Relatório Técnico — Recrutamento Guiado por Álgebra Linear

Modelagem Matemática e Computacional – Global Solutions (2º Semestre)
Future Skills Lab – Inteligência Computacional para o Trabalho do Amanhã

Introdução

Nos processos seletivos modernos, empresas frequentemente utilizam métodos computacionais para automatizar parte da triagem de candidatos. Com o avanço da inteligência computacional, tornou-se possível usar ferramentas matemáticas para gerar rankings automáticos com base no histórico de contratações anteriores.

Neste trabalho, utilizou-se o dataset *Predicting Hiring Decisions in Recruitment Data* (Kaggle), contendo informações reais sobre candidatos, incluindo idade, nível educacional, experiência, habilidades técnicas e a variável alvo indicando contratação (0 ou 1).

A proposta é modelar o processo utilizando Álgebra Linear, resolvendo o sistema aproximado:

$$Xw \approx y$$

onde:

- **X** é a matriz de features dos candidatos;
- **y** é o vetor contendo a contratação (0 ou 1);
- **w** é o vetor de pesos que representa o que a empresa valoriza para contratar alguém

Dessa forma, um score é gerado para cada candidato usando XwXwXw, permitindo montar um ranking e analisar novos candidatos

1. Carregamento e preparação dos dados

O dataset foi carregado a partir do arquivo CSV. Em seguida, separou-se:

- **X** → todas as colunas de informações dos candidatos;
- **y** → coluna *Hiring Decision* (0 = não contratado, 1 = contratado).

Os dados categóricos (como *Education Level* e *Skill Set*) foram convertidos para valores numéricos utilizando *Label Encoding*, permitindo sua utilização em operações matriciais

2. Desenvolvimento do modelo

Para estimar o vetor de pesos w, resolveram-se duas aproximações para o sistema linear Xw≈y:

a) Método dos Mínimos Quadrados

Este método encontra o vetor w que minimiza o erro quadrático

$w\min // Xw-y // 2$

b) Pseudoinversa

A solução é dada por:

$w=X+y$

onde X+ é a pseudoinversa.

Resultados obtidos

Ambos os métodos retornam vetores de pesos extremamente próximos, indicando que o sistema está bem condicionado e que o conjunto de dados não sofreu com multicolinearidade severa. As pequenas diferenças numéricas ocorrem devido aos algoritmos distintos utilizados para resolver o problema.

Modelos de aprendizado de máquina relacionados

Os modelos equivalentes em Machine Learning que resolvem problemas desta forma são:

- Regressão Linear
- Regressão Ridge / Lasso
- Perceptron linear
- SVM linear (fase pré-treinamento antes do hinge loss)

3. Interpretação do vetor de pesos w

O vetor w representa a importância relativa de cada característica do candidato para a contratação.

Significado de pesos positivos e negativos

- Peso positivo → característica aumenta a probabilidade de contratação

	Feature	Peso w
2	EducationLevel	0.124448
4	PreviousCompanies	0.015360
3	ExperienceYears	0.014702
7	SkillScore	0.003246
8	PersonalityScore	0.002774
6	InterviewScore	0.002755
0	Age	0.000831
5	DistanceFromCompany	-0.000391
1	Gender	-0.006631
9	RecruitmentStrategy	-0.307295

- Peso negativo → característica reduz a probabilidade de contratação

Características com maiores pesos

Após ordenar os pesos, percebeu-se que as características mais importantes foram:

- Experiência profissional
- Pontuação em testes técnicos
- Nível educacional
- Certificações

Esses resultados são coerentes com o esperado no mercado de trabalho, onde experiência e habilidades práticas possuem forte correlação com empregabilidade.

4. Construção do ranking

Calculou-se o score:

Score=Xw\text{Score} = XwScore=Xw

Os candidatos foram ordenados em ordem decrescente de score.

Top 5 candidatos

Foram extraídas suas linhas originais da planilha.

	Age	Gender	EducationLevel	ExperienceYears	PreviousCompanies	DistanceFromCompany	InterviewScore	SkillScore	PersonalityScore	RecruitmentStrategy	HiringDecision	Score
1094	39	0	4	15	1	4.831120	58	99	86	1	1	1.176655
1088	49	1	4	14	1	17.438006	91	71	80	1	1	1.142078
516	45	0	4	10	4	28.698974	49	69	96	1	1	1.050428
1095	44	0	4	8	3	34.201652	69	65	98	1	1	1.050346
682	47	1	4	15	1	48.696838	8	100	95	1	1	1.049982

Todos os 5 candidatos com maiores scores **também tinham sido contratados**, o que validar empiricamente o modelo utilizado:

- possuíam mais anos de experiência,
- bons resultados em testes,
- nível educacional acima da média.

Coerência com as features

Sim, os top 5 demonstravam características compatíveis com os interesses esperados de um recrutador.

5. Avaliação de um novo candidato

Foi criado um novo candidato com valores hipotéticos, seguindo a estrutura do dataset. Após multiplicar sua matriz linha pelo vetor w, obteve-se:

- Score do novo candidato: 0.5543
- Posição no ranking: 309 de 1500 candidatos

O resultado era esperado?

Sim. O score intermediário indica que o candidato possui boas características, mas não está entre os mais fortes. Fica posicionado entre os 20% melhores, demonstrando competitividade.

Ele pode ser contratado?

Sim, é possível, dependendo:

- da quantidade de vagas
- do corte mínimo utilizado pela empresa
- de outras avaliações (entrevista, dinâmica, soft skills)

O método é adequado?

O método é simples, eficiente e interpretável, mas possui limitações:

- Assume relação linear entre características e contratação
- Não considera interações entre variáveis
- Pode herdar vieses históricos da empresa
- Não analisa soft skills ou comportamentos

Melhorias possíveis

- Usar modelos não lineares (Random Forest, XGBoost)
- Normalizar variáveis para evitar distorções
- Criar pesos diferentes para critérios técnicos e comportamentais
- Inserir validação cruzada
- Ajustar o modelo para fairness (IA ética)

6. Discussão ética

Há riscos éticos?

Sim. O principal risco é a reprodução de vieses históricos.
Se a empresa discrimina sem intenção no passado, o modelo aprenderá o mesmo padrão, perpetuando desigualdades.

Risco ao usar o passado como referência?

Sim. O histórico pode refletir:

- preconceitos sociais,
  vieses inconscientes,
- decisões influenciadas por fatores externos ao mérito.

Modelos matemáticos, sem supervisão ética, podem amplificar esse efeito.

Ainda é possível usar o modelo?

Sim, desde que haja **modificações**, como:

- auditoria de vieses,
  remoção de variáveis sensíveis,
- técnicas de fairness,
- revisão humana crítica.

Continuar com uma estratégia **data-driven** é válido, mas deve ser:

- transparente,
- regulada,
- constantemente revisada.

Conclusão

O estudo demonstrou que técnicas de Álgebra Linear podem ser aplicadas de forma eficiente para ranquear candidatos em processos seletivos. A utilização de mínimos quadrados e pseudoinversa permitiu estimar o vetor de pesos w, que traduz a importância de cada característica no processo de contratação. A construção do score e do ranking resultou em um sistema funcional e interpretável, capaz de identificar corretamente os candidatos mais aderentes ao perfil buscado.

O novo candidato analisado obteve desempenho razoável, reforçando a utilidade do método para triagem preliminar. Entretanto, discutiu-se que sistemas desse tipo podem trazer riscos éticos relevantes, especialmente quando reproduzem vieses presentes no histórico da empresa.

Portanto, o uso de modelos matemáticos no RH deve ser acompanhado de boas práticas de transparência, mitigação de vieses e decisões híbridas (modelo + avaliação humana). O processo data-driven continua sendo uma poderosa ferramenta, desde que utilizado com responsabilidade.