ITAI 2377 – Data Sciences in Artificial Intelligence
Professor: Viswanatha Rao
Student: Natalia Solórzano Pérez W207818526
Spring 2025 CN: 21229

Final ML Project Report – Salary Prediction

Introduction

The goal of this project is to explore how machine learning can be used to predict an individual's salary based on their years of professional experience. We chose this topic because compensation prediction is a practical and widely relevant use case in HR tech, job boards, career planning, and recruiting. With access to a clean and interpretable dataset from Kaggle, this project allows for hands-on application of regression algorithms, feature engineering, and model evaluation. By the end of the project, we aim to identify which model best captures the relationship between experience and salary and to demonstrate how ML can automate and enhance salary forecasting tools.

Dataset Source

The dataset used was obtained from Kaggle:

Title

- [Salary Dataset -Simple Linear Regression] (https://www.kaggle.com/datasets/abhishek14398/salary-dataset-simple-linear-regression)

Source

- Uploaded by Abhishek Sharma

Columns

- *YearsExperien*ce: Number of years the individual has worked
- S*ala*ry: The corresponding salary (in USD)

This dataset was chosen for its simplicity and clear linear relationship, making it ideal for comparing multiple regression models and exploring model generalization on clean data.

Problem Statement

Can we predict an individual's salary based on their years of professional experience using machine learning regression models? This project aims to implement multiple regression models on a clean, simple dataset and compare their performance.

Dataset Description

The dataset contains two columns: '*YearsExperience*' and '*Salary*'. It was sourced from Kaggle and is ideal for regression modeling due to the strong relationship between experience and salary.

Preprocessing Steps

- Removed duplicate rows
- Detected and removed outliers using Z-score
- Applied log transformation to 'Salary' to reduce skewness
- Engineered polynomial features up to degree 2
- Scaled features using *StandardScaler*
- Split data into Train (70%), Validation (15%), and Test (15%) sets

Models Trained

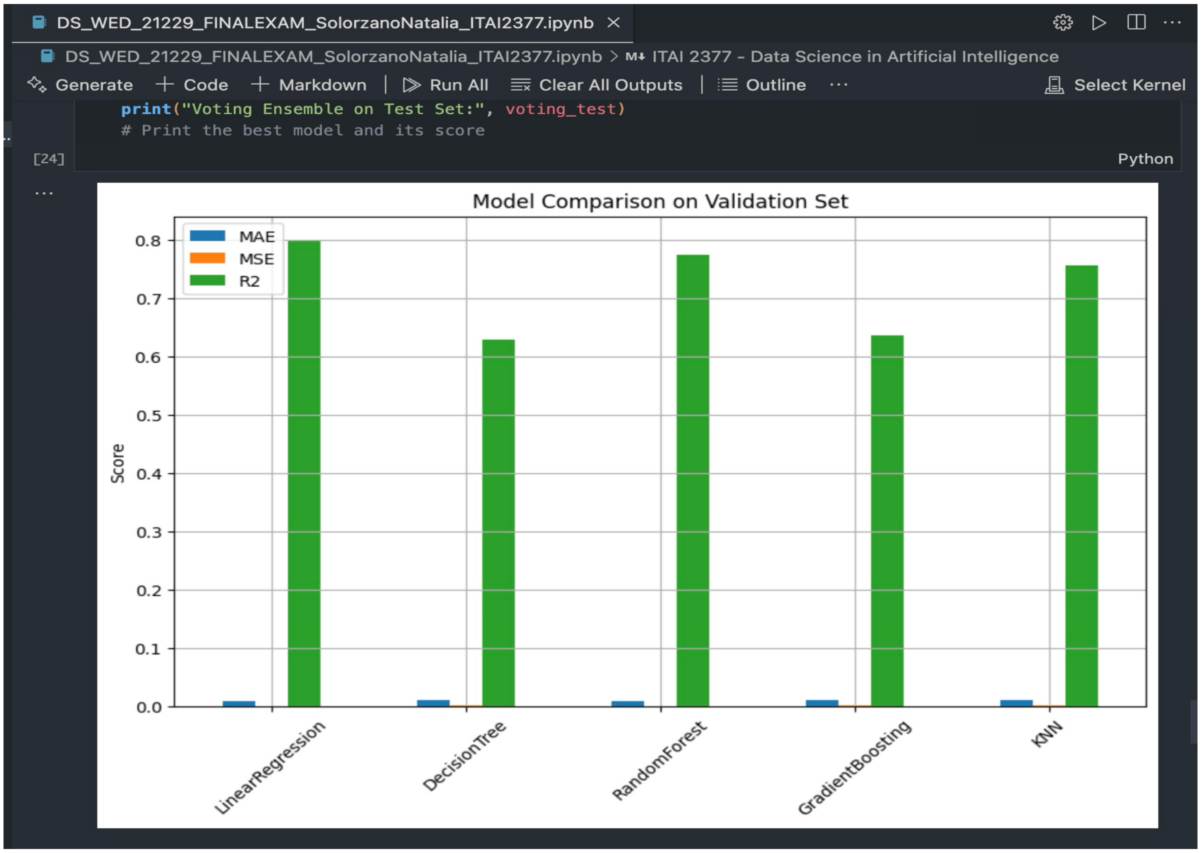
- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- K-Nearest Neighbors Regressor
- Voting Regressor Ensemble (top 3 models)

Evaluation Metrics

Each model was evaluated on the validation set using:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R² Score

A final comparison table was produced including these metrics for both validation and test sets, and the Voting Regressor was also evaluated.



```
print("Voting Ensemble on Test Set:", voting_test)
# Print the best model and its score
```

DS_WED_21229_FINALEXAM_SolorzanoNatalia_ITAI2377.ipynb
DS_WED_21229_FINALEXAM_SolorzanoNatalia_ITAI2377.ipynb > M↓ ITAI 2377 - Data Science in Artificial Intelligence
Generate
+ Code
+ Markdown
Run All
Clear All Outputs
Outline
Select Kernel
[24]
Python
Model Comparison on Validation Set
MAE
MSE
R2
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
LinearRegression
DecisionTree
RandomForest
GradientBoosting
KNN

Best Performing Model

The best-performing individual model was Linear Regression, which achieved the highest R^2 score on the validation set. The ensemble Voting Regressor showed competitive performance and confirmed that the relationship between experience and salary is largely linear.

Student Contribution

All steps from dataset selection, preprocessing, model training, evaluation, visualization, and documentation were completed independently by Natalia Solorzano.

Conclusion

This project demonstrated that even with a small, clean dataset, machine learning can effectively model real-world relationships — in this case, salary prediction based on experience. Linear Regression outperformed more complex models like Gradient Boosting or Random Forest, proving that sometimes simple models work best when the relationship is naturally linear. Through preprocessing (duplicate removal, outlier detection, skewness correction, polynomial expansion), and evaluation (MAE, MSE, R^2), we ensured the models were trained on reliable and normalized data.

Key Takeaway

Predictive modeling doesn't always require big data, it requires clean, relevant data, and an understanding of when to use which model. This kind of ML pipeline can be scaled for HR analytics, salary benchmarking platforms, or integrated into career planning tools for job seekers.

References

- Abhishek Sharma. (n.d.). *Salary Dataset - Simple Linear Regression*. Kaggle. Retrieved April 2025, from https://www.kaggle.com/datasets/abhishek14398/salary-dataset-simple-linear-regression
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
- Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90–95.
- McKinney, W. (2010). *Data structures for statistical computing in Python*. Proceedings of the 9th Python in Science Conference, 51–56. (Pandas)
- Seaborn. (n.d.). *Statistical data visualization*. https://seaborn.pydata.org
- XGBoost Developers. (n.d.). *XGBoost: Scalable and Flexible Gradient Boosting*. https://xgboost.readthedocs.io/
- LightGBM Developers. (n.d.). *LightGBM: A fast, distributed, high-performance gradient boosting framework*. https://lightgbm.readthedocs.io/