جامعة الملك سعود
King Saud University
1957

Data Science Salaries prediction

presented by: Group 4

problem

Data

pre-processing

Data Mining Techniques

Findings

Problem

Salary is a crucial component for employees, job hunters, and organizations, especially in the field of data science, which is in constant growth,

understanding the factors that affect data science job salaries could be a way to help individuals and businesses keep track of the market trends, making informed choices. It also allows organizations to set up fair and competitive salary ranges, as well as support discussions while hiring employees for a particular job.

Data

we applied our data mining tasks on the dataset of:

NUMBER OF ATTREBUTES:11

TYPE OF ATTRIBUTES: NOMINAL, ORDINAL, NUMERIC

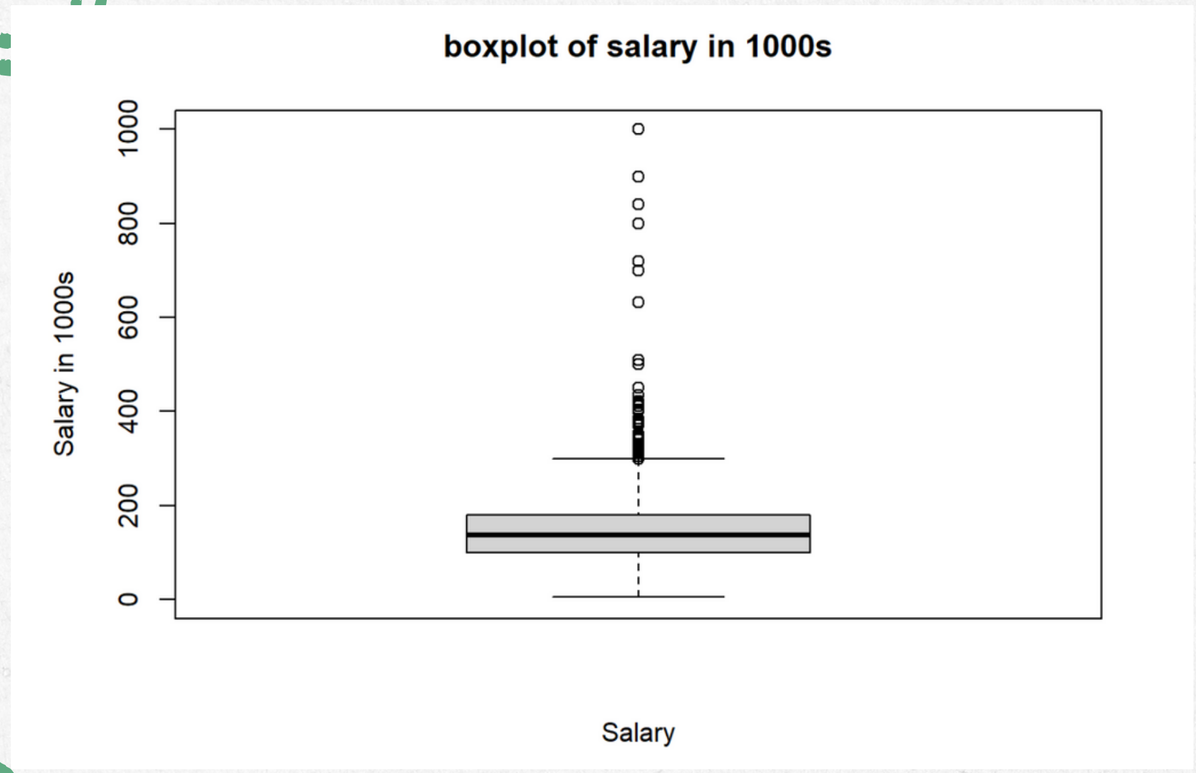
NUMBER OF OBJECTS: 3755

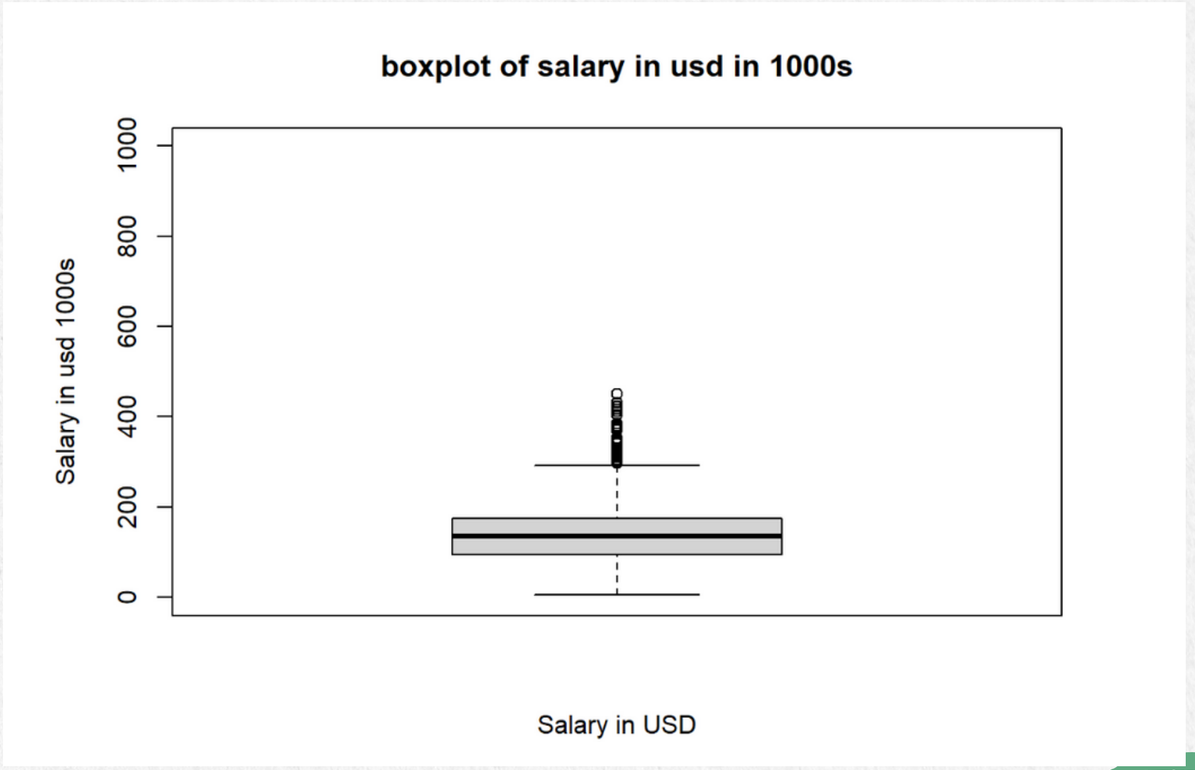
CLASS LABEL: SALARY IN USD

1. work_yeae
2. experience_level
3. employment_type
4. job_title
5. salary
6. salary_currency
7. salary_in_usd
8. employee_residence
9. remote_ratio
10. company_location
11. company_size

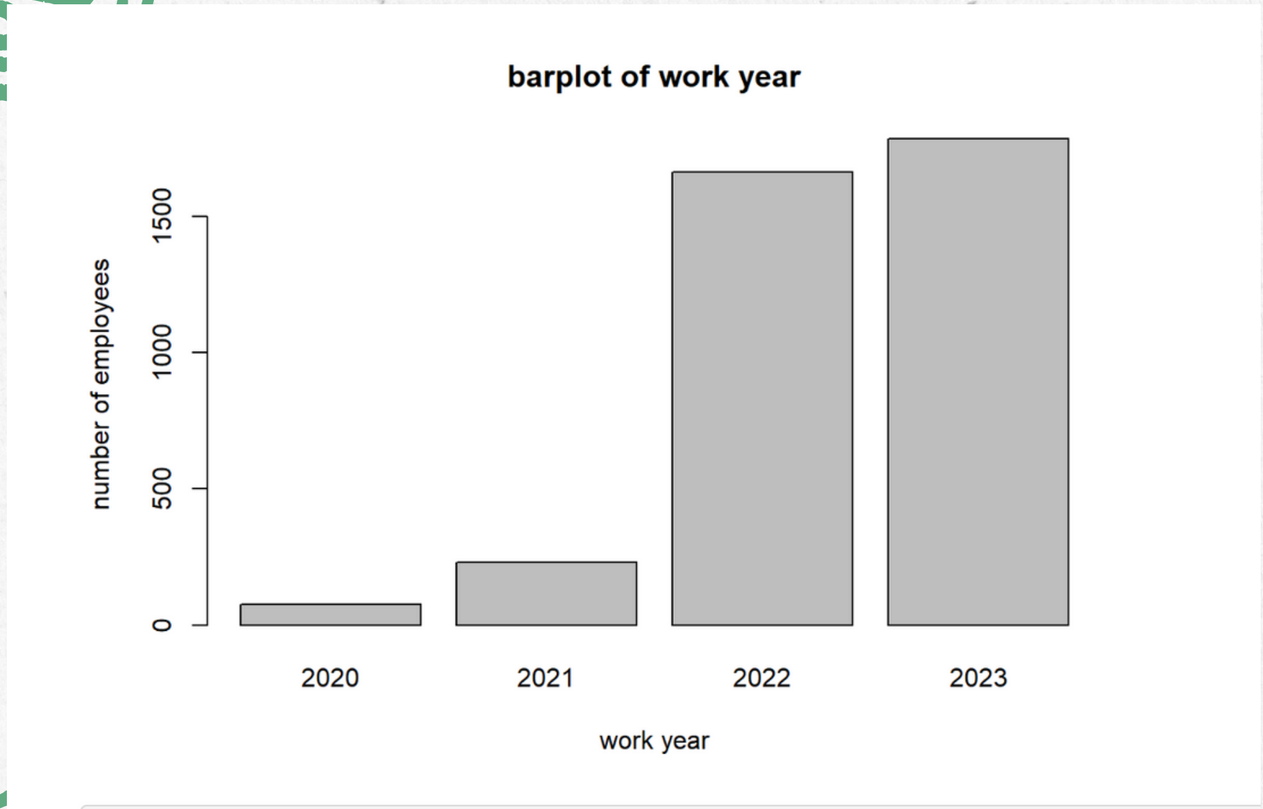
source Dataset:
https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023?resource=download

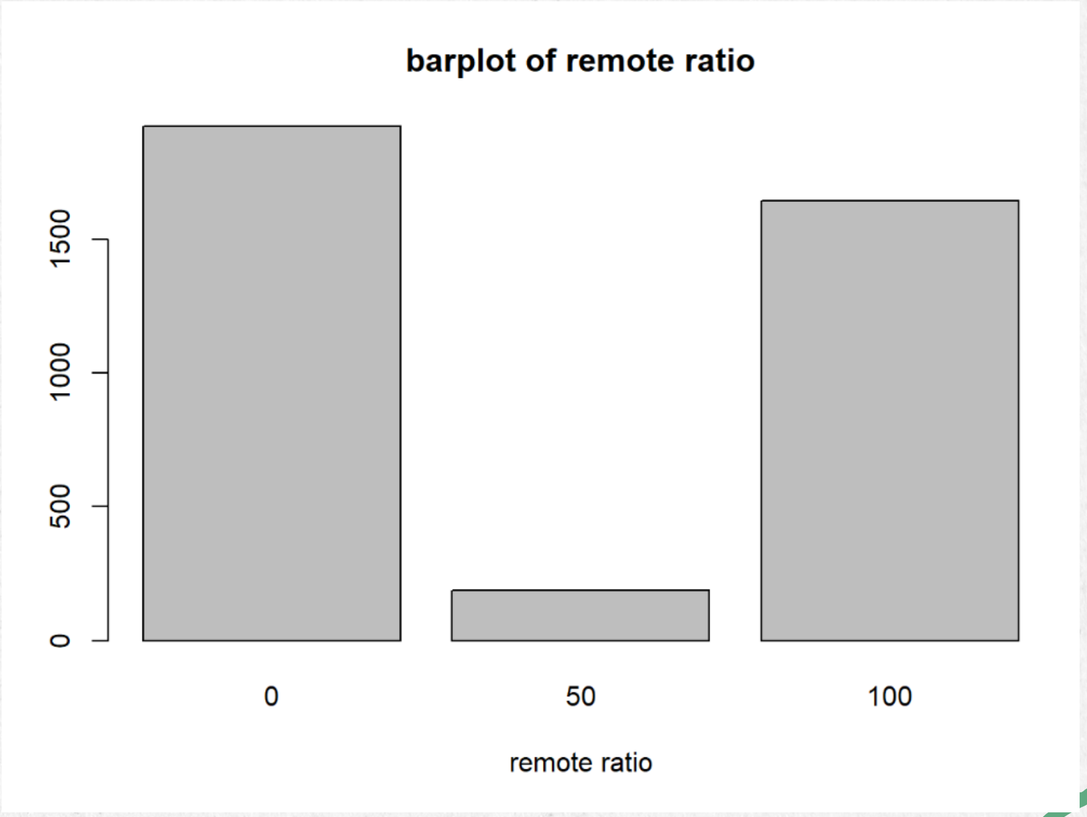
Graphs


boxplot of salary in 1000s
Salary in 1000s
0
200
400
600
800
1000
Salary

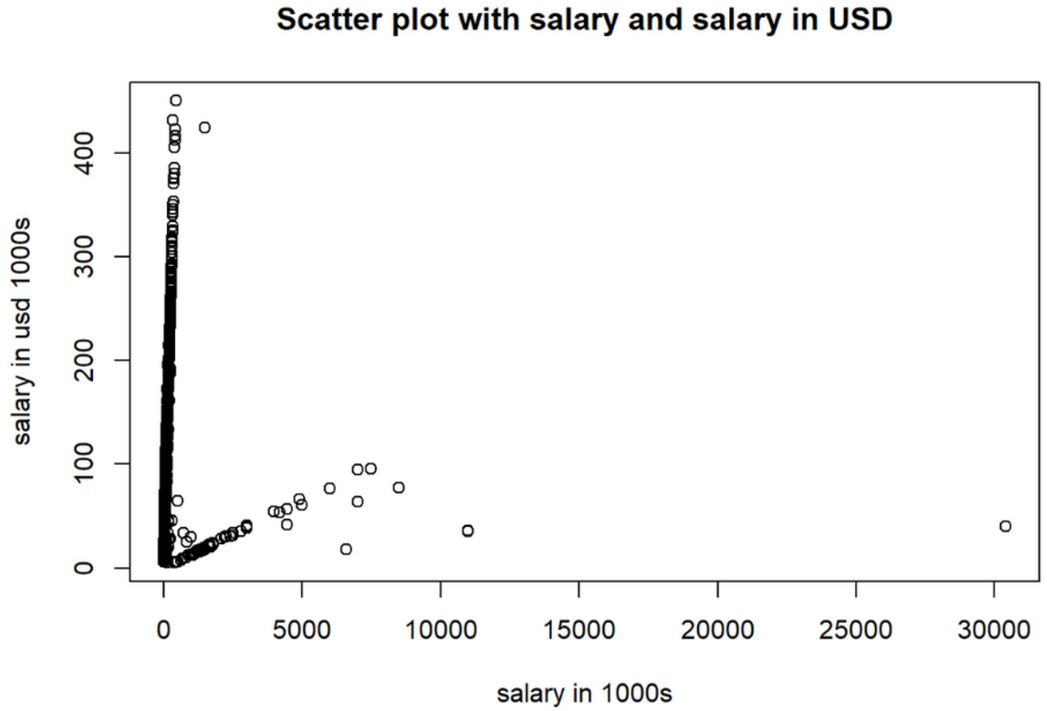

boxplot of salary in usd in 1000s
Salary in usd 1000s
0
200
400
600
800
1000
Salary in USD

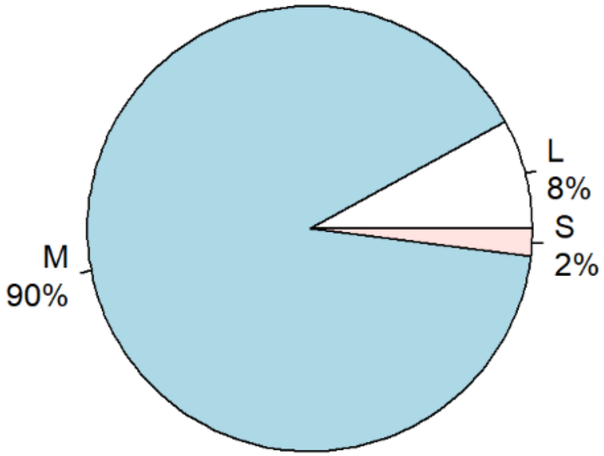
Graphs


barplot of work year
number of employees
0
500
1000
1500
2020
2021
2022
2023
work year


barplot of remote ratio
0
500
1000
1500
0
50
100
remote ratio

Graphs


Scatter plot with salary and salary in USD
salary in usd 1000s
0
100
200
300
400
0
5000
10000
15000
20000
25000
30000
salary in 1000s


L
8%
S
2%
M
90%

pre-processing

Data preprocessing is an essential step in the data analysis process. It involves cleaning and transforming raw data into a format suitable for analysis

data cleaning

data transformation

preprocessing

removing the outliers

- Outliers can introduce bias and skewness in the data.
- we detected and removed the outliers using the interquartile range (IQR) method.

Normalization

- normalize() function was applied to the salary and salary_in_usd attributes. Normalization ensures that these variables are on a similar scale

Discretization

- the cut() function was utilized to discretize the salary_in_usd attribute into three categories: “low,” “mid,” and “high.” Discretization simplifies the analysis and interpretation of the data.

Encoding

- we encoded our ordinal and nominal variables using factor() function
- many algorithms and statistical techniques require numerical inputs

categorizing

- The original dataset had 93 different job title categories,
- The dataset had 72 different company location categories
- which could make analysis and grouping challenging

Removing irrelevant and duplicate attributes

to avoid redundancy, we removerd redundent columns

Data Mining Techniques

Classification

- Classification is supervised learning which mean is need a class lapel to classify the objects.
- We trained our models to be able to predict if the salary is (low, mid, or high) using (salary in usd) class labe.

Clustering

- Clustering is unsupervised learning it will group objects in cluster based on similarity and dissimilarity.
- Our model will create a set of clusters for the employes who have similar characteristics, then these clusters will be used to predict.

Classification

we tried three different split ratios each divided into two sets:

- Training dataset: Used for building the decision tree.
- Testing dataset: Used to evaluate the constructed model.

1	2	3
Training 70% Testing 30%	Training 80% Testing 20%	Training 90% Testing 10%

Classification

to buld our model we used three decision tree algorithms:

- Gini index
- Gain ratio
- Information gain

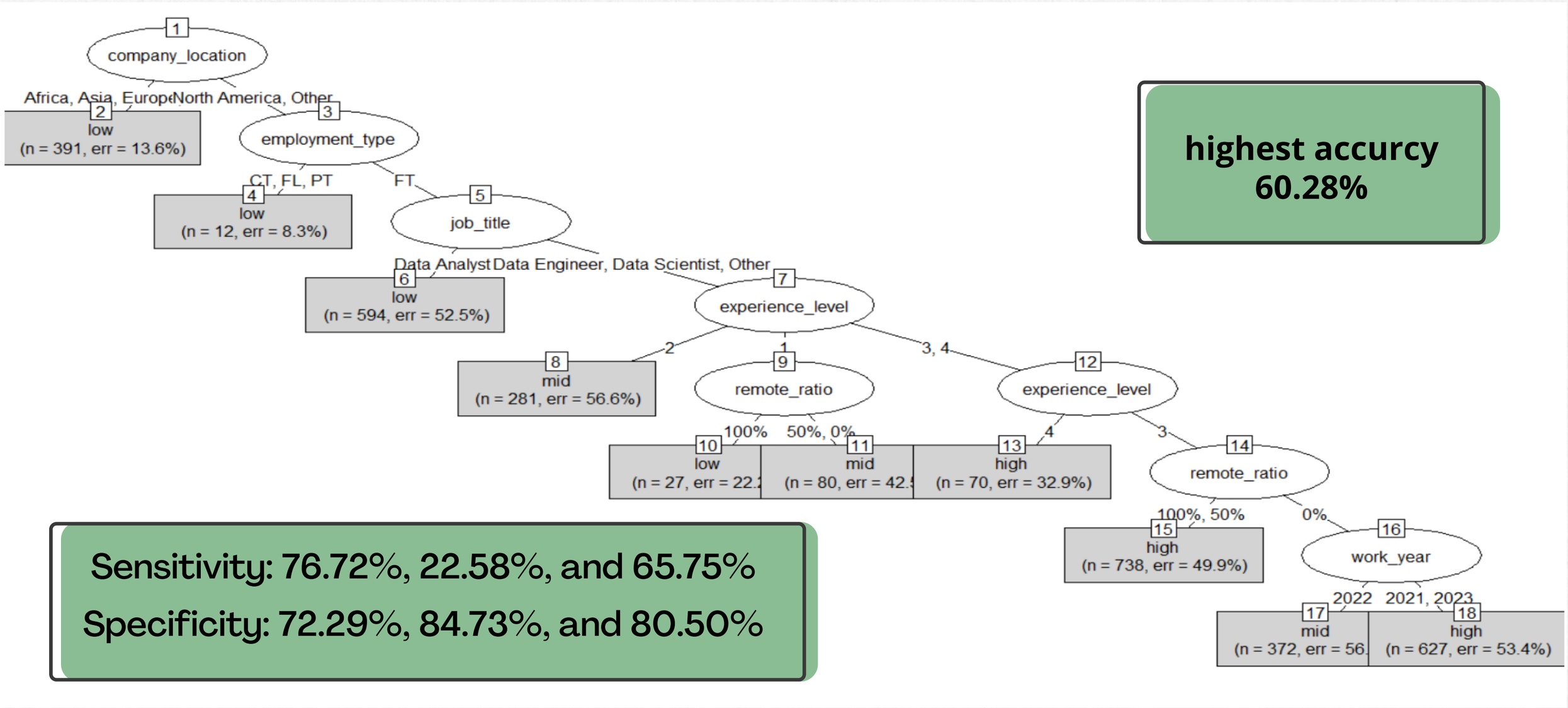
Classification Findings

The accuracy of the model in predicting employee salaries ranges from 53.24% to 60.28%. The gain ratio model consistently outperforms other models, such as Gini Index and Information Gain.

90% training and 10% testing split generally yields better performance in terms of sensitivity and specificity compared to other splits (70-30 and 80-20).

The decision tree provides insights into how the model makes predictions based on work year, experience level, job title, and company location. Based on its superior performance, it is recommended to use the gain ratio model with a 90-10 split for accurate salary predictions.

Classification Findings



highest accurcy 60.28%

Sensitivity: 76.72%, 22.58%, and 65.75%

Specificity: 72.29%, 84.73%, and 80.50%

Clustering

Clustering Methods

we used K-means algorithm, generates K clusters by determining the center point for each cluster and then assigning each object to the cluster with the closest center point.

Clustering evaluation

we evaluate the results using clustering evaluation such as: Elbow method, Silhouette Coefficient, BCubed precision and recall metrics, Total within-cluster sum of squares.

Clustering

we tried three different k's (number of clusters):

1

k=2

- Based on observation (clearly clustering)

2

k=3

- Based on the results of Elbow method

3

k=10

- Based on the results of silhouette coefficient method

Clustering Findings

the best cluster k=2
k=2 successfully captures a comprehensive representation of items from the same category within their clusters.

The usage of the optimal k value, specifically set at 10, results in undesirable overlapping clusters.

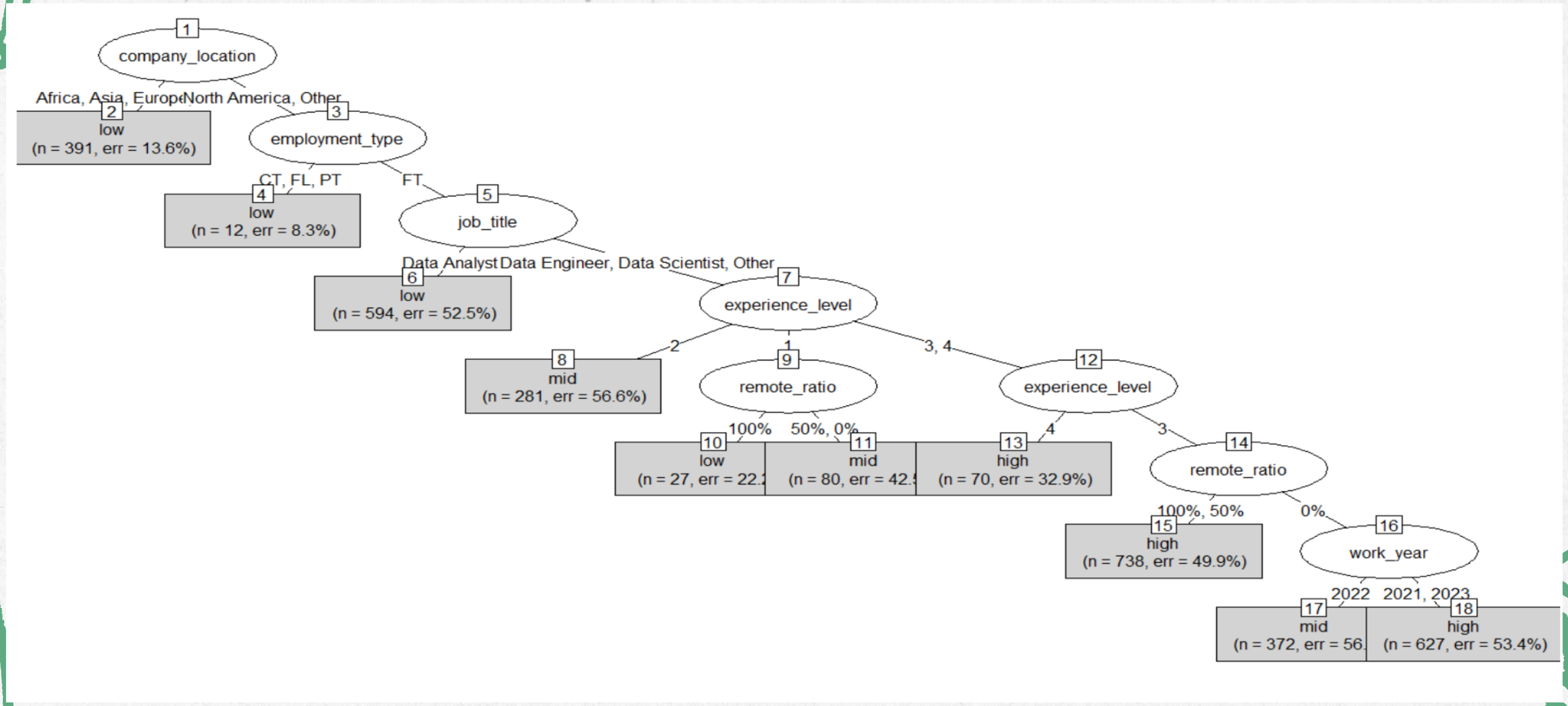
k-3 considered as a reasonable alternative as it strikes balance between precision and recall.

Clustering Findings

less overlap between clusters

highest Recall value

the final tree choosen structure



Thank you for listening!

- Deem Alshaye 443200583
- Norah mohammed Alwohaibi 443200753
- Jana Aljomaih 443200860
- Khloud Mohammed Al-doayan 443201002