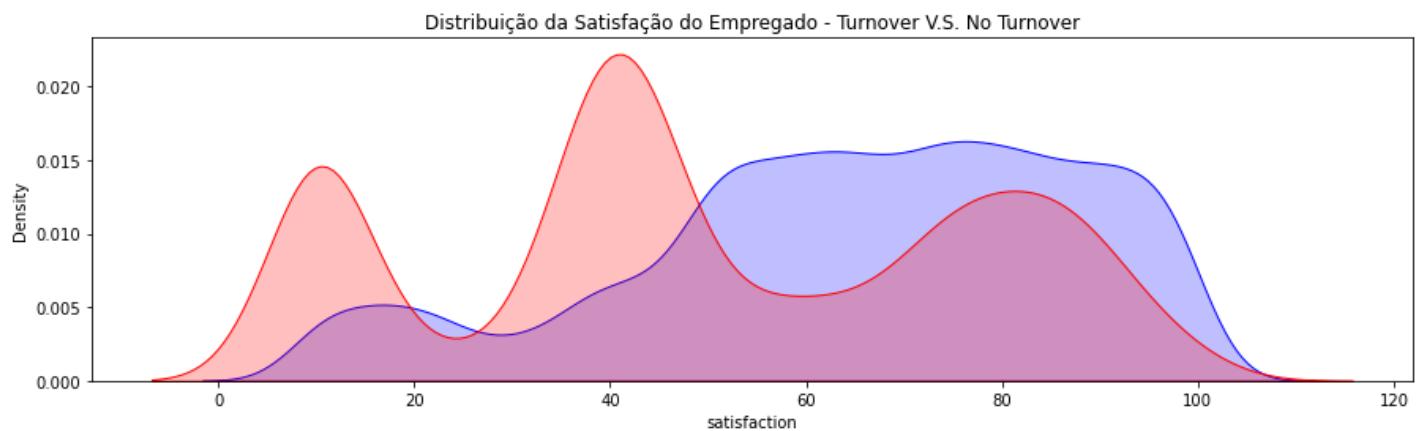
Insights interessantes que encontramos aqui:

- Mais da metade dos empregados com **2, 6 e 7 projetos** deixam a empresa.
- A maioria dos empregados que permanecem na empresa estão envolvidos de **3 à 5 projetos**.
- Todos os empregados que estavam inseridos **7 projetos** deixaram a empresa.
- Existe uma pequena **tendência de crescimento no índice de turnover** em relação à quantidade de projetos.

RESULTADOS

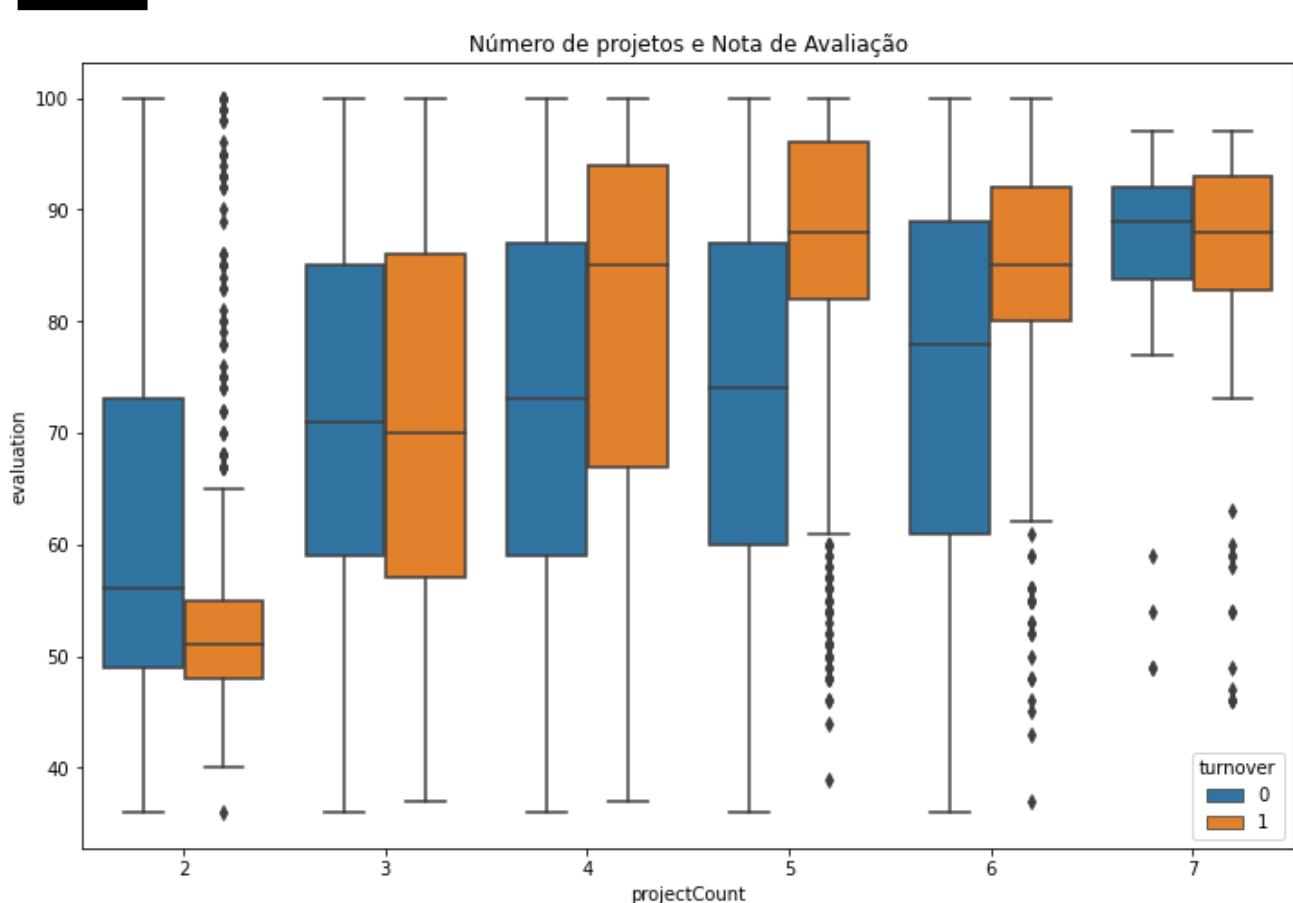


- Temos uma distribuição bimodal para o conjunto que deixou a empresa.
- Colaboradores com **baixa performance** tendem a deixar a empresa.
- Colaboradores com **alta performance** tendem a deixar a empresa.
- O **ponto ideal** para os funcionários que permaneceram está dentro da avaliação de 60 à 80.



- Empregados com o nível de satisfação em 20 ou menos tendem a deixar a empresa.
- Empregados com o nível de satisfação em até 50 tem maior probabilidade de deixar a empresa.

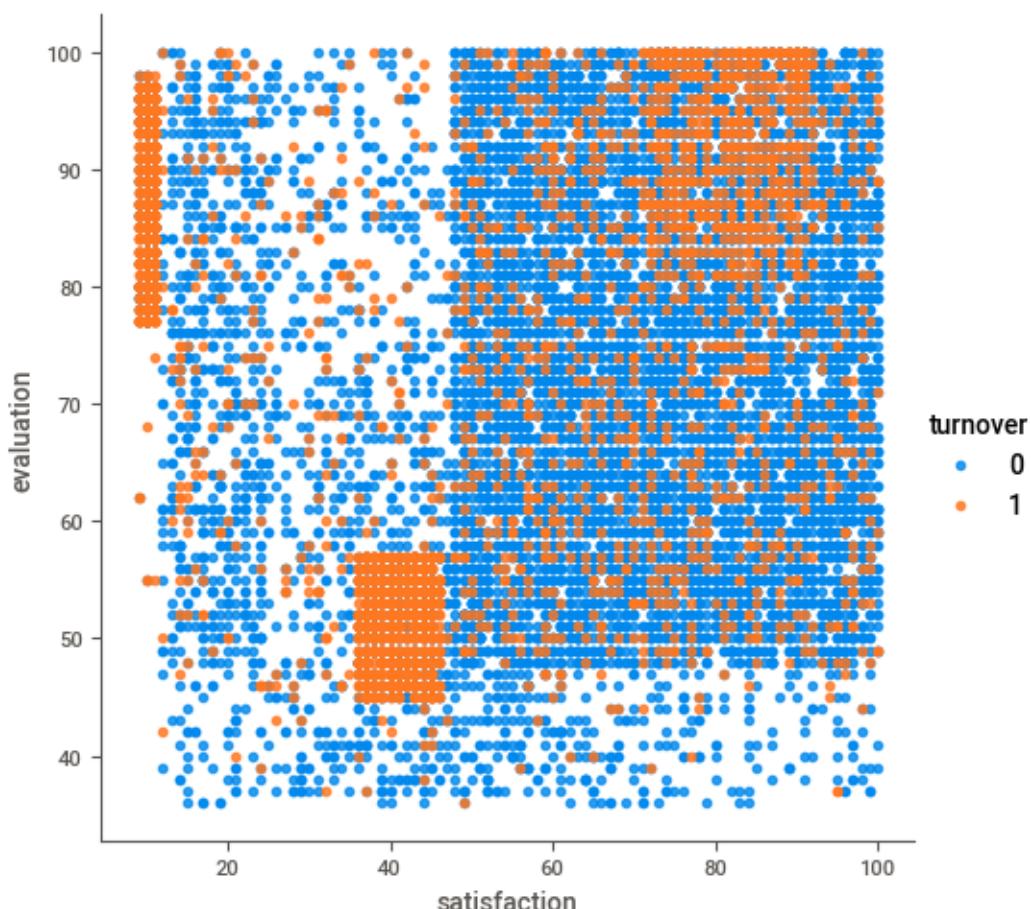
RESULTADOS



Números de projetos VS Nota de Avaliação

- Há um aumento na avaliação para os funcionários que realizaram mais projetos dentro do grupo de quem deixou a empresa.
- Para o grupo de pessoas que permaneceram na empresa, os empregados tiveram uma pontuação de avaliação consistente, apesar do aumento nas contagens de projetos.
- Empregados que permaneceram na empresa tiveram uma avaliação média em torno de 70%, mesmo com o número de projetos crescendo.
- Esta relação muda drasticamente entre os empregados que deixaram a empresa. A partir de 3 projetos, as médias de avaliação aumentam consideravelmente.
- Empregados que tinham dois projetos e uma péssima avaliação saíram.
- Empregados com mais de 3 projetos e avaliações altas deixaram a empresa.

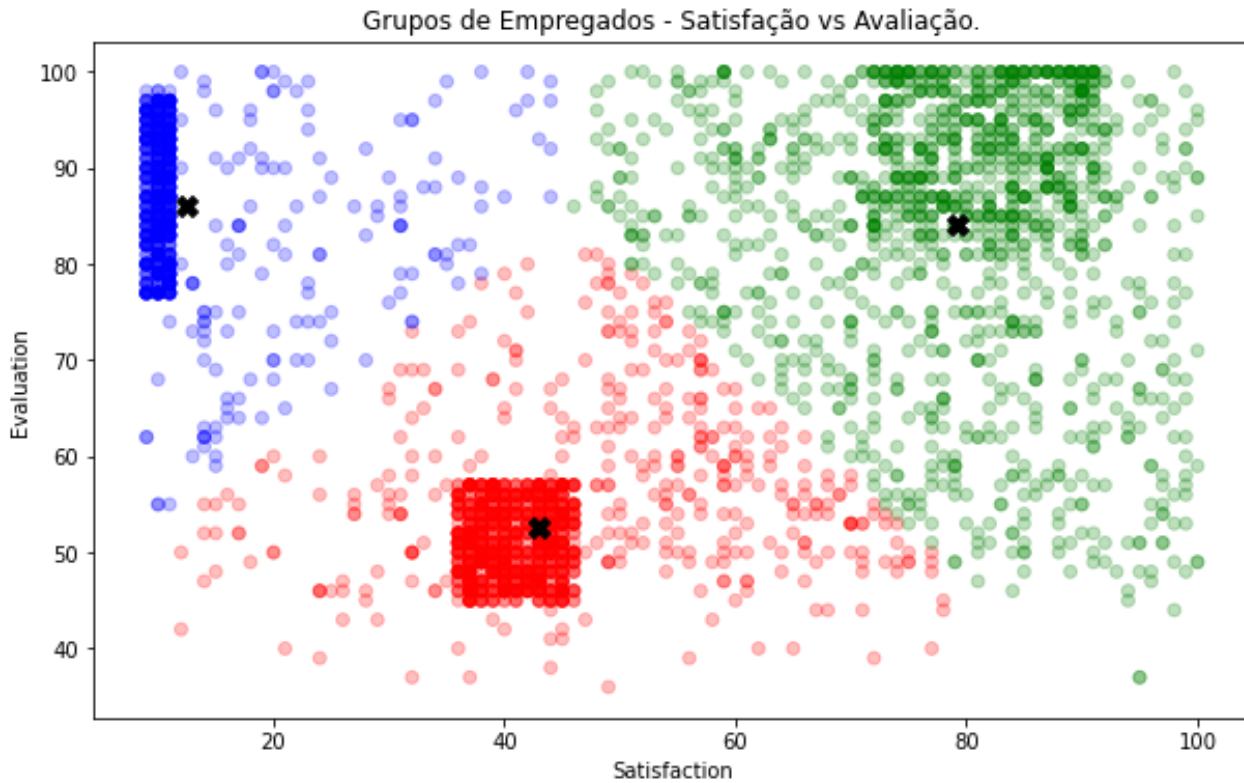
RESULTADOS



Através da análise foi possível desenvolver 3 grupos distintos para agrupar colaboradores que deixaram a empresa por comportamentos similares que são:

- **Grupo 1 (Empregados insatisfeitos e trabalhadores):** A satisfação foi inferior a 20 e as avaliações foram superiores a 75.
- Que corresponde ao grupo de funcionários que deixaram a empresa e eram bons trabalhadores, mas se sentiam péssimos no trabalho.
- **Grupo 2 (Empregados ruins e insatisfeitos):** Satisfação entre 35 à 50 e suas avaliações abaixo de ~ 58.
- Corresponde aos empregados que foram mal avaliados e se sentiram mal no trabalho.
- **Grupo 3 (Empregados satisfeitos e trabalhadores):** Representa os empregados ideais, que gostam do seu trabalho e são bem avaliados por seu desempenho. Este grupo pode indicar os empregados que deixaram a empresa porque encontraram outra oportunidade de trabalho.

RESULTADOS



Análise do Cluster: Satisfação Vs Avaliação

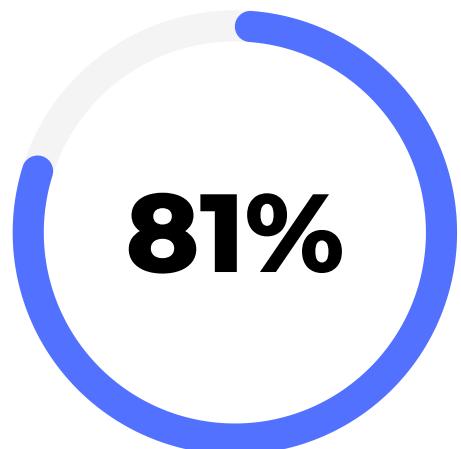
- **Cluster 0 (Verde):** Empregados trabalhadores e satisfeitos.
- **Cluster 1 (Vermelho):** Empregados ruins e insatisfeitos.
- **Cluster 2 (Azul):** Empregados trabalhadores e tristes.

Avaliação do algoritmo de Machine Learning

Para a estimativa com o objetivo de predizer se um empregado vai deixar a empresa, foi implementado um modelo utilizando o algoritmo Gradient Boosting Classifier que atingiu uma performance de AUC em 0.81.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.8283	0.8102	0.6947	0.6257	0.6583	0.5441	0.5455	0.6840
lightgbm	Light Gradient Boosting Machine	0.8453	0.8063	0.6575	0.6815	0.6692	0.5683	0.5685	0.1960
catboost	CatBoost Classifier	0.8466	0.8050	0.6591	0.6849	0.6717	0.5717	0.5719	5.3080
ada	Ada Boost Classifier	0.8021	0.8031	0.7047	0.5681	0.6290	0.4961	0.5016	0.2720
xgboost	Extreme Gradient Boosting	0.8363	0.8009	0.6191	0.6685	0.6427	0.5367	0.5375	0.6440
rf	Random Forest Classifier	0.8113	0.7784	0.5654	0.6126	0.5879	0.4658	0.4666	0.9280
qda	Quadratic Discriminant Analysis	0.7296	0.7618	0.7583	0.4589	0.5718	0.3912	0.4176	0.0820
knn	K Neighbors Classifier	0.7411	0.7577	0.6867	0.4705	0.5581	0.3841	0.3980	0.1320
et	Extra Trees Classifier	0.8020	0.7519	0.5346	0.5934	0.5623	0.4349	0.4359	0.6880
nb	Naive Bayes	0.7177	0.7240	0.5931	0.4324	0.5001	0.3101	0.3176	0.0560
lr	Logistic Regression	0.6816	0.7041	0.6391	0.3956	0.4887	0.2756	0.2923	2.2780
lda	Linear Discriminant Analysis	0.6854	0.7029	0.6339	0.3989	0.4896	0.2789	0.2945	0.0640
dt	Decision Tree Classifier	0.7294	0.6552	0.4878	0.4389	0.4618	0.2818	0.2826	0.0500
svm	SVM - Linear Kernel	0.6600	0.0000	0.6375	0.3760	0.4724	0.2464	0.2650	0.1540
ridge	Ridge Classifier	0.6854	0.0000	0.6335	0.3988	0.4894	0.2787	0.2943	0.1380

**Gradient Boosting Classifier
atingiu uma performance de
AUC em 0.81.**



Data App

Abaixo temos um uma imagem do Data App, projetado para ser uma interface interativa para o usuário enviar dados, e o modelo predizer se o funcionário tem a probabilidade de sair ou não da empresa.

Defina os atributos do empregado para previsão de turnover

Satisfação
50,00 - +

Avaliação
60,00 - +

Média de Horas por mês
168,00 - +

Anos na Empresa
2,00 - +

Realizar Classificação

Human Resource Analytics

Este é um Data App utilizado para exibir a solução de Machine Learning para o problema de Human Resource Analytics.

	dept	salary	averageMonthlyHours	project_satisfaction	evaluation	yearsAtCompany	work	turnover
0	7	1	266	2	38.0000	53.0000	3	0
1	7	2	252	5	80.0000	86.0000	6	0
2	7	2	289	7	11.0000	88.0000	4	0
3	7	1	275	5	72.0000	87.0000	5	0
4	7	1	292	2	37.0000	52.0000	3	0

	satisfaction	evaluation	averageMonthlyHours	yearsAtCompany	Label	Score
0	50.0000	60.0000	168.0000	2.0000	0	0.8595

Grupos de Empregados - Satisfação vs Avaliação.

CONCLUSÃO

Através desse projeto foi possível praticar e implementar conceitos importantes da Ciência e Engenharia de Dados, e propor uma solução para um problema latente e recorrente de qualquer empresa que é a retenção de talentos através da Análise de Dados de Recursos Humanos.

Como um processo de melhoria contínua podemos desenvolver uma automação para executar não só o pipeline de coleta e transformação de dados como automatizar os passos da etapa de Machine Learning e Deploy.

Contato



João Victor Soares Saraiva



Victorjoaosoares88@gmail.com



<https://bit.ly/3hNBU9t>



<https://bit.ly/36k88nF>