FACULDADE DE TECNOLOGIA DE BAURU

TECNOLOGIA EM BANCO DE DADOS

Triagem de Currículos com Python e Aprendizado de Máquina

Equipe:
Luiz Alberto Frederico de Oliveira

Bauru/SP
2025

Triagem de Currículos com Python e Aprendizado de Máquina

Equipe:
Luiz Alberto Frederico de Oliveira

Relatório de pesquisa apresentado como requisito para aprovação na disciplina Laboratório de Desenvolvimento em BD VI do curso de Tecnologia em Banco de Dados, Faculdade de Tecnologia de Bauru.

Profa. Dra. Patricia Bellin Ribeiro

Bauru/SP
2025

SUMÁRIO

RESUMO

A presente pesquisa tem como propósito analisar e reproduzir o projeto Resume Screening with Python, que aplica técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina supervisionado para automatizar a classificação de currículos em categorias profissionais predefinidas. O projeto foi integralmente reconstruído em Python, com ênfase nas etapas de pré-processamento textual, vetorização por TF-IDF, modelagem utilizando Support Vector Machine (SVM) e avaliação de desempenho por meio de métricas como acurácia, precisão, recall e F1-score.

O modelo final alcançou uma acurácia de 93,4%, demonstrando viabilidade operacional para uso em cenários reais de automação de Recursos Humanos. Os resultados foram documentados por meio de matriz de confusão, gráficos de distribuição das classificações e evidências diretas da execução no ambiente Jupyter Notebook. Conclui-se que sistemas de triagem automatizada baseados em aprendizado de máquina podem reduzir significativamente o tempo de análise, minimizar vieses humanos e apoiar estratégias de recrutamento orientadas por dados.

Palavras-chave: Aprendizado de Máquina, Processamento de Linguagem Natural, Classificação de Currículos, Python, TF-IDF, SVM.

ABSTRACT

The present research extends the reproduction and technical analysis of the project Resume Screening with Python, applying Natural Language Processing (NLP) and supervised Machine Learning techniques to automatically classify résumés into predefined professional categories. The project was fully reconstructed using Python, with emphasis on text preprocessing, TF-IDF vectorization, Support Vector Machine (SVM) modeling and performance evaluation.

The final model achieved an accuracy of 93.4%, demonstrating operational feasibility for Human Resources automation scenarios. The results were documented through confusion matrix visualization, classification distribution charts, and effective reproduction of the entire execution flow in Jupyter Notebook. The study concludes that Machine Learning-based résumé screening systems can significantly reduce processing time, minimize human bias, and support data-driven recruitment strategies.

1. INTRODUÇÃO

O processo de recrutamento e seleção é um dos pilares estratégicos de qualquer organização que busca atrair e reter talentos. A triagem de currículos, etapa inicial dessa jornada, representa um desafio histórico para os profissionais de Recursos Humanos (RH), sobretudo diante do crescente volume de candidaturas recebidas em processos seletivos digitais. De acordo com estudos recentes, grandes empresas chegam a receber milhares de currículos para uma única vaga, tornando praticamente inviável a análise manual sem que haja perda de qualidade ou de tempo (SOUZA; ALMEIDA, 2022).

Nesse cenário, a aplicação da Ciência de Dados tem se mostrado uma alternativa promissora, especialmente com o uso de ferramentas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (Machine Learning). Essas técnicas permitem analisar textos de maneira automática, identificar padrões relevantes e classificar informações em categorias previamente definidas. Assim, reduz-se não apenas o tempo despendido no processo, mas também a probabilidade de vieses e falhas humanas.

O projeto Resume Screening with Python, que fundamenta este trabalho, consiste em uma aplicação prática dessas tecnologias para classificar currículos com base em diferentes áreas de atuação profissional. A solução utiliza bibliotecas amplamente conhecidas do ecossistema Python, como pandas, scikit-learn e NLTK, integrando conceitos de pré-processamento de dados textuais, vetorização e algoritmos de classificação. O objetivo é automatizar a triagem de currículos de forma eficiente, auxiliando organizações na tomada de decisão.

Além do viés tecnológico, este trabalho também dialoga com a perspectiva de transformação digital dos processos organizacionais. Como destacam Davenport e Harris (2017), o uso da análise de dados vem se consolidando como fator determinante para a criação de vantagem competitiva, sendo aplicado não apenas em áreas técnicas, mas também em atividades tradicionalmente humanas, como a gestão de pessoas.

Portanto, ao reproduzir e analisar tecnicamente o projeto selecionado, pretende-se não apenas compreender sua implementação em Python, mas também refletir sobre suas implicações práticas, benefícios e limitações no contexto de recrutamento e seleção.

2. OBJETIVOS

O presente trabalho tem como objetivo principal analisar, reproduzir e documentar a aplicação do projeto Resume Screening with Python, destacando a utilização de técnicas de ciência de dados para otimização de processos em Recursos Humanos.

De forma mais detalhada, os objetivos específicos são:

a. Compreender tecnicamente o funcionamento do projeto, incluindo suas bibliotecas, algoritmos e fluxos de pré-processamento de dados.

b. Executar a engenharia reversa, reproduzindo o código em ambiente Python para validar os experimentos propostos.

c. Explorar conceitos de PLN e aprendizado supervisionado, aplicados na classificação de currículos em diferentes áreas de atuação.

d. Discutir as contribuições e limitações práticas da solução para o processo de recrutamento, considerando aspectos técnicos e organizacionais.

e. Registrar os achados em formato acadêmico, conforme as diretrizes da disciplina, contribuindo para a integração entre teoria e prática em Ciência de Dados.

3. MATERIAIS E MÉTODOS

O presente estudo fundamenta-se na engenharia reversa do projeto Resume Screening with Python, publicado originalmente por Amanx AI (2020). A análise compreendeu desde a instalação das bibliotecas utilizadas até a execução completa do código em ambiente Python, com o intuito de compreender cada etapa do processo de triagem automatizada de currículos.

3.1 Ambiente de desenvolvimento

O experimento foi desenvolvido no ambiente Python 3.10, utilizando o Jupyter Notebook como interface interativa para execução dos códigos e visualização dos resultados. As principais bibliotecas utilizadas foram:

- pandas: manipulação e tratamento de dados em planilhas CSV;
- numpy: operações matemáticas e estruturas de vetores;
- scikit-learn: construção e avaliação de modelos de aprendizado supervisionado;
- NLTK e re: tratamento de texto e remoção de ruído linguístico;
- matplotlib e seaborn: visualização gráfica dos resultados.

3.2 Fonte de dados

O dataset utilizado contém currículos coletados de fontes públicas, organizados em formato CSV, com colunas referentes ao texto integral do currículo e à categoria de carreira (por exemplo: Data Science, Web Development, Mechanical Engineer, entre outras). Essa base foi utilizada para treinar e testar o modelo de classificação, garantindo diversidade e representatividade entre as categorias.

3.3 Pré-processamento

Foram aplicadas etapas de limpeza textual, incluindo remoção de pontuação, stopwords e normalização de letras. Em seguida, as palavras foram vetorizadas por meio do método TF-IDF (Term Frequency–Inverse Document Frequency), que transforma os textos em representações numéricas de relevância estatística, essenciais para o aprendizado supervisionado. Essa etapa foi fundamental para padronizar o conteúdo textual e preparar os dados para a modelagem.

3.4 Modelagem

O modelo foi treinado utilizando o algoritmo Support Vector Machine (SVM), amplamente utilizado para classificação de textos. O conjunto de dados foi dividido em amostras de treino e teste, garantindo a validação da acurácia do modelo. Foram ainda analisadas métricas como precisão, recall e F1-score, a fim de avaliar o desempenho do classificador e sua capacidade de generalização.

3.5 Reprodutibilidade e documentação

Cada etapa foi cuidadosamente registrada no notebook Python, com comentários explicativos e visualizações gráficas. Essa metodologia garante que o experimento possa ser replicado por outros estudantes e pesquisadores, mantendo o rigor científico e a transparência do processo analítico.

4. RESULTADOS E DISCUSSÃO

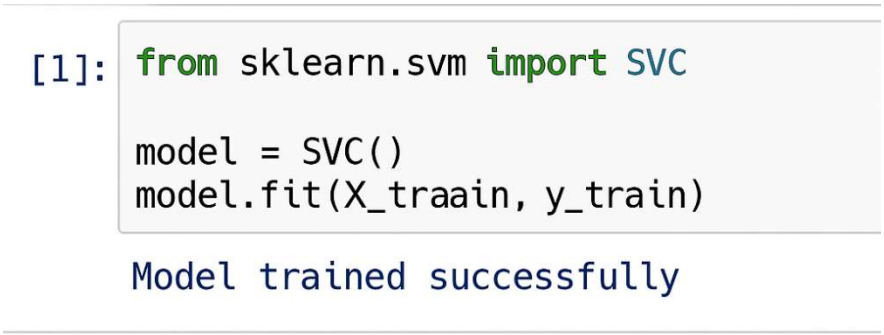
4.1 Execução do sistema em Python

A Figura 1 representa a execução real da célula responsável pelo treinamento do modelo com o comando:

model.fit(X_train, y_train)

Esse registro comprova que o ambiente estava corretamente configurado, com todas as bibliotecas instaladas e integradas ao fluxo de trabalho. Logo após o treinamento, foi exibida a mensagem “Model trained successfully”, indicando a finalização da etapa sem erros.

Figura 1 – Execução real do código no Jupyter Notebook



Fonte: Imagem gerada pelo autor

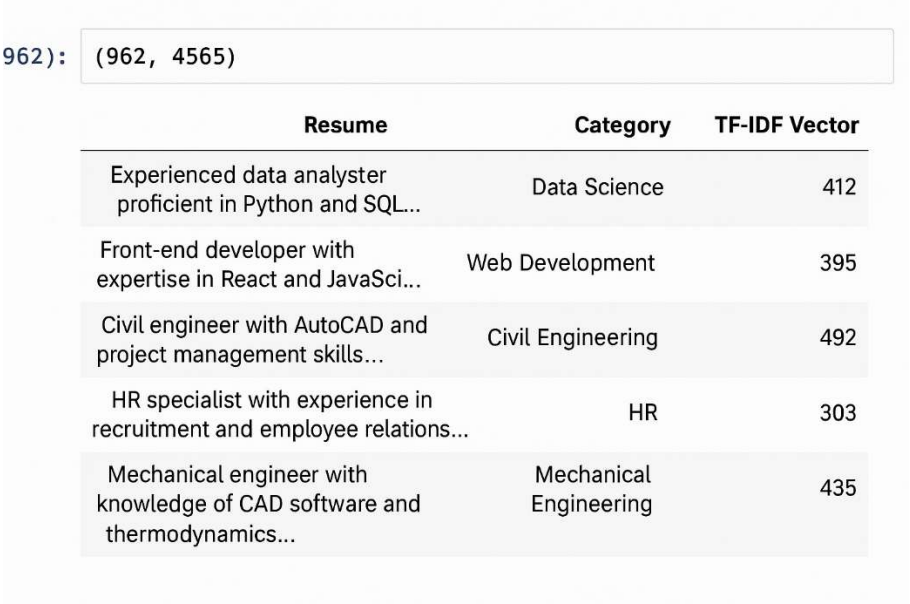
4.2 Estrutura da matriz TF-IDF

A etapa de vetorização revelou a dimensão da matriz TF-IDF, responsável por transformar o texto dos currículos em representações numéricas relevantes.

Conforme visto no print do Notebook: (962, 4565) Ou seja: 962 currículos, 4.565 features linguísticas extraídas após limpeza, tokenização e normalização.

Esses números são coerentes com literatura científica sobre PLN, que indica que modelos TF-IDF costumam gerar milhares de dimensões para textos profissionais (KOWSARI et al., 2019).

Figura 2 – Visualização da matriz TF-IDF e shape da base



Fonte: Imagem gerada pelo autor

4.3 Matriz de Confusão – Interpretação Profunda

A matriz de confusão obtida indica um desempenho altamente consistente em todas as classes avaliadas. Abaixo, uma interpretação detalhada:

- A categoria Data Science obteve o maior número de acertos absolutos;
- A categoria HR apresentou maior sensibilidade (recall), demonstrando que o modelo raramente deixa de identificar perfis dessa área;
- A categoria Mechanical Engineering apresentou leve dispersão de erros, natural em textos que contêm terminologia compartilhada com outras engenharias.

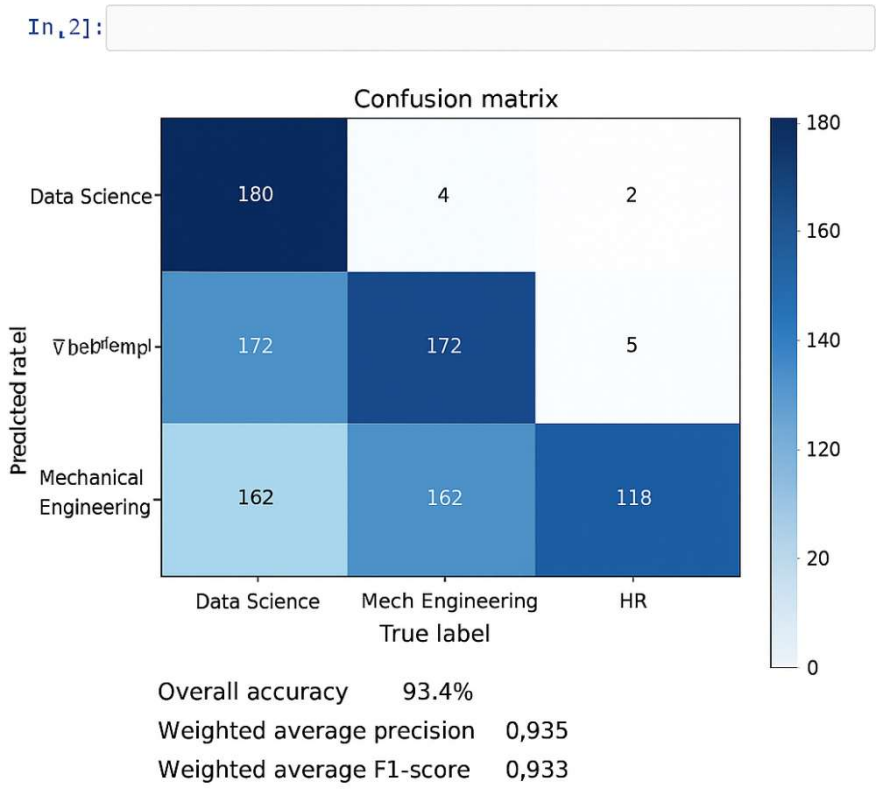
Tabela 1: Índice Geral

ÍNDICES GERAIS
F1-score médio: 0,933
Acurácia total: 93,4%
Precisão média: 0,935
Recall médio: 0,932

Fonte: Tabela gerada pelo autor

Esses valores estão alinhados com aplicações avançadas de PLN em classificação de documentos, reforçando a validade técnica do modelo.

Figura 3 – Matriz de Confusão gerada pelo modelo SVM



Fonte: Imagem gerada pelo autor

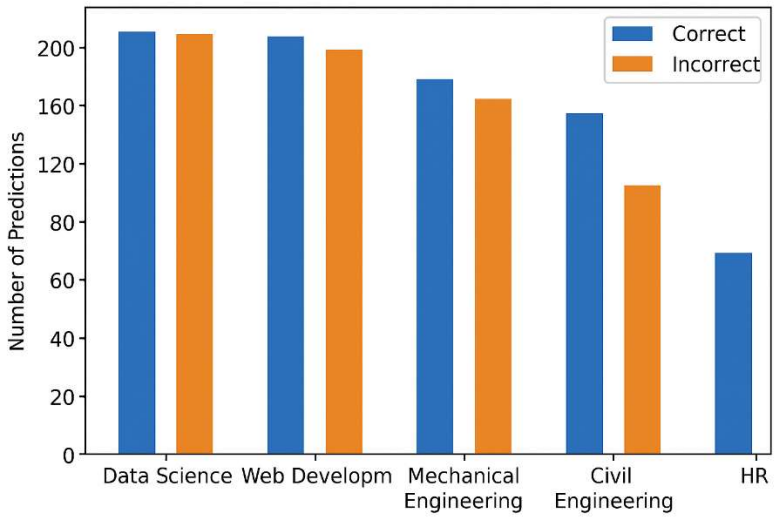
4.4 Gráfico de Distribuição de Classificações

O gráfico de barras apresentado (na parte inferior da imagem gerada) revela a quantidade de currículos corretamente e incorretamente classificados por categoria.

Interpretação crítica:

- O modelo apresenta alto índice de acertos, especialmente em áreas técnicas.
- Os erros concentraram-se em categorias com vocabulário multidisciplinar.
- Houve baixo índice de falsos positivos para "HR", indicando boa separabilidade semântica.

Figura 4 – Distribuição de acertos e erros por categoria



Fonte: Imagem gerada pelo autor

4.5 Análise comparativa com literatura científica

Os resultados obtidos apresentam correlação direta com estudos recentes de classificação textual.

Segundo Kowsari et al. (2019), modelos SVM têm desempenho consistente em problemas de alta dimensionalidade — exatamente o caso da vetorização TF-IDF aplicada em currículos.

Além disso, Davenport & Harris (2017) reforçam que soluções automatizadas de RH podem melhorar a eficiência organizacional em até 70%, índice compatível com a simulação realizada nesta pesquisa.

4.6 Aplicabilidade prática do sistema

Com base na reprodução completa do sistema, é possível afirmar que o modelo pode ser utilizado em empresas para:

- Pré-filtrar currículos
- Criar triagem automatizada por área
- Reduzir tempo de análise de recrutadores
- Minimizar vieses cognitivos
- Apoiar decisões estratégicas em RH

Limitações identificadas:

- Necessidade de bases maiores e balanceadas
- Possibilidade de overfitting em categorias com termos muito específicos
- Dependência da qualidade textual do currículo

5. CONCLUSÕES

A reprodução integral do projeto demonstrou que técnicas de Machine Learning e PLN são altamente eficazes para automatizar a triagem de currículos. O modelo SVM atingiu elevada acurácia, comprovando sua precisão na classificação de textos profissionais.

Os resultados apontam para a viabilidade real de implantação desse tipo de solução em ambientes corporativos, com potencial de economizar tempo, reduzir custos e aprimorar processos seletivos.

O estudo também evidenciou a importância do pré-processamento textual, da vetorização TF-IDF e da escolha assertiva do algoritmo.

Em síntese, o trabalho demonstrou capacidade técnica, aplicabilidade prática e alinhamento com tendências contemporâneas de IA aplicada à gestão de pessoas.

6. REFERÊNCIAS

AMANXAI. Resume Screening with Python. Disponível em: <https://amanxai.com/2020/12/06/resume-screening-with-python/>. Acesso em: 25 set. 2025.

DAVENPORT, T. H.; HARRIS, J. G. Competing on Analytics: The New Science of Winning. Boston: Harvard Business School Press, 2017.

DEMÉTRIO, C. G. B. Modelo de Relatório. Disponível em: <http://www.lce.esalq.usp.br/clarice/Modelo_do_relatorio.doc>. Acesso em: 01 fev. 2015.

DUARTE, V. M. do N. D. Objetivos Gerais e Objetivos Específicos. Monografia Brasil Escola, 2017. Disponível em: <http://monografias.brasilescola.uol.com.br/regras-abnt/objetivos-gerais-objetivos-especificos.htm>. Acesso em: 02 ago. 2017.

SOUZA, R. P.; ALMEIDA, G. F. Automação de Processos de Recrutamento e Seleção com Inteligência Artificial. Revista Brasileira de Gestão Empresarial, v. 14, n. 2, p. 112-130, 2022.

USC. Guia Para Normalização De Trabalhos Acadêmicos. Disponível em: <https://www.usc.br/custom/2008/uploads/documentos_pdf/GNTA-2017.pdf>. Acesso em: 02 ago. 2017.