Practical work 1

Student: Ej Sobrepeña 2107019

Title of the work: Salary Prediction Using Machine Learning

Used algorithm(s): Linear Regression, Decision Tree, Random Forest

1 DESCRIPTION OF THE WORK

In this project, I have employed machine learning algorithms to predict salaries based on a dataset containing information such as age, gender, education level, job title, years of experience, and salary. My goal is to build and evaluate predictive models to estimate an individual's salary using this data.

2 DATA PREPARATION FOR THE TRAINING

Dataset Description:

The dataset used in this project contains 374 rows and 5 columns. These columns include information about individuals' age, gender, education level, job title, years of experience, and salary.

Screenshot of Training Data:

traning_data - DataFrame

Index	Age	Gender	ducation Levi	Job Title	s of Experi	Salary
0	32	Male	Bachelor's	Software Engineer	5	90000
1	28	Female	Master's	Data Analyst	3	65000
2	45	Male	PhD	Senior Manager	15	150000
3	36	Female	Bachelor's	Sales Associate	7	60000
4	52	Male	Master's	Director	20	200000
5	29	Male	Bachelor's	Marketing Analyst	2	55000
6	42	Female	Master's	Product Manager	12	120000
7	31	Male	Bachelor's	Sales Manager	4	80000
8	26	Female	Bachelor's	Marketing Coordinator	1	45000
9	38	Male	PhD	Senior Scientist	10	110000

To prepare the data for training, I applied the following steps:

- Checked for null values in the data.
- Created dummy variables for categorical features, including Gender, Education Level, and Job Title, to convert them into numerical format for model training.
- Split the dataset into a training set and a test set with an 80:20 ratio.
- Saved a copy of the data specifically for Linear Regression before scaling to ensure correct results in this model.
- Scaled the features using StandardScaler for models other than Linear Regression to ensure that all features have the same scale, making it easier for machine learning models to process.

3 Relevant Metrics for the Cases:

Regression Metrics:

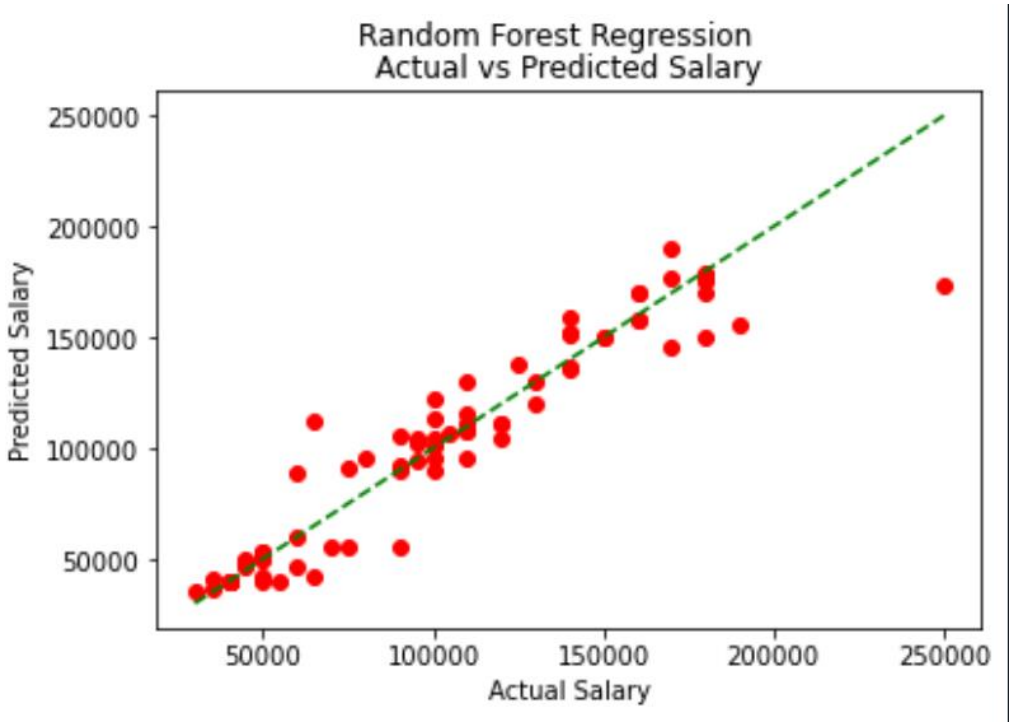
I evaluated the models using the following metrics:

- R-squared (r2): R-squared measures the proportion of the variance in the dependent variable (salary) that is predictable from the independent variables (features). A higher r2 indicates a better fit of the model to the data.
- Mean Absolute Error (MAE): MAE represents the average absolute difference between the predicted and actual salaries. It quantifies the average magnitude of errors.
- Root Mean Squared Error (RMSE): RMSE calculates the square root of the average of the squared differences between predicted and actual salaries. It provides a measure of the typical error in salary predictions.

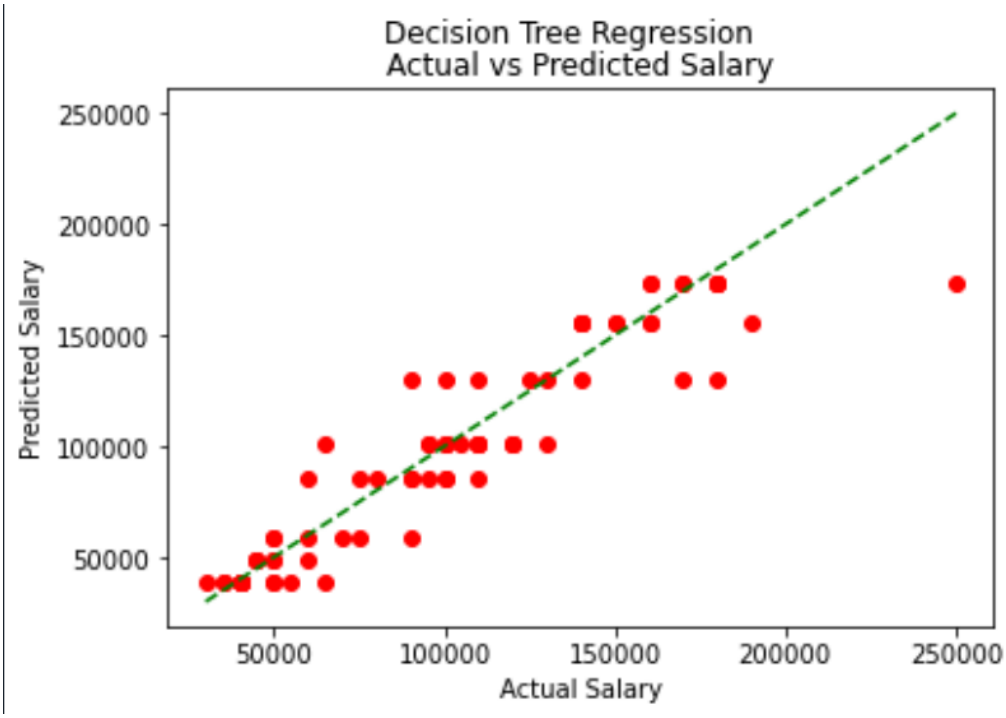
These metrics help us understand how well the models are at predicting salaries. I used them to compare and assess the models.

I also tested each model by using it to predict the salary of a new employee. This employee is a 45-year-old male with 28 years of experience, a Master's degree, and the job title "Director of Marketing." This practical test helps us see how well the models work in real-life situations.

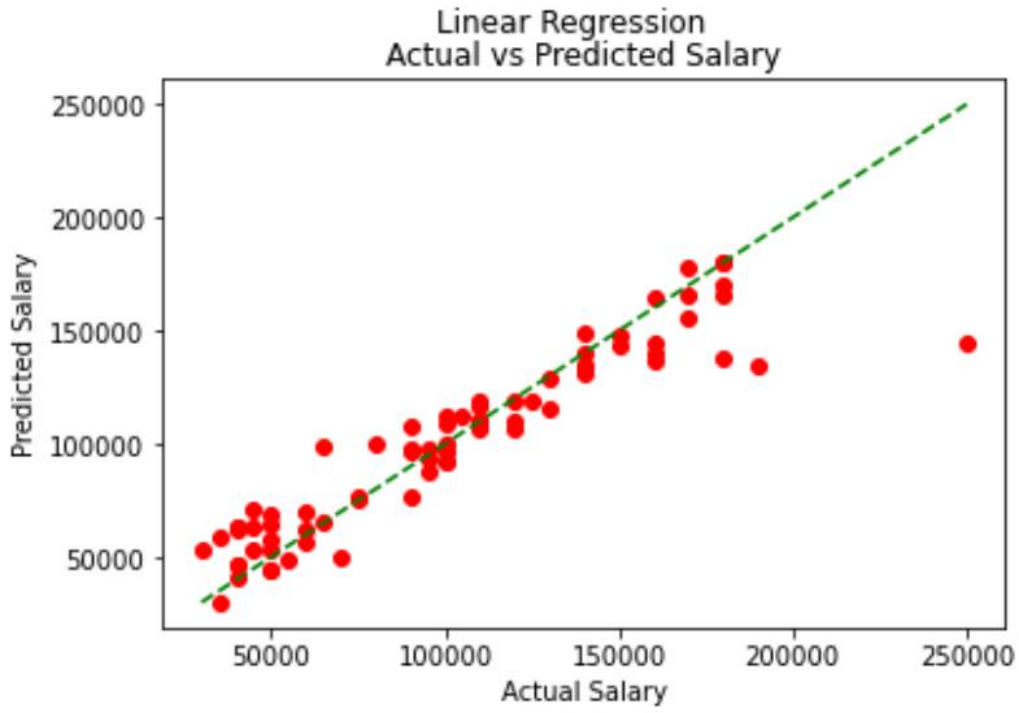
Random Forest Regression Graph:



Decision Tree Regression Graph:



Linear Regression Graph:



4 CONCLUSIONS OF THE RESULTS:

Model Evaluation:

I evaluated three distinct machine learning models: Random Forest Regression, Decision Tree Regression, and Linear Regression. Each model was rigorously scrutinized using essential evaluation metrics.

```
Random Forest Regression:
r2 0.8997113573674219
mae 9790.666666666672
rmse 15506.458869344311
Predicted salary for new employee: 170700.00

Decision Tree Regression:
r2 0.864450593545959
mae 12428.172341336109
rmse 18027.50252106439
Predicted salary for new employee: 129743.59

Linear Regression:
r2 0.852246935672009
mae 11596.559274069365
rmse 18821.53342083603
Predicted salary for new employee: 258733.45
```

Model Comparison:

Among these models, the Random Forest Regression model exhibited the highest r2 score and the lowest MAE and RMSE values. It consistently provided the most accurate salary predictions. The Decision Tree Regression model also displayed competitive results, while Linear Regression, while still respectable, achieved slightly lower performance in comparison.

Usability of the Models:

All models demonstrated decent to good performances, making them potentially usable for salary predictions in real-world applications.

Recommendations for Improvement:

An area for potential improvement is the dataset's size, comprising approximately 370 rows. Moreover, the dataset contains numerous "Job title" categories, but some categories lack sufficient samples. To enhance the models' performance, acquiring a more extensive dataset is recommended. A larger dataset can improve model robustness and predictive accuracy.