# Milestone Report 1

# Preprocessing and Dataset Overview

Group I

Beema Sajeev

Jaisy Joy

ParminderKaur

Harsh Kundal

Diksha

Capstone-project/milestone-1.ipynb at main · Beema20/Capstone-project (github.com)

## Table of Contents

## Introduction

The project aims to analyze the LinkedIn Job Postings dataset from 2023 to 2024 to extract insights into job market trends, salary predictions, and remote work dynamics. By exploring this dataset, we seek to provide valuable insights for job seekers, employers, and policymakers.

## Progress Report

We have made significant progress in the initial phase of the project. Key tasks completed include:

- **Data Preparation**: We successfully loaded and cleaned the datasets, handling missing values, duplicates, and inconsistencies. The datasets were merged to create a unified dataset for analysis.

## Dataset Overview

- Companies: Contains information about various companies including their name, description, size, location, and website URL.
- Company Industries: Provides details about the industries to which companies belong.
- Company Specialties: Lists specialties associated with different companies.
- Employee Counts: Includes data on the number of employees and follower counts for different companies.
- Benefits: Describes the benefits offered by companies, linked to specific job positions.
- Job Industries: Specifies industries associated with different job positions.
- Job Skills: Lists skills required for different job positions.
- Salaries: Provides salary information for different job positions.
- Industries: Describes various industries along with their IDs.
- Skills: Specifies skills along with their abbreviations and full names.

- Postings: Contains postings for job positions, including titles, descriptions, locations, and salary details.

## Preprocessing Steps

1. Data Cleaning: Removed any duplicate entries, missing values, or irrelevant columns.
2. Standardization: Ensured consistency in formatting across datasets (e.g., company names, job titles).
3. Integration: Combined related datasets where applicable (e.g., joining company information with job postings).

## Key Findings

The data provided includes information about companies, job postings, industries, skills, salaries, employee counts, benefits, and job specialties.

Companies are described with details like their name, description, size, location, and website URL. Job postings contain information such as job title, description, salary, pay period, location, company ID, and various other details related to the job.

Industries and skills are listed along with their respective IDs and names. Salaries are detailed with salary IDs, job IDs, maximum, median, and minimum salaries, currency, and compensation type.

Employee counts provide data on the number of employees and followers for each company, along with the time the data was recorded. Benefits include types of benefits inferred from job postings. Job specialties are listed with company IDs and specialties.

Overall, the data covers a wide range of information about companies, job postings, industries, skills, salaries, employee counts, benefits, and job specialties, providing a comprehensive overview of various aspects related to employment and companies.

### Next Steps

- Exploratory Data Analysis (EDA): Conduct in-depth analysis to uncover further insights.
- Feature Selection: Identify the most relevant features for modeling.
- Modeling: Develop predictive models to forecast salaries, analyze job trends, or classify companies based on various attributes.
- Evaluation: Assess model performance and iterate as necessary.
- Deployment: Deploy models into production for real-world applications.

## Deviation from the Original Plan

While the initial timeline was followed closely, more time was allocated to data cleaning and preprocessing due to the complexity of the datasets. As a result, the timeline for EDA and feature engineering was slightly adjusted.

### Modified Timeline Table

| **Milestone** | **Description** | **Timeline** |
|---|---|---|
| Project Kickoff | Initial meeting, project planning, data acquisition | Week 1 |
| Data Preparation | Understanding the data and preprocessing | Weeks 5 |
| Preliminary Analysis | Initial insights and trend identification | Weeks 6 |
| Model Development | Developing prediction models | Weeks 10 |
| Data Visualization | Creating visualizations and developing dashboards | Weeks 13 |
| Reporting | Compiling the final report | Week 14 |

| | | |
|---|---|---|
| Final Presentation | Presenting findings to stakeholders | Week 15 |

## Expected Outcomes and Goals for the Next Phases

- Deployment of predictive models into production for real-world applications.
- Presentation of findings and recommendations to stakeholders.
- Finalization of the project report detailing methodology, findings, and visualizations.

## Challenges Faced

While progress has been steady, some challenges encountered during the project include:

- Handling complex data structures and inconsistencies in the datasets.
- Balancing time between data cleaning, analysis.
- Ensuring accuracy and performance with limited data.

## Conclusion

The preprocessing phase has provided a clean and structured dataset ready for further analysis and modeling. The next steps will involve diving deeper into the data to extract actionable insights and build predictive models to support decision-making processes.