Project Report

1. Introduction

This project focuses on analyzing employment outcomes of engineering graduates with respect to salaries, specialization, and demographic attributes. The study integrates data preprocessing, exploratory data analysis (EDA), statistical inference, and optimization modeling to derive meaningful business insights and decision-support tools.

2. Data Understanding

- Dataset: 4000 records × 40 attributes (graduate demographics, academic scores, job information, salaries).
- Target Variable: Salary.
- Features: Gender, Specialization, Academic Scores (cognitive, technical, personality skills), Job Profile, etc.

3. Data Preprocessing

- **Imputation:** Handled missing values using mean/median for numerical variables and mode for categorical variables.
- **Outlier Detection:** Applied IQR method to identify and cap extreme values.
- **Encoding:** Implemented One-Hot Encoding and Label Encoding for categorical features.
- **Feature Engineering:** Derived new features such as standardized score indices and categorized salary ranges.

4. Exploratory Data Analysis (EDA)

- **Univariate Analysis:** Histograms, Boxplots, and CDFs to study distribution of salary and test scores.
- **Bivariate Analysis:** Scatterplots, Crosstabs, Pivot tables, and Barplots to examine relationships between variables.

- **Observations:**
    - Salary variation across specializations.
    - Countplot analysis revealed city-based job concentration.
    - Boxplots showed outliers in salary distribution.

5. Statistical Inference

- **t-tests:** One-sample t-tests applied to benchmark job-specific salaries against expected salary ranges.
- **Chi-Square Test:** Analyzed dependency between Gender and Specialization; results showed significant association ($p < 0.05$).

6. Machine Learning & Optimization Models

- **Supervised Learning:** Applied Linear Regression and k-Nearest Neighbours for salary prediction.
- **Unsupervised Learning:** Implemented k-Means clustering and PCA for dimensionality reduction and pattern discovery.
- **Optimization Models:**
    - Developed **constrained optimization models** in Pyomo.
    - Solved using CBC and GLPK solvers.
    - Implemented **unconstrained optimization algorithms** (Gradient Descent, Newton’s Method, BFGS) coded from scratch.
- **Metaheuristic Comparison:** Benchmarked **MILP models** against Genetic Algorithms (GA) and Simulated Annealing (SA) for solving complex optimization problems, analyzing convergence speed and solution quality.

7. Results & Insights

- Salary distribution differs significantly across specialization and gender categories.
- Statistical tests confirmed dependencies between categorical features.

- Optimization experiments revealed that **metaheuristics (GA, SA)** provided near-optimal solutions faster than MILP in large-scale scenarios.

- ML models highlighted key predictors of salary outcomes.

8. Conclusion

This study demonstrates an end-to-end data analytics and optimization pipeline, from **data cleaning & statistical inference** to **machine learning & mathematical programming**. The integration of statistical methods and optimization frameworks provides both descriptive insights and prescriptive decision-making tools for workforce and salary analysis.