Veridia Internship Data Analysis Report

This project was completed as part of Veridia's internship program to enable data-driven decision-making in recruitment and operations. The goal was to analyze, visualize, and derive insights from a large collection of resumes to improve recruitment strategies.

Dataset Overview

The dataset was sourced from Kaggle. It contains over 2400 resumes collected from livecareer.com, categorized into various job domains such as Information Technology, HR, Finance, Engineering, and more. The dataset includes the following key fields:

• ID – Unique identifier for each resume

• Resume_str – Resume content in text format

• Category – Job category of the resume

Steps Performed

• Data Cleaning & Preprocessing – Removed duplicates, handled missing values, and cleaned text using regex, stopword removal, and lemmatization.

• Exploratory Data Analysis (EDA) – Conducted statistical and visual analysis to identify patterns in resume categories and skill distributions.

• Data Visualization – Generated bar charts, word clouds, and interactive visuals using Matplotlib, Seaborn, and Plotly.

• Model Training – Used TF-IDF Vectorization and trained an SVM model for category prediction.

• Model Evaluation – Achieved 72% accuracy using Support Vector Machine (SVM) classifier.

Key Insights

• IT, Engineering, and HR are the most common resume categories.
• Candidates with Python, SQL, and Machine Learning skills were mostly categorized under IT and Data Science.
• WordCloud analysis showed frequent use of keywords like 'management', 'project', and 'customer'.
• Predictive modeling can help recruiters automatically classify resumes into relevant job domains.

Tech Stack & Tools Used

• Python (Pandas, NumPy, Matplotlib, Seaborn, Plotly)

• NLTK for text preprocessing

• Scikit-learn for machine learning modeling

• KaggleHub for accessing datasets directly

• Jupyter Notebook for analysis and documentation

Conclusion

The project successfully demonstrated the use of Natural Language Processing and Machine Learning in resume categorization. With 72% model accuracy, the system can assist HR departments in screening resumes faster and improving hiring efficiency. Future enhancements could include deep learning models, automated dashboards, and integration with applicant tracking systems.