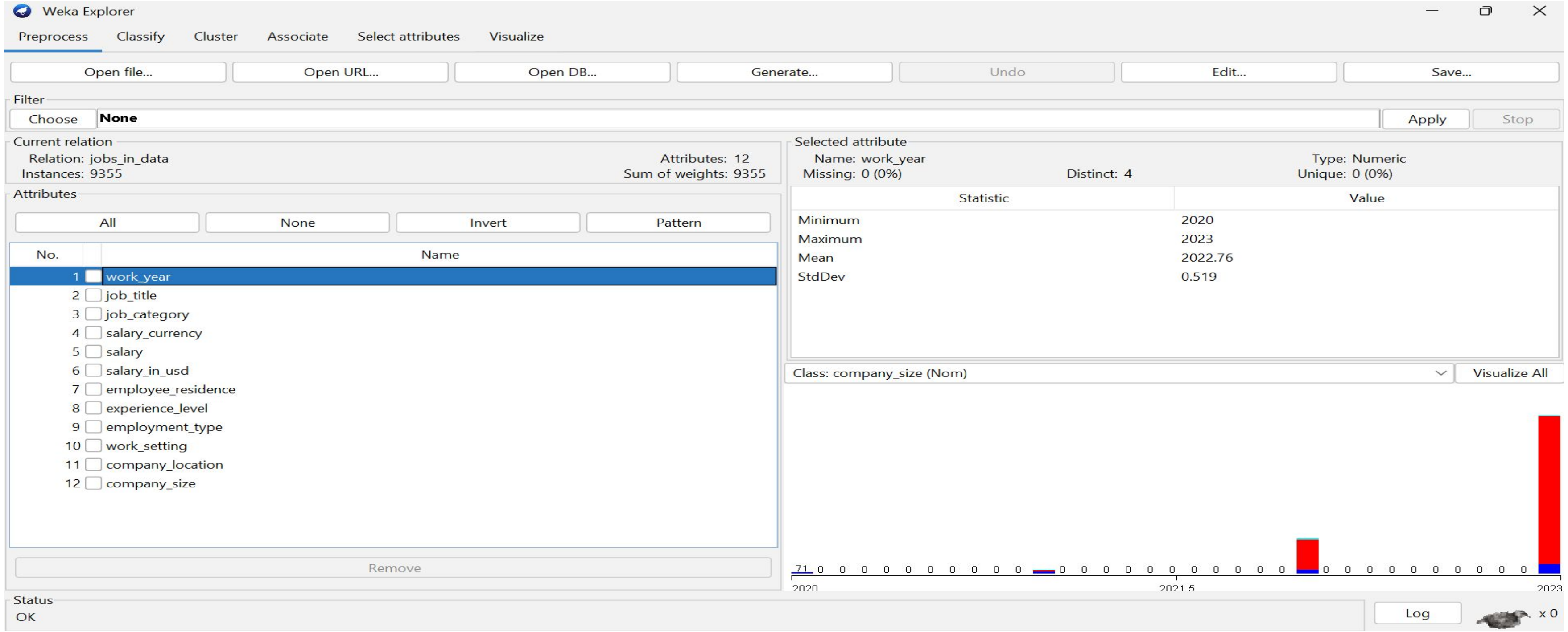
HW4 - Classification

Team Members: Puja Shah & Sanjida Chowdhury

Dataset: “Jobs and Salaries in Data Science”



Filter

None

Current relation

Relation: jobs_in_data

Instances: 9355

Attributes: 12

Sum of weights: 9355

Attributes

No.	Name
1	work_year
2	job_title
3	job_category
4	salary_currency
5	salary
6	salary_in_usd
7	employee_residence
8	experience_level
9	employment_type
10	work_setting
11	company_location
12	company_size

Selected attribute

Name: work_year

Type: Numeric

Missing: 0 (0%)

Distinct: 4

Unique: 0 (0%)

Statistic	Value
Minimum	2020
Maximum	2023
Mean	2022.76
StdDev	0.519

Class: company_size (Nom)

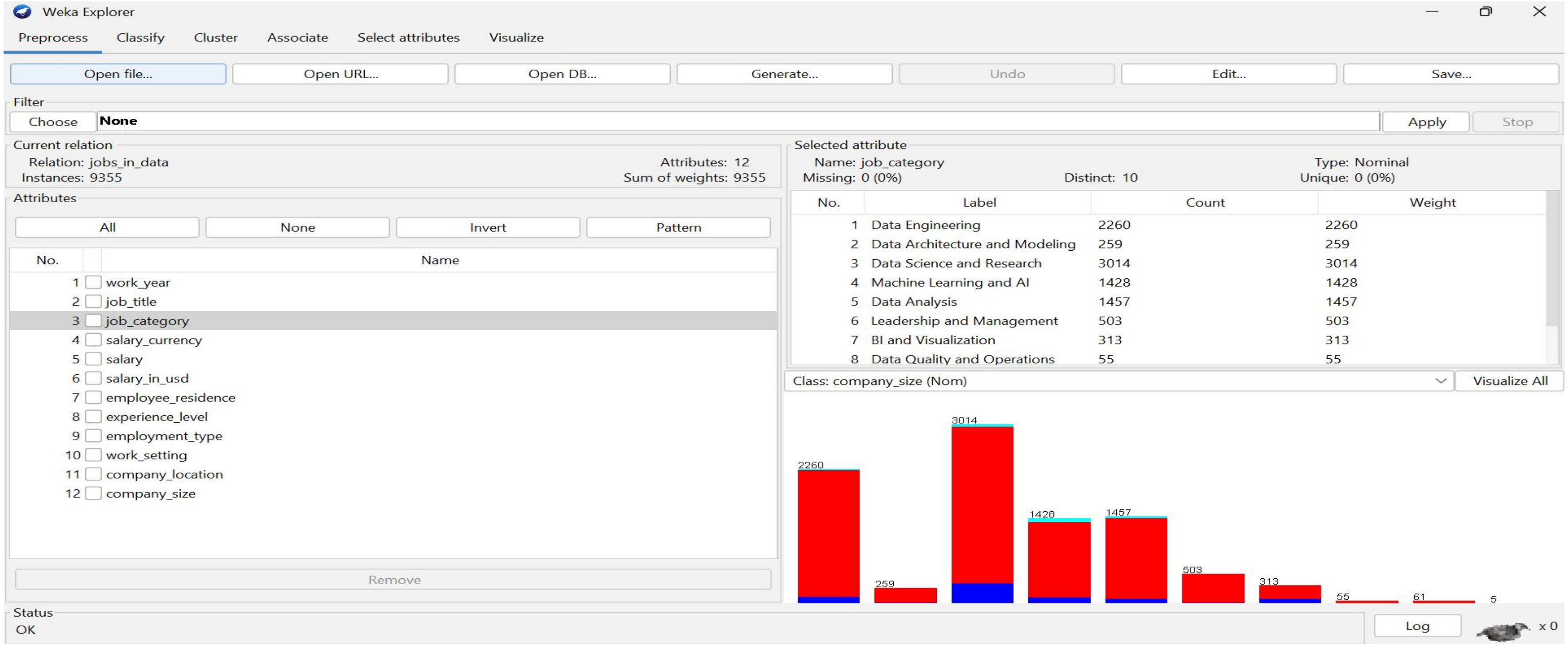
Status

OK

Dataset: “Jobs and Salaries in Data Science”

- The dataset used for this analysis contains information related to job postings, including attributes such as work year, job title, job category, salary, experience level, and more. It comprises a total of 9355 instances, each representing a unique job posting. The dataset attributes provide valuable insights into the characteristics of different job opportunities, such as the type of role, salary range, and required experience level. This information can be valuable for job seekers, employers, and researchers interested in understanding trends and patterns in the job market.
- The “Jobs and Salaries in Data Science” dataset sourced from Kaggle (same data from the pre-processing). The link to the data set is here: https://www.kaggle.com/datasets/hummaamqaasim/jobs-in-data

Distribution of the Target Class



Filter: None

Current relation

Relation: jobs_in_data
Instances: 9355
Attributes: 12
Sum of weights: 9355

Attributes

No.		Name
1	□	work_year
2	□	job_title
3	□	job_category
4	□	salary_currency
5	□	salary
6	□	salary_in_usd
7	□	employee_residence
8	□	experience_level
9	□	employment_type
10	□	work_setting
11	□	company_location
12	□	company_size

Selected attribute

Name: job_category
Type: Nominal
Missing: 0 (0%)
Distinct: 10
Unique: 0 (0%)

No.	Label	Count	Weight
1	Data Engineering	2260	2260
2	Data Architecture and Modeling	259	259
3	Data Science and Research	3014	3014
4	Machine Learning and AI	1428	1428
5	Data Analysis	1457	1457
6	Leadership and Management	503	503
7	BI and Visualization	313	313
8	Data Quality and Operations	55	55

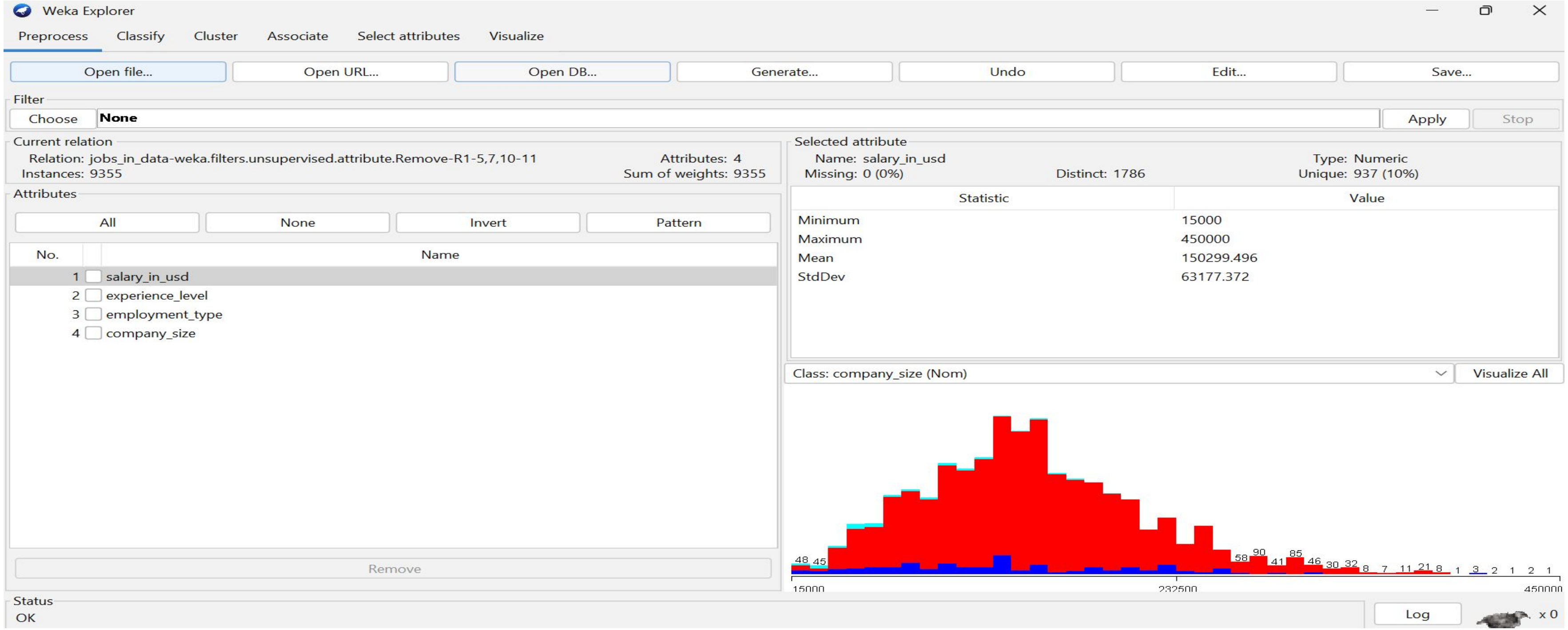
Class: company_size (Nom)

Status
OK

Distribution of the Target Class

- The target class for this classification task is the job category.
- It serves as the class label that we aim to predict or classify.
- The job category is chosen because the goal is to categorize job listings into different job categories.
- Attributes like salary, experience level, employment type, and company size are used as features for classification.
- These attributes help determine the nature of the job and are input features for the classification model.
- Therefore, the job category becomes the target class as it encapsulates the main objective of classifying jobs into specific categories.
- By predicting the job category, we can effectively categorize job listings based on various attributes.

Attributes to Train the Model



Filter: None

Current relation

Relation: jobs_in_data-weka.filters.unsupervised.attribute.Remove-R1-5,7,10-11
Attributes: 4
Instances: 9355
Sum of weights: 9355

Attributes

No.		Name
1	□	salary_in_usd
2	□	experience_level
3	□	employment_type
4	□	company_size

Selected attribute

Name: salary_in_usd
Type: Numeric
Missing: 0 (0%)
Distinct: 1786
Unique: 937 (10%)

Statistic	Value
Minimum	15000
Maximum	450000
Mean	150299.496
StdDev	63177.372

Class: company_size (Nom)

Status

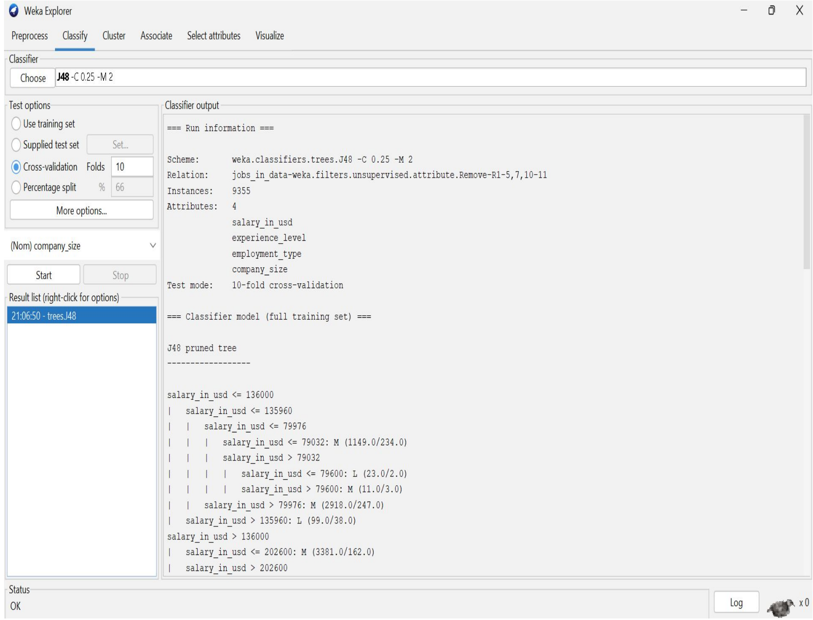
OK

Attributes to Train the Model

Chosen attributes for training the model are:

- **salary_in_usd:** Represents standardized salary in USD, crucial for job seekers and employers.
- **experience_level:** Indicates the level of experience required, impacting salary and job responsibilities.
- **employment_type:** Specifies employment terms (e.g., Full-time, Part-time), influencing work conditions.
- **company_size:** Denotes the size of the company, providing insights into organizational dynamics and career growth opportunities.
- These attributes were selected because they collectively cover essential aspects of job listings relevant to both job seekers and employers, including financial factors, job requirements, employment terms, and organizational characteristics. Other attributes excluded because of less informative for predicting job suitability and salary expectations.

Classifier Model: J48 -C 0.25 -M 2



J48 -C 0.25 -M 2

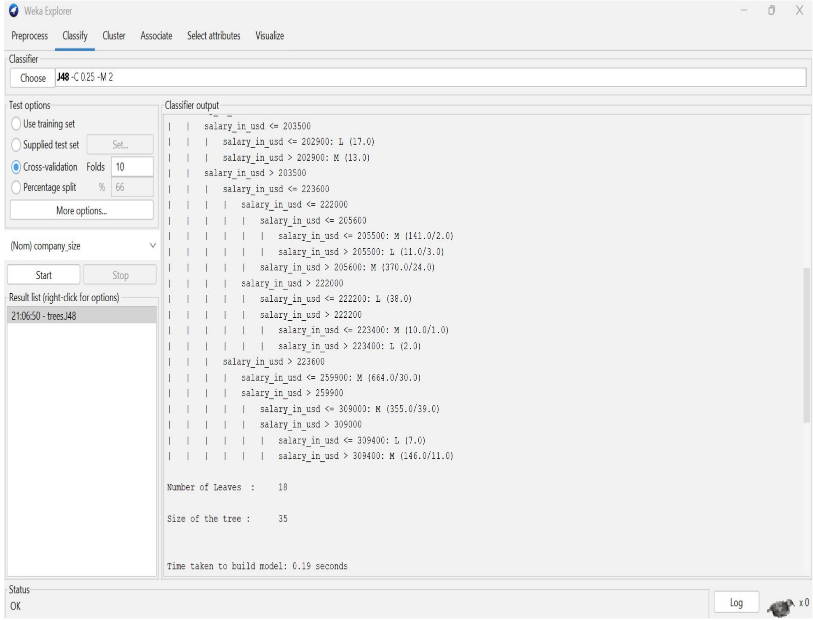
○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) company_size

21:06:50 - trees.J48

Status OK

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     jobs_in_data-weka.filters.unsupervised.attribute.Remove-R1-5,7,10-11
Instances:    9355
Attributes:   4
              salary_in_usd
              experience_level
              employment_type
              company_size
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

salary_in_usd <= 136000
|   salary_in_usd <= 135960
|   |   salary_in_usd <= 79976
|   |   |   salary_in_usd <= 79032: M (1149.0/234.0)
|   |   |   salary_in_usd > 79032
|   |   |   |   salary_in_usd <= 79600: L (23.0/2.0)
|   |   |   |   salary_in_usd > 79600: M (11.0/3.0)
|   |   salary_in_usd > 79976: M (2918.0/247.0)
|   salary_in_usd > 135960: L (99.0/38.0)
salary_in_usd > 136000
|   salary_in_usd <= 202600: M (3381.0/162.0)
|   salary_in_usd > 202600
```



J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66

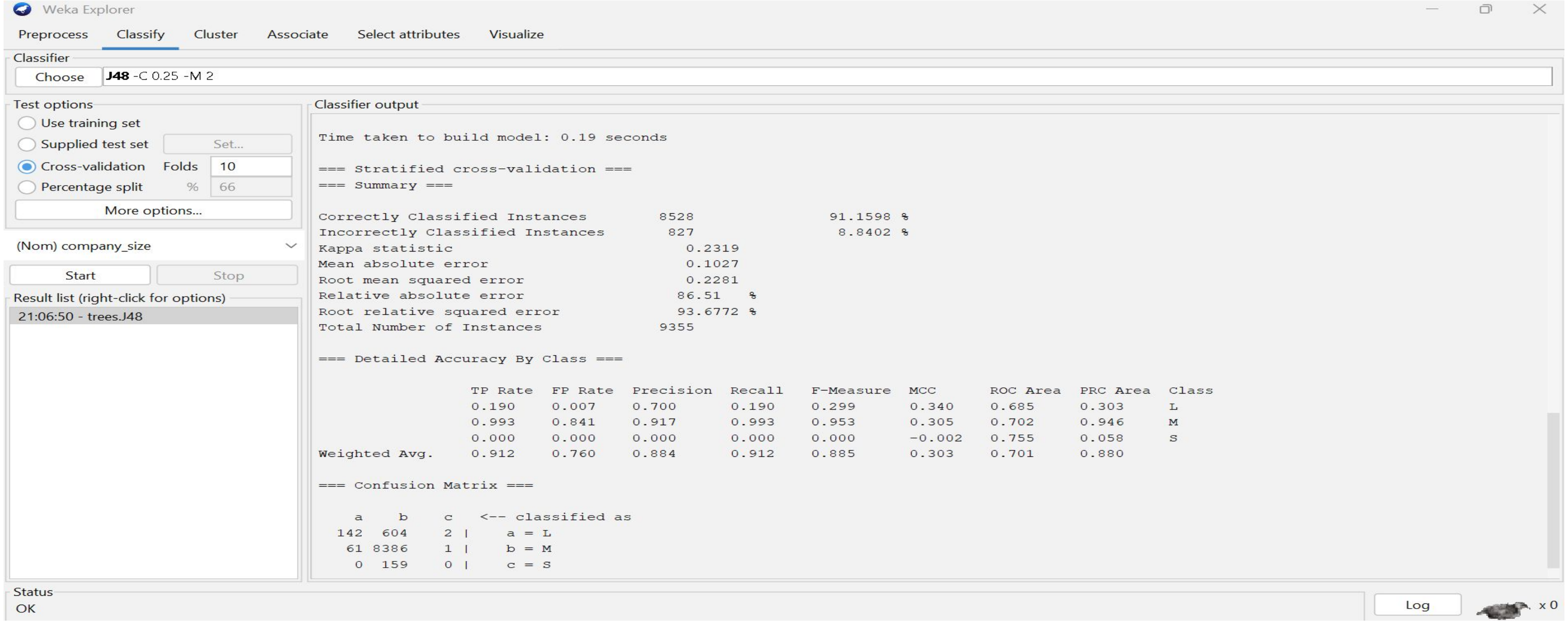
(Nom) company_size

21:06:50 - trees.J48

Status OK

```
|   |   salary_in_usd <= 203500
|   |   |   salary_in_usd <= 202900: L (17.0)
|   |   |   salary_in_usd > 202900: M (13.0)
|   |   salary_in_usd > 203500
|   |   |   salary_in_usd <= 223600
|   |   |   |   salary_in_usd <= 222000
|   |   |   |   |   salary_in_usd <= 205600
|   |   |   |   |   |   salary_in_usd <= 205500: M (141.0/2.0)
|   |   |   |   |   |   salary_in_usd > 205500: L (11.0/3.0)
|   |   |   |   |   salary_in_usd > 205600: M (370.0/24.0)
|   |   |   |   salary_in_usd > 222000
|   |   |   |   |   salary_in_usd <= 222200: L (38.0)
|   |   |   |   |   salary_in_usd > 222200
|   |   |   |   |   |   salary_in_usd <= 223400: M (10.0/1.0)
|   |   |   |   |   |   salary_in_usd > 223400: L (2.0)
|   |   |   salary_in_usd > 223600
|   |   |   |   salary_in_usd <= 259900: M (664.0/30.0)
|   |   |   |   salary_in_usd > 259900
|   |   |   |   |   salary_in_usd <= 309000: M (355.0/39.0)
|   |   |   |   |   salary_in_usd > 309000
|   |   |   |   |   |   salary_in_usd <= 309400: L (7.0)
|   |   |   |   |   |   salary_in_usd > 309400: M (146.0/11.0)

Number of Leaves  :     18

Size of the tree :      35



Time taken to build model: 0.19 seconds
```

Classifier Model: J48 -C 0.25 -M 2



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) company_size

Result list (right-click for options)

21:06:50 - trees.J48

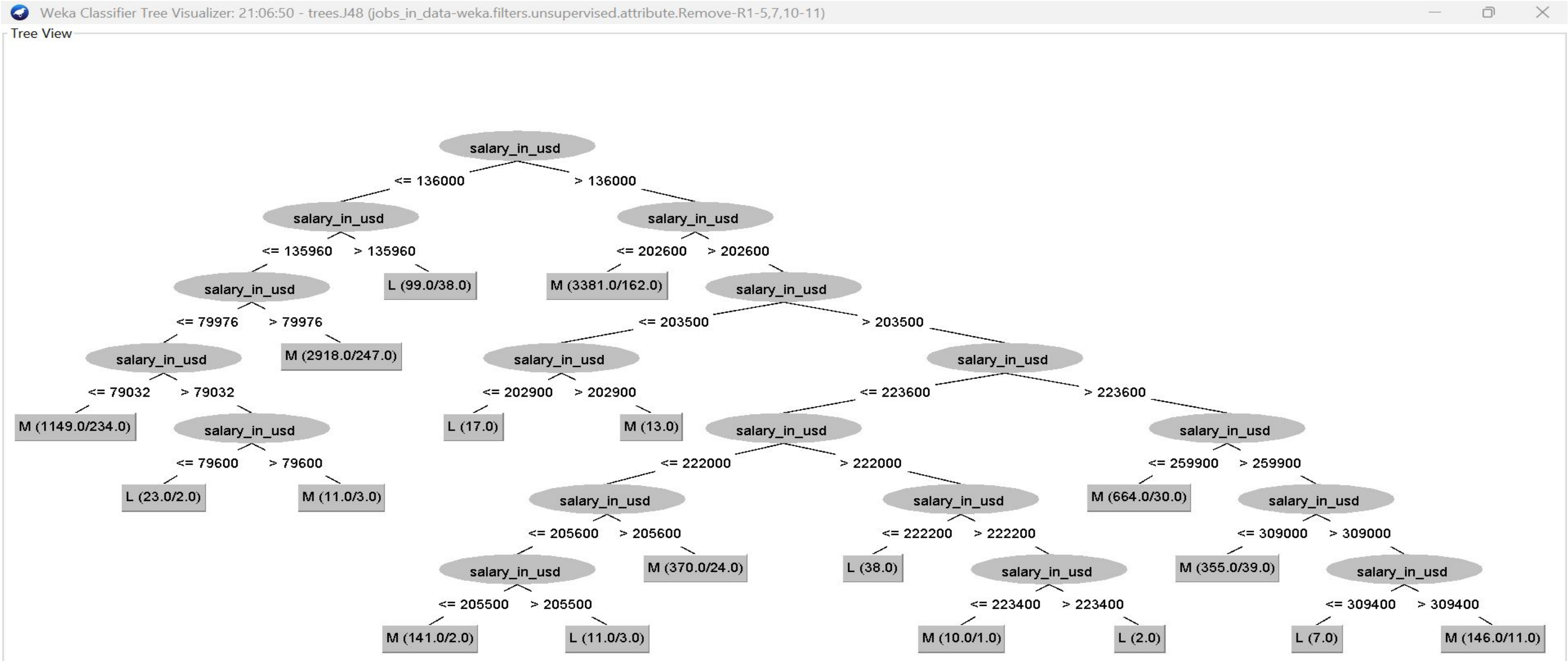
Classifier output

```
Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8528               91.1598 %
Incorrectly Classified Instances       827                8.8402 %
Kappa statistic                          0.2319
Mean absolute error                      0.1027
Root mean squared error                  0.2281
Relative absolute error                 86.51   %
Root relative squared error             93.6772 %
Total Number of Instances             9355

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.190    0.007    0.700      0.190    0.299      0.340    0.685     0.303     L
                 0.993    0.841    0.917      0.993    0.953      0.305    0.702     0.946     M
                 0.000    0.000    0.000      0.000    0.000      -0.002   0.755     0.058     S
Weighted Avg.    0.912    0.760    0.884      0.912    0.885      0.303    0.701     0.880

=== Confusion Matrix ===

    a    b    c   <-- classified as
  142  604    2 |    a = L
   61 8386    1 |    b = M
    0  159    0 |    c = S
```

Status

OK

Decision Tree



Weka Classifier Tree Visualizer: 21:06:50 - trees.J48 (jobs_in_data-weka.filters.unsupervised.attribute.Remove-R1-5,7,10-11)

Interpreting a J48 Decision Tree Model

- The J48 decision tree model achieved an overall accuracy of approximately 91.16% in the stratified 10-fold cross-validation.
- The model's decision tree is relatively complex, with 35 nodes and 18 leaves, indicating multiple decision paths based on the attributes salary_in_usd, experience_level, employment_type, and company_size.
- Specific rules generated by the tree involve conditions on salary_in_usd to predict job categories. For example, if salary_in_usd is less than or equal to $136,000, the model predicts 'L' (low salary job). If salary_in_usd is greater than $136,000, the model further examines the salary range to make predictions.
- The model has 827 incorrect classifications out of 9355 instances in total, indicating that it misclassified approximately 8.84% of instances.
- The specific errors for each possible classification outcome vary: For class 'L' (low salary), there were 142 instances misclassified as 'L', 604 instances correctly classified as 'L', and 2 instances misclassified as 'M'.
- For class 'M' (medium salary), there were 61 instances misclassified as 'L', 8386 instances correctly classified as 'M', and 1 instance misclassified as 'S'.
- For class 'S' (low frequency), there were no instances correctly classified, indicating a significant error for this class.

Evaluation and Summary

- Recall for class 'L' (low salary): 0.190
- Recall for class 'M' (medium salary): 0.993
- Recall for class 'S' (low frequency): 0.000
- Precision for class 'L': 0.700
- Precision for class 'M': 0.917
- Precision for class 'S': 0.000
- Accuracy: 91.16%
- The model performed relatively well in predicting classes with medium salaries (Class M), achieving a high recall and precision. However, it struggled with classes of low salaries (Class L) and low frequency (Class S), as indicated by low recall and precision scores.
- The model met expectations in some aspects, such as accurately predicting medium salary jobs, but it fell short in predicting low salary and low frequency jobs.
- The limited predictive power for low salary and low frequency jobs could be due to the complexity of real-world job market dynamics.

Thank You