Veridia Internship Data Analysis Report

This report summarizes the data analysis and predictive modeling performed on the Resume Dataset (https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset) as part of the Veridia Internship project. The goal was to analyze, visualize, and derive insights to support data-driven recruitment decisions.

Key Steps Performed:
• Data loading from Kaggle using KaggleHub
• Data cleaning and preprocessing (duplicates, missing values, stopword removal, lemmatization)
• Exploratory Data Analysis (EDA) – visualized resume categories and text distribution
• Feature extraction using TF-IDF Vectorizer
• Predictive Modeling using Support Vector Machine (SVM)
• Evaluation of model performance using accuracy and classification report
• Prediction of job category from new resume input

Model Results:
• Algorithm Used: Support Vector Machine (LinearSVC)
• Achieved Accuracy: ~92% on test data
• Model successfully predicts job category from unseen resumes.

Insights & Recommendations:
• The dataset shows clear separation among technical and non-technical categories.
• Data Science, Engineering, and IT resumes share overlapping skill patterns.
• HR and Business Development roles show strong textual indicators around communication and leadership.
• The SVM model can be integrated into a resume screening tool for automated candidate classification.

Tools & Libraries Used:
Python, Pandas, NLTK, Scikit-learn, Matplotlib, Plotly, WordCloud, KaggleHub, ReportLab