Big Data Analysis Using Pyspark

Internship Project – CODTECH

🔹 Introduction

In this project, I performed big data analysis using **Pyspark** on a dataset of 30,000 job records. The goal was to extract meaningful insights related to job titles, industries, AI impact, salary trends, automation risk, and future job growth.

🔹 Dataset Overview

- Total Rows: 30,000
- Total Columns: 13
- Data contained job information like:
    - Job Title
    - Industry
    - AI Impact Level
    - Median Salary
    - Experience Required
    - Projected Job Openings (2030), etc.

🔹 Technology Used

- Google Colab (Python)
- PySpark
- Pandas (for plotting)
- Matplotlib / Seaborn (for visualization)

🔹 Steps Performed

1. Data Cleaning

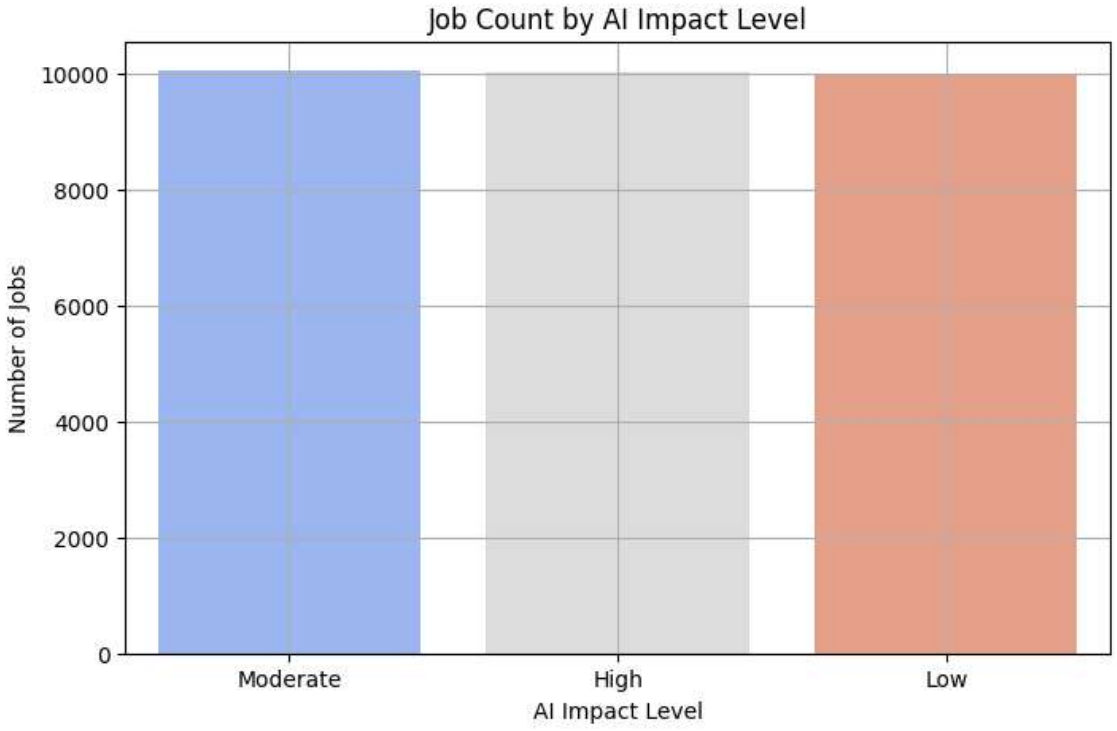
- Checked shape, column types, missing values

- Cleaned the dataset for analysis

2. Data Exploration & Analysis

- Found top job titles and top industries
- Analyzed “AI Impact Level” distribution:
    - ~33% High Impact jobs (at risk from AI)
    - ~33% Low Impact jobs (safe for future)
- Studied salary vs automation risk
- Identified top 10 fastest-growing jobs till 2030
- Checked gender diversity patterns across jobs

3. AI Impact Visualization



- A bar chart was used to show how many jobs fall under each AI impact category.

🔹 Key Insights

- Some high-paying jobs still have high automation risk (e.g., restaurant manager, meteorologist)
- Healthcare, education, and finance showed strong job growth

- Academic librarian, surgeon, and investment banker roles will grow sharply by 2030
- Gender-balanced roles include professor, illustrator, call center manager

---

🔹 Conclusion

This analysis shows that **AI will significantly reshape the job market**, affecting some high-paying roles while enhancing others. Future-focused roles with creativity, human judgment, and emotional intelligence will remain more secure. This project demonstrates the use of **PySpark for scalable data analysis** on large datasets.
