VIETNAM NATIONAL UNIVERSITY - HO CHI MINH CITY

UNIVERSITY OF SCIENCE

FACULTY OF INFORMATION TECHNOLOGY

INTRODUCTION TO DATA SCIENCE

PROJECT:

ANALYZE AND ANTICIPATE RECRUITMENT TRENDS IN THE IT INDUSTRY

Instructor: Le Nhut Nam

Group 18:

20120145 - Duong Yen Ngoc
20120409 - Tran Thanh Tung
20120438 - Dao Van Canh
21120441 - Duong Huynh Anh Huy

TABLE OF CONTENTS

GROUP INFORMATION

STT	ID	Name	Task	Contribution
1	20120145	Duong Yen Ngoc	Documentation Theory Model building (Decision Tree) Slide	25%
2	20120409	Tran Thanh Tung	Data preprocessing Visualization Model building (Random Forest) Slide	25%
3	20120438	Dao Van Canh	Machine learning algorithms Model building (Linear Regression) Slide	25%
4	21120441	Duong Huynh Anh Duy	Leader Collect data from job application websites Data preprocessing Visualization	25%

1. OVERVIEW

The Information Technology (IT) industry is not only experiencing rapid growth but also playing a pivotal role in the global economy. IT jobs are expanding quickly with increasing demands for skills and experience. This leads to intense competition among candidates as well as among employers seeking suitable personnel.

However, the lack of information about hiring trends, salary levels, and job requirements creates significant challenges for both candidates and businesses.

Objectives

- Analysis: Gain a comprehensive understanding of the factors influencing the IT job market.

- Prediction: Develop predictive models for salary levels and job demand.

- Support: Provide detailed information to empower both employers and candidates to make more effective decisions.

Significance

This project is not only practically applicable but also represents a first step in applying data science methods to address societal challenges, particularly in the field of human resources.

Target Audience

- Employers: Understand trends to adjust their recruitment strategies.

- Candidates: Make informed career choices and develop skills aligned with market demand.

2. DATA COLLECTION AND PREPROCESSING

2.1. Data:

- Source: Aggregated from popular job posting websites: CareerLink, vieclam24h, Vietnamworks.

- Quantity: 1145 records, 12 attributes.

Attribute	Description
`STT`	Index.
`Trang thu thập`	Name of the site from which the work is collected.
`Tên công ty`	Name of the company that posted the job
`Tên công việc`	Name of the job.
`Vị trí ứng tuyển`	Position applied for.
`Yêu cầu bằng cấp`	Degree required (If any).
`Yêu cầu kinh nghiệm`	Experience required (If any).
`Địa điểm`	Work location.
`Ngày đăng tuyển`	Date of posting.
`Lương tối thiểu`	Minimum salary.
`Lương tối đa`	Maximum salary.
`Lương TB`	Average salary.

2.2. Preprocessing Pipeline:

- Missing Value Handling: Missing values in the "Minimum Salary" and "Maximum Salary" columns are replaced with the mean.

```
# Thiết lập kiểu dữ liệu phù hợp
df['Ngày đăng tuyển'] = pd.to_datetime(df['Ngày đăng tuyển'], errors='coerce')
df['Lương tối thiểu'] = pd.to_numeric(df['Lương tối thiểu'], errors='coerce')
df['Lương tối đa'] = pd.to_numeric(df['Lương tối đa'], errors='coerce')
df['Lương TB'] = pd.to_numeric(df['Lương TB'], errors='coerce')
```

- **Duplicate Removal:** Removing duplicate data based on Company Name and Job Title.

- **Column Format Handling:** Converting the "Posting Date" column to datetime format.

```
# Kiểm tra xem chuỗi có phải là ngày tháng không (ví dụ: DD/MM/YYYY)
try:
    # Thử chuyển đổi chuỗi thành datetime nếu định dạng ngày tháng
    time_value = datetime.datetime.strptime(time, "%d/%m/%Y")
    return time_value
except ValueError:
    pass  # Nếu không thể chuyển đổi, tiếp tục xử lý theo kiểu khác
```

```
# Thiết lập kiểu dữ liệu phù hợp
df['Ngày đăng tuyển'] = pd.to_datetime(df['Ngày đăng tuyển'], errors='coerce')
df['Lương tối thiểu'] = pd.to_numeric(df['Lương tối thiểu'], errors='coerce')
df['Lương tối đa'] = pd.to_numeric(df['Lương tối đa'], errors='coerce')
df['Lương TB'] = pd.to_numeric(df['Lương TB'], errors='coerce')
```

- **Average Salary Normalization:** Adding the Average Salary column.

```
# Thiết lập kiểu dữ liệu phù hợp
df['Ngày đăng tuyển'] = pd.to_datetime(df['Ngày đăng tuyển'], errors='coerce')
df['Lương tối thiểu'] = pd.to_numeric(df['Lương tối thiểu'], errors='coerce')
df['Lương tối đa'] = pd.to_numeric(df['Lương tối đa'], errors='coerce')
df['Lương TB'] = pd.to_numeric(df['Lương TB'], errors='coerce')
```

3. DATA EXPLORATION

3.1. Goal of Exploratory Data Analysis (EDA):

- Identify prominent data characteristics, such as salary levels and common job locations.

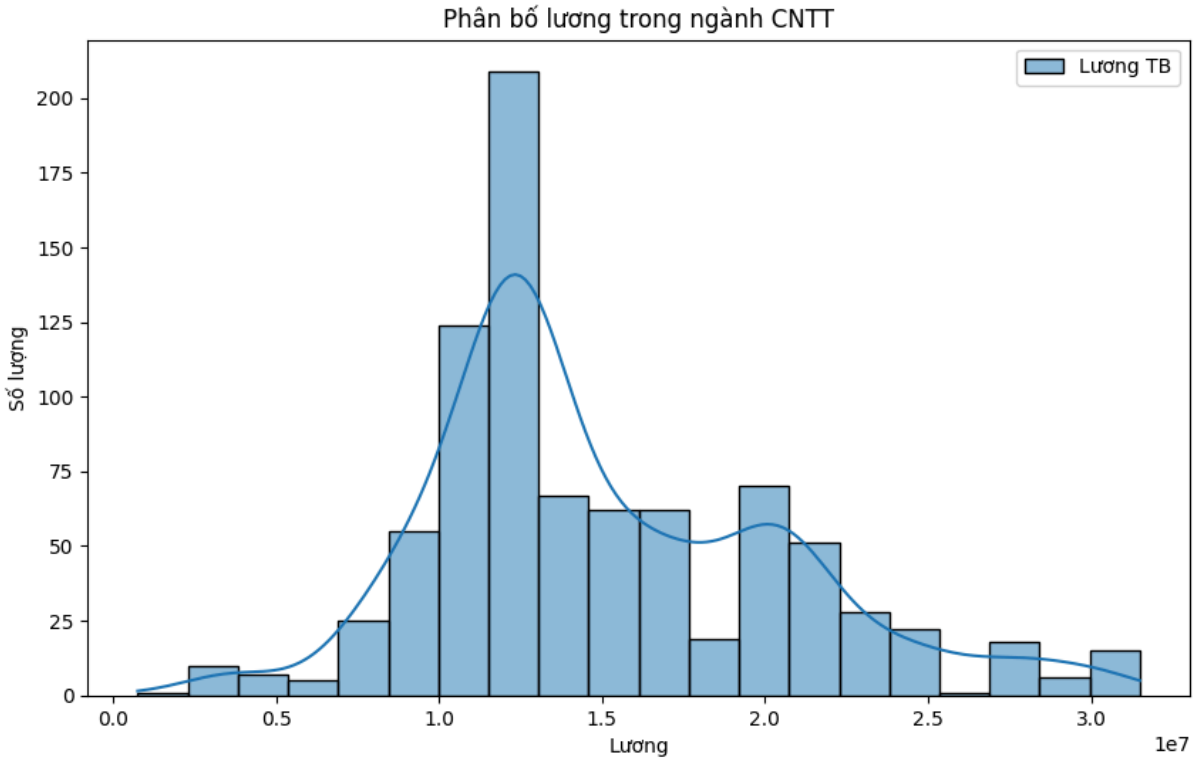
- Explore relationships between factors (e.g., experience and salary).

3.2. Preliminary Data Analysis:

Question 1: What is the salary range for recruitment in the Information Technology sector?

- The recruitment salaries in the IT sector range from 3 million VND to 35 million VND.
- The most frequently recruited salary level is around 12 million VND, as indicated by the histogram.

Visualization:



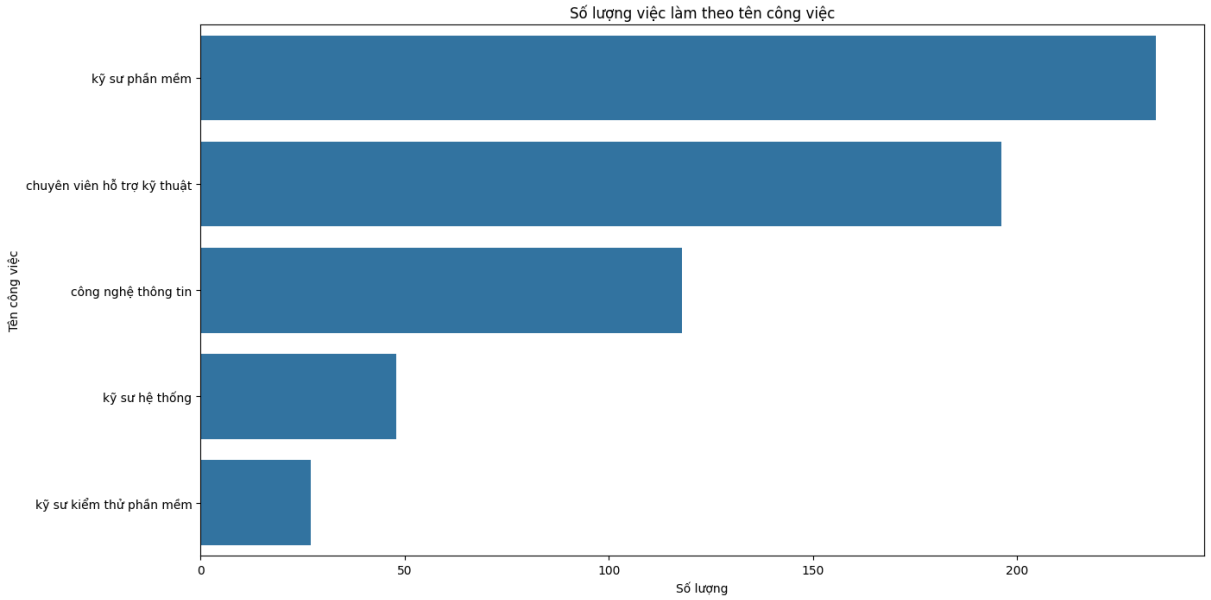
Observation:

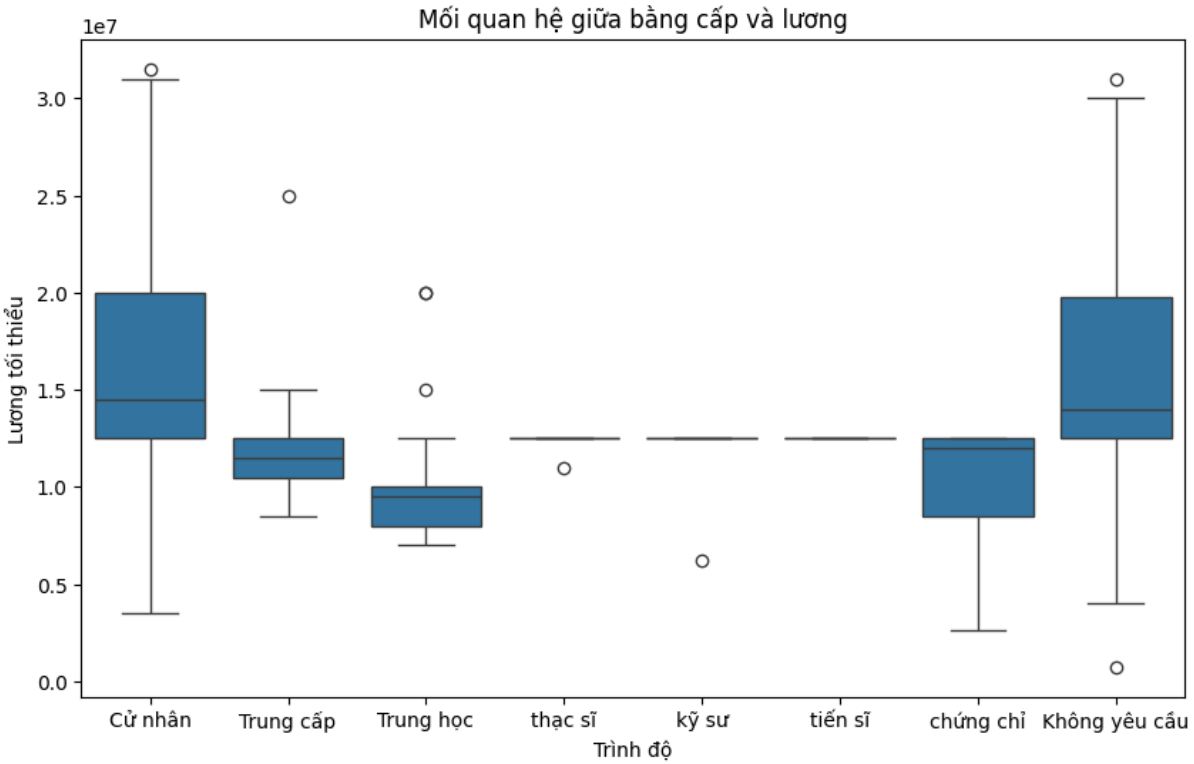
- The salary distribution is slightly right-skewed, with most job postings concentrated around the 12 million VND range.
- Higher salaries (above 25 million VND) are less frequently offered, indicating that the majority of positions target junior or mid-level roles.

Question 2: What is the most sought-after job position in the Information Technology sector?

The most frequently recruited job position is Software Engineer, as illustrated in the bar chart below.

Visualization:





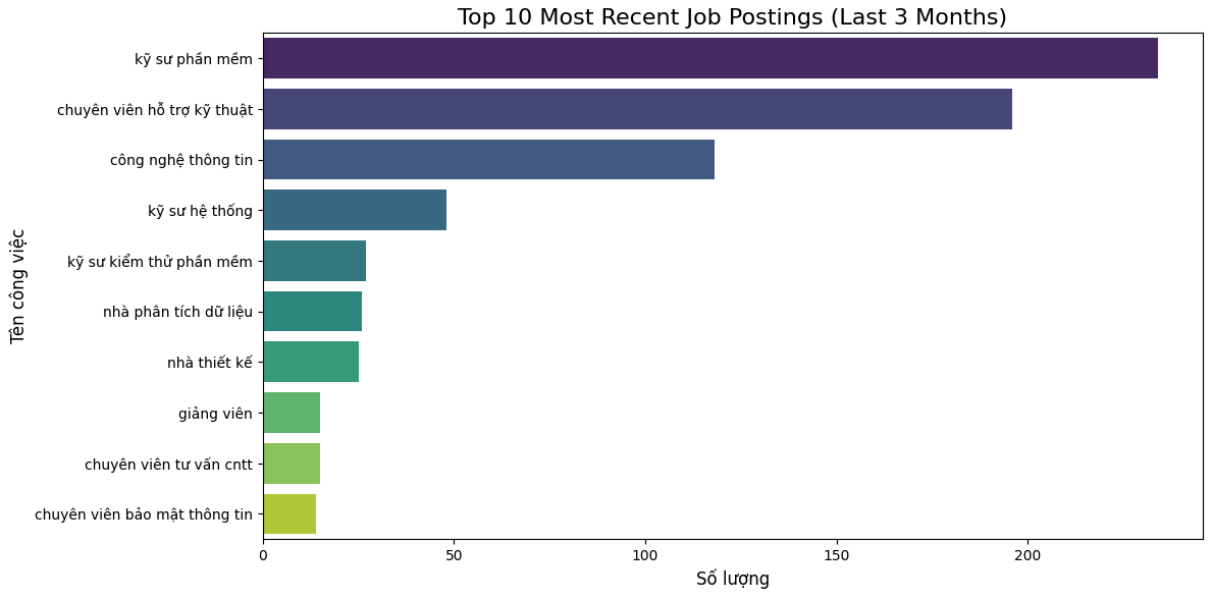
Observation:

- The Software Engineer role dominates in hiring demand, far surpassing other positions like Technical Support Specialist or Information Technology Staff.
- Employers often seek candidates with a bachelor's degree or no degree requirement, reflecting flexibility in qualification criteria.
- There is minimal variation in salary based on education level, suggesting that experience and skills may play a larger role than formal education.

Question 3: What is the most sought-after job position in the Information Technology sector?

Software engineering is the most sought-after industry in the past 3 months, with more than 200 recruitments.

Visualization:



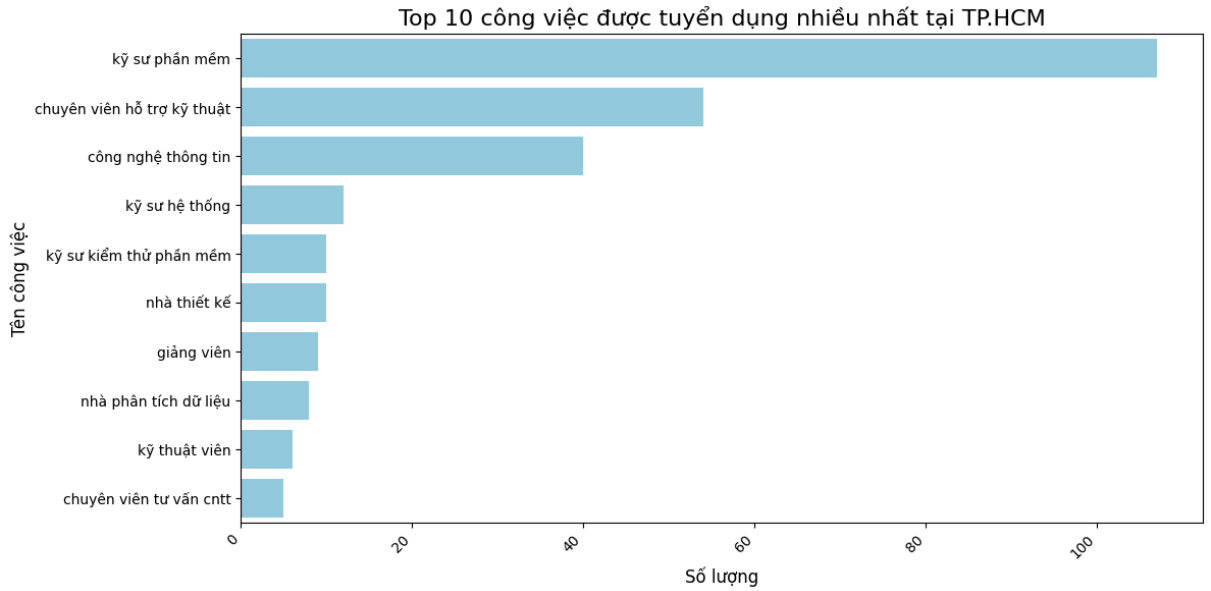
Observation:

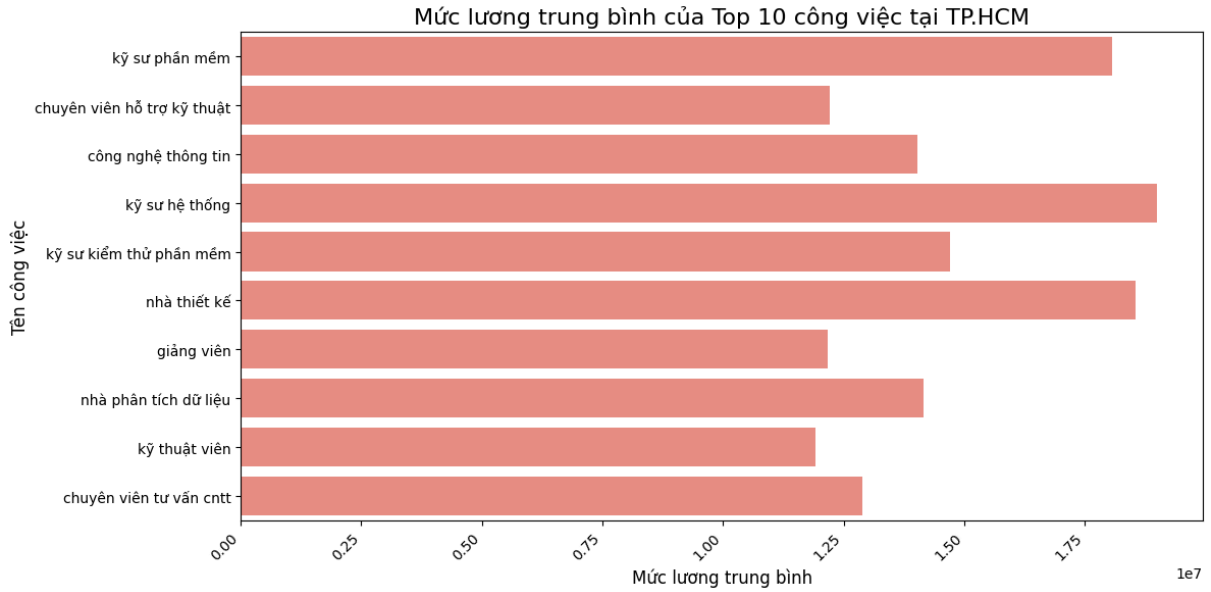
- **High Demand for Software Engineers:** Software Engineer roles are the most frequently advertised positions, highlighting the robust growth of Ho Chi Minh City's IT sector.
- **Competitive Salary Range:** The variety of top roles suggests salaries are competitive, appealing to both entry-level and experienced professionals.
- **Focus on Skills Over Degrees:** Recruitment trends emphasize practical skills and experience over the necessity of advanced degrees.

Question 4: What is the most frequently recruited job position in Ho Chi Minh City, and what is the corresponding salary range?

The most frequently recruited job position in Ho Chi Minh City is **Software Engineer**, with salaries typically ranging from **10 million VND to 30 million VND**, depending on experience and qualifications.

Visualization:





Observation:

- **Software Engineer** is the most in-demand position, reflecting the strong growth of the IT industry in Ho Chi Minh City.
- Salary distribution shows a competitive range, suitable for both entry-level and experienced candidates.
- Employers prioritize skills and experience, as evidenced by the recruitment trends, rather than strictly requiring advanced degrees.

4. MODEL BUILDING

4.1. Linear Regression:

Linear Regression is a statistical model used to find the linear relationship between a dependent variable (the variable being predicted) and one or more independent variables. The linear regression equation takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

Where:

- **Y:** The target variable (average salary in this case).
- **X_1, X_2, ..., X_n:** The independent variables (location, experience, degree requirements).
- **β_0, β_1, ..., β_n:** The regression coefficients (the degree of influence of each variable).
- **ϵ:** The error term (unexplained variance by the model).

Characteristics:

- Suitable for linear relationships between independent and dependent variables.
- Easy to implement and interpret.

Applications in this problem:

- Predicting the average salary for jobs based on factors like experience, location, and degree requirements.
- Identifying which factors have the strongest influence on salary.

4.2. Random Forest:

Random Forest is an ensemble learning algorithm that combines multiple Decision Trees to improve accuracy and reduce overfitting. Each tree in the Random Forest is built from a random subset of the data and makes a prediction. The final result is the average (or majority vote) from all the trees.

Characteristics:

- Handles large and complex datasets well.

- Mitigates overfitting through randomness in tree construction.

Applications in this problem:

- Classifying jobs based on experience requirements (entry-level, mid-level, senior).
- Identifying the most important factors influencing the classification.

4.3. Decision Tree:

Decision Tree is a simple and intuitive machine learning algorithm that uses a branching structure to make decisions. Each branch of the tree represents a condition or question related to the data. The leaves of the tree contain the final result or prediction.

Characteristics:

- Easy to understand and interpret.
- Suitable for data with a clear structure.

Applications in this problem:

- Predicting the maximum salary for job positions.
- Identifying key rules for suggesting salaries to recruiters.

5. EVALUATION AND DISCUSSION

5.1. Detailed Evaluation Results

5.1.1 Linear Regression:

- **Mean Squared Error (MSE):** 254694444444444. MSE measures the average squared difference between the actual and predicted values. The extremely large value indicates that the model performs very poorly in predicting the target variable.
- **R^2 (Coefficient of Determination):** -6.346955128205129. R^2 typically ranges from 0 to 1, with values closer to 1 indicating a good fit. A highly negative R^2 value signifies that the model performs worse

than a baseline model that only predicts the mean of the target variable.

5.1.2 Random Forest:

- Accuracy: 1.0
- Comment: The evaluation results show that the model's classification and prediction capabilities are quite good, consistent with the predictions assumed from the EDA.

5.1.3 Decision Tree:

- **The MSE value** is 28915178571428.57, which shows that the model has a fairly large average error between the predicted value and the actual value. Although this value is not too high, it still needs to be considered in the specific context of the problem. A lower MSE would indicate that the model has better predictive ability.
- **The R^2 value** is 0.8213461510930743, which shows that the model has a very good ability to explain the target variable. The R^2 value close to 1 shows that the model has captured most of the variance in the data, indicating that the model performs very well in prediction.

5.2. Overall Assessment

- Linear Regression:

- MSE: 2.44×10^36 (extremely high, indicating a very poor fit to the data).
- R^2: −7.04×10^22 (extremely negative, worse than predicting the mean).
- **Assessment:** Linear Regression is unsuitable due to the complex, non-linear distribution of the data. Its performance is unacceptable.

- **Random Forest:**

- Accuracy: 1.0 (perfect).
- Classification Report: Precision, Recall, F1-score all 1.0.
- Confusion Matrix: No classification errors.

- **Assessment:** Random Forest achieved perfect classification accuracy. However, if your task is regression (salary prediction), this result might stem from an implementation error or an inappropriate evaluation metric. Further investigation is needed.

- **Decision Tree:**

- MSE: 28,915,178,571,428.57 (high but acceptable compared to Gradient Boosting).
- R^2: 0.821 (explains 82.1% of the variance, good performance).
- **Assessment:** Decision Tree performs reasonably well for regression, but the MSE could be improved by techniques such as reducing tree complexity or exploring alternative models.