PROJECT ABSTRACT

Salary Prediction using Regression

Salary prediction from years of experience using linear, ridge, polynomial and random forest regression on the real Kaggle Salary Dataset, with train/test evaluation on MAE, RMSE and R2, delivered built-to-order with code, notebook, demo app, report, PPT and viva kit.

01. Title

Salary Prediction using Regression

02. Introduction / Background

Salary benchmarking matters to students choosing careers, HR teams framing offers and analysts studying labor markets, yet most regression teaching stops at textbook equations with toy numbers that never touch a real dataset. This project makes regression concrete end to end: a genuine public dataset of employee experience and salary figures, exploratory analysis, four real regressors, and comparison with proper train/test evaluation. Because the experience-pay relationship is intuitive, every modeling decision is visible in the plots and directly discussable in the viva.

03. Problem Statement

Regression is usually taught as mathematics without data, or demonstrated on fabricated CSVs whose 'results' cannot be questioned. Students need a project where the dataset is real and public, the evaluation protocol is fixed and reproducible, and the comparison between models is fair: same split, same metrics, same test set. This project is built exactly that way around the well-known Kaggle Salary Dataset.

04. Objectives

- Use the public Kaggle 'Salary Dataset - Simple Linear Regression' (30 records: YearsExperience, Salary) as published, with the source cited.
- Perform exploratory analysis (distributions, scatter, correlation, outliers) before modeling.
- Train Linear, Ridge, second-degree Polynomial and Random Forest regressors on a fixed 80/20 split (random_state=42) and compare on test-set MAE, RMSE and R2 with residual plots.
- Build an interactive demo that predicts salary from experience with the fitted curve shown, and deliver the complete student kit: code, notebook, report, PPT and viva Q&A.

05. Proposed Methodology

The CSV is loaded, checked for missing values and profiled. EDA covers the experience-salary scatter, salary distribution and correlation. Features and target are split 80/20 with a fixed seed. Four regressors are trained; Ridge alpha and forest depth are chosen by cross-validation on the training data only. Each model predicts the 6-record held-out test set and is scored with MAE, RMSE and R2; residual plots expose systematic errors. Fixed seeds and pinned requirements make every number reproducible.

06. System Architecture / Working

The analysis notebook runs the full pipeline: load, EDA, split, train four models, evaluate on the test set, and plot fits with residuals. The comparison script renders the metric table used in the report. The Flask demo loads the trained models, accepts years of experience through a slider, and returns the predicted salary together with the model's test metrics and the regression curve drawn over the data scatter.

07. Hardware / Software Requirements

- Python 3.10 with pandas, NumPy
- scikit-learn (LinearRegression, Ridge, PolynomialFeatures, RandomForestRegressor, metrics)
- Matplotlib and Seaborn for EDA and regression plots
- Jupyter notebook for the buyer-run analysis
- Flask demo application
- Standard laptop/desktop; CPU only

08. Expected Outcomes

A complete, reproducible regression study on real data: EDA, four fitted models, test-set MAE/RMSE/R2 computed live by the notebook, residual analysis, and an interactive predictor. No metric is pre-claimed; the report quotes the numbers the buyer's own notebook run produces under the fixed-seed protocol. The bias-variance discussion (why the forest overfits 30 rows, what the polynomial buys) is grounded in the student's own plots.

09. Applications

- Teaching demonstrator for regression coursework: OLS, regularization, bias-variance, metrics
- Template for any univariate tabular regression mini-project
- Viva-ready case study in honest small-data evaluation
- Foundation for a multi-feature salary study with a richer dataset (optional customization)

10. Future Scope

- Multi-feature extension with role, education and location on a richer dataset
- Cross-validation and learning-curve analysis across sample sizes
- Quantile regression for salary-band prediction intervals
- Deployed web demo with saved model artifacts
- Fairness audit discussion: what a single-feature pay model omits