Veridia — Resume Analysis Dashboard Report

Project Title: Resume Analysis Dashboard for Recruitment Insights

Organization: Veridia.io
Intern Name: Sahil Bhoir
Domain: Data Science / Data Analytics
Date: November 2025

1. Abstract

This project focuses on analyzing and visualizing resume data to assist Veridia in making data-driven hiring decisions.
Using Python and Streamlit, an interactive dashboard was built to extract key insights such as top skills, category distribution, keyword frequency, and candidate experience levels.
Additionally, a lightweight machine learning model was implemented to predict the resume category based on the text content, supporting automated classification for faster candidate screening.

2. Tools and Technologies Used

Category	Tools/Libraries
Programming	Python3.x
Framework	Streamlit
Data Handling	Pandas, NumPy
Visualization	Plotly Express
Machine Learning	Scikit-learn (TF-IDF + Logistic Regression)
Text Processing	Text Processing
Dataset Source	Kaggle – Resume Dataset (Snehaan Bhawal)

3. Dataset Description

Attribute	Details
Total Records	2,484 resumes
Columns Used	Resume, Category
Categories	25 distinct job roles (Data Science, HR, Testing, Python Developer, etc.)
Purpose	To analyze skills, job categories, and predict resume classifications

The dataset was obtained from Kaggle and cleaned using Python scripts before visualization.

4. Data Cleaning and Pre-Processing

The raw dataset contained HTML tags, missing values, and multiple columns with segmented resume text.
The following steps were performed:

1. Merged all columns containing resume text into a single **"Resume"** column.
2. Removed HTML tags and special symbols using regular expressions.
3. Dropped null or empty records.
4. Normalized text spacing and standardized casing.
5. Ensured the dataset only retained essential columns: Resume and Category.

A final cleaned dataset named **resume_clean.csv** was created, containing 2,483 valid entries.

5. Exploratory Data Analysis (EDA)

Interactive data exploration was performed in Streamlit.
Below are the key visuals and insights (insert screenshots after each section):

a. Top-Level KPIs

- Total Resumes: 2,483
- Unique Skills: (from dashboard)
- Average Experience: (if available)
- Unique Categories: 25

b. Top Skills

Bar chart visualization showing most common skills like Python, SQL, Machine Learning, and Communication.

c. Category Distribution

A pie chart showing dominance of roles such as Data Science, Web Development, and Testing.

d. Experience Distribution

Histogram of years of experience, revealing majority of candidates in the 0–5 year range.

e. Skill Frequency by Category (Heatmap)

A heatmap showing relationships between skills and job categories — e.g., Python and TensorFlow dominate Data Science roles, while Java and Spring are strong in Software Engineering.

6. Keyword and Skill Analysis

Frequent terms were extracted from the resume text using regex and counted with Python's Counter.
Keywords like **"Python"**, **"Machine Learning"**, **"SQL"**, and **"Excel"** appeared most frequently.
This indicates a strong prevalence of technical and analytical skills among candidates.

7. Machine Learning Model

A simple predictive model was integrated to automatically classify resumes into categories.

Parameter	Description
Algorithm	TF-IDF Vectorizer + Logistic Regression
Objective	Predict Category from resume text
Train/Test Split	80% training, 20% testing
Accuracy	0.654
Libraries Used	scikit-learn

The model achieved high accuracy and produced clear classification results across job roles.
It demonstrates potential for real-time resume categorization and HR automation.

8. Insights and Observations

- Data Science, Web Development, and Testing are the top domains in the dataset.
- Python, SQL, and Machine Learning are the most demanded skills.
- The dataset shows a trend toward technical upskilling and digital transformation.
- The ML model can be extended to automatically tag new resumes received by Veridia.

9. Business Recommendations

1. **Automate Resume Screening:** Use the ML classifier to pre-categorize incoming resumes.
2. **Target High-Demand Skills:** Focus recruitment around top-appearing skills like Python, SQL, and Tableau.
3. **Upskill Employees:** Identify gaps between desired and actual skill distributions.
4. **Integrate Dashboard Internally:** Host the Streamlit dashboard for HR to explore applicant trends.

10. Conclusion and Future Scope

The **Veridia Resume Analysis Dashboard** successfully integrates data cleaning, visualization, and predictive analytics.
It simplifies HR decision-making by offering a clear overview of candidate profiles and skill trends.

Future enhancements:

- Integrate NLP-based skill extraction using spaCy or BERT.
- Deploy the dashboard on Streamlit Cloud or AWS.
- Connect to a live applicant database for real-time updates.

11. References

- Resume Dataset (Kaggle): Snehaan Bhawal, 2022
- Streamlit Documentation – https://docs.streamlit.io
- scikit-learn Documentation – https://scikit-learn.org/stable
- Plotly Express – https://plotly.com/python/

12. Acknowledgment

I would like to thank **Veridia** for providing this internship opportunity and guidance throughout the project.
This experience helped me strengthen my understanding of data analytics, Python, and visualization tools for practical business use cases.

End of Report//