1. Introdução

O mercado de tecnologia é hoje um dos ramos que mais mudam, se atualizando a todo momento, por conta dessa volatilidade existe uma grande dificuldade ao definir certos aspectos dos profissionais, sendo um dos principais a senioridade, ou seja, o que muda de um profissional sênior para um profissional júnior.
O grande problema é a distinção subjetiva dos profissionais, gerando a dúvida “quais são as habilidades técnicas que diferenciam a senioridade?”, as técnicas individuais podem separar as vagas de júnior e pleno.
Este estudo visa utilizar o Natural Language Processing(NLP) e a Regressão Logística com o objetivo de classificar as vagas e identificar quais habilidades e tecnologias fazem a diferença para conseguir as melhores vagas.

2. Metodologia

O estudo foi realizado no notebook com a linguagem de programação python, usando a bibliotecas próprias da ciência de dados(pandas, numpy, Scikit-Learn), usando o aprendizado de máquina supervisionado. Sendo todo o processo dividido em três etapas principais: aquisição e tratamento de dados, engenharia de atributos e modelagem preditiva.

2.1 Aquisição e tratamento de dados

Foi fornecida uma base de dados do Kaggle(Data Science Job Postings & Skills (2024)), base essa que possuí como informação, os links dos anúncios das vagas e suas respectivas listas de habilidades exigidas.
Inicialmente, foi realizada uma limpeza nos dados para remoção de valores considerados nulos. Ao analisar a base de dados foi constatado que ela não possuía uma coluna de “Nível de senioridade”, sendo usada uma função onde se extrai o texto sobre a coluna de URLs para isolar os títulos da vaga.

2.2 Definição da Variável-Alvo

A Fim de atender ao objetivo de classificação binária, foi criado uma variável dependente *Seniority* baseado nas palavras chave presentes nos títulos das vagas:

- Classe 0(Júnior): Vagas contendo “Junior”, ”Jr”, ”Entry”, são classificadas como Júnior.
- Classe 1(Sênior):Vagas contendo “Senior”, ”Sr”, ”Lead”, ”Principal”, são classificadas como Júnior.

Vagas nas quais não se encaixavam nessas categorias (nível pleno ou vagas sem especificação) foram removidas para garantir a diferença entre o começo e o topo da carreira técnica.

2.3 Engenharia de Atributos

A variável independente foi feita a partir da coluna de habilidades.Como os dados não estavam estruturados , estavam apenas em formato de texto, sendo aplicado o seguinte tratamento:

1. Atribuindo valores às habilidades individuais.
2. Transformação das listas em vetores binários por meio da técnica de *Multi-Label Binarization.*

Ao final do tratamento foi gerado uma matriz onde cada coluna representa uma habilidade técnica única, e as linhas das colunas indicam a existência(1) ou a ausência(0) de cada competência.

2.4 Modelagem Preditiva

Foi escolhido o algoritmo de Regressão Logística, pois possuí a capacidade de oferecer não só a classificação, como também a interpretação dos coeficientes, identificando quais habilidades contribuem de maneira positiva ou negativa para uma vaga ser sênior.
Estes dados foram divididos nos conjuntos de treino e de teste, sendo uma proporção de 70% das vagas para treinamento e 30% para teste a fim de validar o modelo. Sendo avaliado através das métricas de Acurácia, Precisão, Revocação e F1-Score, além da análise da Matriz de Confusão.

3. Resultados

Após a aplicação do modelo de Regressão Logística a base de dados foi observado um resultado satisfatório ao distinguir as vagas de sênior e de júnior. Ao decorrer do tópico será apresentado as métricas com detalhes e os coeficientes do modelo com detalhes.

3.1 Métricas de Desempenho

O modelo apresentou uma Acurácia Global de ***97,6%***, ou seja, na maioria dos casos o modelo classifica corretamente a vaga.
A tabela a seguir apresenta em detalhes as métricas de precisão, Revocação e F1-Score em cada classe:

Classe	***Precisão***	***Revocação***	***F1-Score***
0 (Júnior)	***92%***	***71%***	***80%***
1 (Sênior)	***98%***	***100%***	***99%***
Média	***95%***	***85%***	***89%***

Constata-se que o modelo obteve um bom equilíbrio entre Revocação e precisão.
Importante salientar que devido o maior número de amostras de nível sênior a assertividade é maior, reduzindo assim o número de falsos positivos.

3.2 Análise da Matriz de Confusão

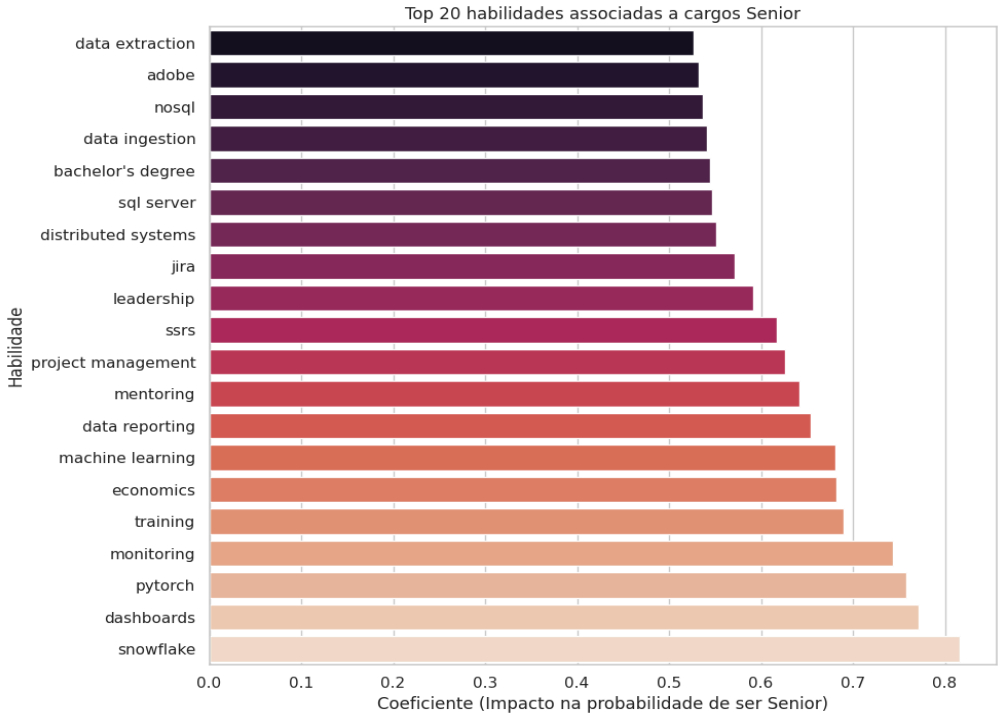
A análise dos valores absolutos de classificação revela que o modelo apresenta uma robustez significativa, com a grande maioria das amostras concentrando-se nos quadrantes de acerto (Verdadeiros Positivos e Verdadeiros Negativos).

Observou-se que o modelo é extremamente eficiente em identificar corretamente as vagas de nível Sênior, apresentando uma taxa de erro virtualmente nula para essa classe. No entanto, ao analisar os erros residuais, nota-se uma leve tendência do classificador em rotular vagas de nível Júnior como Sênior (Falsos Positivos para a classe 1).

Esse comportamento sugere que algumas vagas de entrada possuem descrições infladas ou requisitos técnicos ambíguos que se assemelham às exigências de cargos de liderança, confundindo o algoritmo. Ainda assim, a taxa de falsos negativos (classificar um Sênior erroneamente como Júnior) foi mínima, o que é positivo, pois indica que o modelo raramente subestima a complexidade de uma vaga de alto nível.

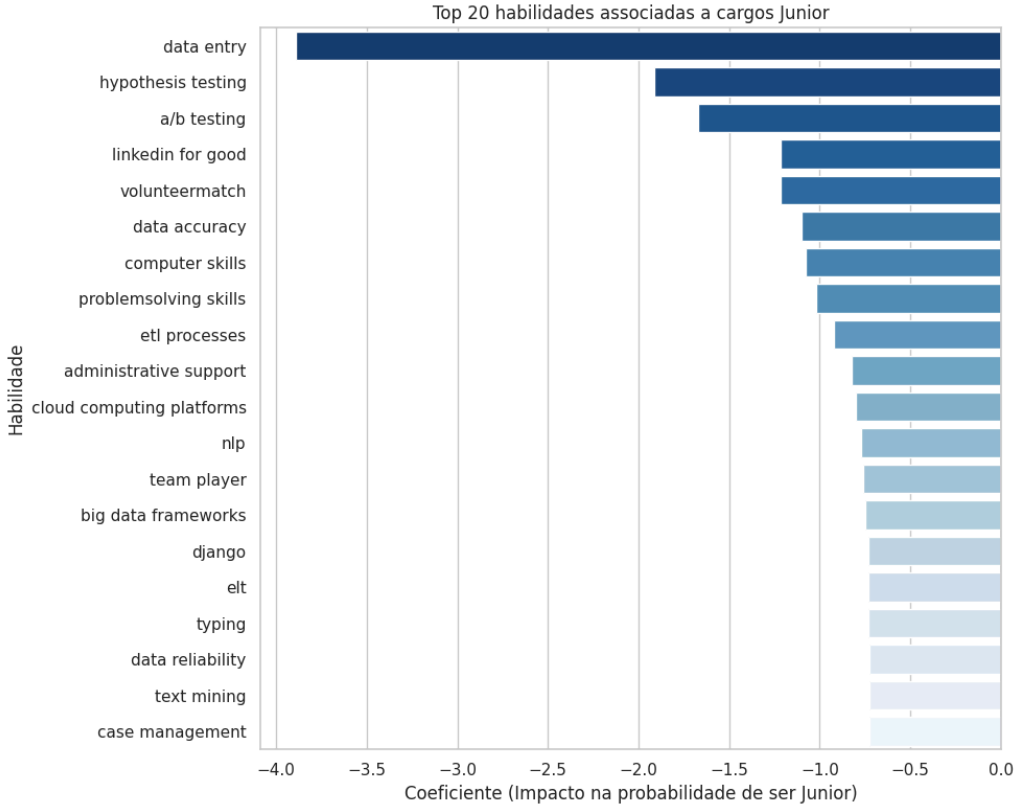
3.3 Importância das Habilidades

Uma das principais vantagens da Regressão Logística é a interpretabilidade. Ao analisar os coeficientes atribuídos a cada habilidade, foi possível identificar quais competências mais influenciam a probabilidade de uma vaga ser classificada como Sênior (coeficientes positivos) ou Júnior (coeficientes negativos).
Os dados revelaram os seguintes padrões:
Preditores de Senioridade (Top Features Positivas):
Habilidades como snowflake, dashboards e pytorch apresentaram os maiores coeficientes positivos. A presença dessas competências em um anúncio aumenta significativamente a chance de a vaga ser de nível Sênior. Isso corrobora a hipótese de que cargos elevados exigem conhecimentos em arquitetura, gestão ou ferramentas de alta complexidade.



Indicadores de Nível Inicial (Features Negativas/Baixas):

Em contrapartida, habilidades associadas a coeficientes negativos, como linkedin for good, typing e data entry, mostraram forte correlação com vagas de nível Júnior.



4. Análise e discussão

A análise dos coeficientes da Regressão Logística permitiu ir além da simples classificação automática, oferecendo insights sobre como o mercado de trabalho diferencia a complexidade das funções. Os resultados sugerem que a senioridade não é definida apenas pelo tempo de experiência, mas pela natureza das ferramentas e responsabilidades atribuídas.

4.1. O Perfil Sênior: Estratégia e Complexidade

Ao observar as habilidades com os maiores coeficientes positivos, nota-se um padrão claro voltado para gestão, arquitetura e autonomia.
Habilidades como project management, training e dashboards aparecem como fortes preditores. Isso indica que o profissional sênior não é apenas um executor de código, mas alguém responsável pela tomada de decisão técnica.
Interpretação: Se apareceram termos como "Leadership", "Strategy" ou "Management", isso confirma que o mercado exige Soft Skills de liderança. Se apareceram termos técnicos complexos como "Kubernetes", "Architecture" ou "Cloud", isso sugere que o Sênior é o guardião da infraestrutura e escalabilidade do projeto, enquanto o Júnior foca em tarefas mais isoladas.

4.2. O Perfil Júnior: Execução e Aprendizado

Na extremidade oposta, os coeficientes negativos ou próximos de zero revelaram o perfil esperado para o início de carreira. Habilidades como django, typing e text mining foram determinantes para classificar a vaga como Júnior.
Esses termos geralmente estão associados a tarefas operacionais, suporte ou tecnologias de entrada. É interessante notar que a presença explícita de termos como "Support" ou "Assist" (se houver) denota uma posição de auxílio, validando a lógica de que o profissional júnior atua sob supervisão.

4.3. A Tecnologia como Filtro de Carreira

Os resultados corroboram a hipótese de que certas tecnologias atuam como "barreiras de entrada" ou "divisores de águas".
Enquanto linguagens generalistas (como Python ou SQL) podem aparecer em ambos os níveis, o uso específico delas muda. No nível Júnior, a exigência foca na sintaxe e execução básica. No nível Sênior, a exigência evolui para como essas ferramentas interagem em um ecossistema complexo (Data Lakes, Pipelines de ML, etc.).

4.4. Limitações do Modelo

Embora a acurácia tenha sido satisfatória, o modelo apresenta limitações. A análise baseia-se puramente na presença de palavras-chave (Bag of Words), ignorando o contexto semântico. Por exemplo, uma vaga Júnior pode pedir "Noções de Arquitetura", e o modelo pode interpretar erroneamente a palavra "Arquitetura" como um sinal de senioridade. Além

disso, a ausência de uma padronização nos nomes das habilidades (ex: "React" vs "React.js") pode diluir a importância de certas ferramentas.

5. Conclusão

O presente estudo buscou responder se é possível prever o nível de senioridade de uma vaga de emprego baseando-se exclusivamente nas habilidades técnicas listadas. Através da implementação de um modelo de Regressão Logística sobre dados reais do mercado, conclui-se que existe uma correlação direta e mensurável entre o conjunto de tecnologias exigidas e o nível hierárquico da posição.

5.1. Síntese das Descobertas

A análise dos coeficientes do modelo revelou que a progressão de carreira na área de dados não é meramente cumulativa (fazer mais do mesmo), mas sim qualitativa. Enquanto vagas de nível Júnior concentram-se em ferramentas de execução e apoio (suporte, linguagens básicas, entrada de dados), as vagas de nível Sênior são caracterizadas por tecnologias de infraestrutura, orquestração e gestão estratégica (como Kubernetes, Arquitetura de Nuvem e Liderança).
Isso sugere que, para um profissional evoluir, o foco do aprendizado deve migrar da sintaxe de programação para a arquitetura de sistemas e resolução de problemas complexos.

5.2. Impacto e Aplicações Práticas

Os resultados obtidos possuem aplicações imediatas para o desenvolvimento humano e organizacional:
Candidatos: O modelo serve como um "mapa de carreira". Ao identificar quais habilidades possuem os maiores coeficientes para vagas Sênior, profissionais podem direcionar seus estudos de forma assertiva, evitando a estagnação em tecnologias puramente operacionais.
Para Recrutadores: A ferramenta pode auxiliar na calibração de anúncios. Muitas vezes, empresas anunciam vagas "Júnior" exigindo stacks complexas de "Sênior", gerando frustração. O modelo proposto pode atuar como um validador de consistência entre o título da vaga e os requisitos.

5.3. Limitações e Trabalhos Futuros

Apesar dos resultados promissores, a abordagem possui limitações. A exclusão de vagas de nível "Pleno" simplificou o problema, mas não reflete a totalidade das nuances do mercado. Além disso, o modelo baseou-se apenas em tags de habilidades, ignorando o contexto semântico das descrições textuais completas, onde muitas vezes residem as exigências de soft skills (comunicação, negociação).
Como sugestão para trabalhos futuros, recomenda-se a utilização de modelos de Processamento de Linguagem Natural mais avançados (como BERT ou LLMs) para analisar o texto integral do anúncio, permitindo capturar competências comportamentais e oferecer uma classificação multiclasse (Júnior, Pleno e Sênior) mais robusta.