

Classifying AI and Data Science Salaries and Mining Salary Association Rules with Weka

Fouad Bouabdi-Benaatou

Personal Information

Student ID: 2242740
Student Name: Fouad Bouabdi-Benaatou
Course: Data Mining and Visualisation ACCA7015
Assessment: Component 2 – Data Mining Report
Teacher: Dr Nouh Elmitwally
Address: University of Wales Trinity Saint David
MSc Data and Artificial Intelligence
Birmingham

Contents

1	Introduction	3
2	Literature Review	3
2.1	The AI and Data Science Job Market	3
2.2	Predicting Salaries with Data Mining	4
2.3	Classification Techniques	4
2.4	Evaluating Classifiers	4
2.5	Association Rule Mining	5
2.6	Summary and Research Gap	5
3	Dataset and Preparation	5
3.1	The Dataset	5
3.2	Cleaning	5
3.3	Feature Engineering	5
3.4	Findings from the Exploratory Analysis	6
4	Methodology	6
4.1	Process and Tools	6
4.2	Classification Models	6
4.3	Evaluation	7
4.4	Association Rules	7
5	Results	7
5.1	Model Comparison	7
5.2	Results by Class	7
5.3	What the Models Learned	8
5.4	Effect of Removing Redundant Attributes	8
5.5	Association Rules	8
6	Discussion	9
6.1	Answer to the Research Question	9
6.2	Why the Medium Band Is Hard	9
6.3	Agreement Between the Two Parts	10
6.4	Limitations	10
7	Business Recommendations	10
7.1	For Employers and Recruiters	10
7.2	For Job Seekers	10
7.3	For Training Providers	11
7.4	For Future Analysis	11
8	Conclusion	11
	References	12
	A Weka Evidence	14
	List of Figures	17
	List of Tables	17
	List of Abbreviations	17

Abstract

Demand for AI and data skills has grown fast, but pay in these jobs varies a lot. This report applies data mining to the Data Science Salaries dataset, which has 565 unique records of AI and data jobs from 2020 to 2022. Salary was split into three equal bands: Low, Medium and High. In Weka, a J48 decision tree, Naive Bayes and Random Forest were compared with two baselines using ten runs of ten-fold cross-validation and a corrected paired t-test. Random Forest reached the highest accuracy (64.5%), but it was not significantly better than J48 (63.0%) or Naive Bayes (62.3%). All three beat the baselines. Low and High salaries were predicted well, while Medium salaries were often confused with the other bands. The Apriori algorithm, with lift as an extra measure, found rules that link High pay to North America and Senior level, and Low pay to Asia and hybrid work. Both methods show that region and experience level are the main factors. The report ends with recommendations for employers, job seekers and training providers.

Keywords: data mining, classification, association rules, Weka, salary prediction, data science jobs

1 | Introduction

Demand for artificial intelligence (AI) and data skills has grown fast in the last decade. In US online job postings, demand for AI skills rose sharply between 2010 and 2019. Postings that ask for AI skills also offer about 11% higher pay than other postings in the same firm [1]. AI-related vacancies grew quickly, mostly at establishments whose tasks suit current AI. However, the effect on wider employment and wages is still too small to detect [2]. The way people work has also changed. The share of US job postings that offer at least one remote day per week more than tripled between 2019 and 2025 [3].

These changes make salary an important question for several groups. Employers and recruiters need pay offers that are fair and competitive. Training providers need to know which skills and career levels the market rewards. Job seekers want to know what affects their pay. Data mining can help, because it finds patterns in salary data that are hard to see by hand. For example, Martín *et al.* [4] studied 4,000 IT job offers in Spain. They found that experience is rewarded more than education, and they used tree-based ensemble methods to classify salary ranges.

This report continues the exploratory analysis of Component 1. It uses the Data Science Salaries dataset [5], which has 607 records of AI and data jobs from 2020 to 2022. After cleaning, 565 records remain. The exploratory analysis showed that company location and experience level were by far the strongest factors linked to salary. This report tests whether these and other job-market attributes can predict salary, and which combinations of attributes go together with high pay.

The research question is: *How well can job-market attributes classify AI and Data Science salaries into Low, Medium and High bands, and which combinations of attributes are most strongly linked to a High salary?*

The report has two parts. In the supervised part, three classifiers (a J48 decision tree, Naive Bayes and Random Forest) are built in Weka [6] and compared with two simple baselines. In the unsupervised part, the Apriori algorithm [7] is used to find association rules. The work follows the CRISP-DM process [8].

The rest of the report is organised as follows. Section 2 reviews the literature. Section 3 describes the dataset and its preparation, and Section 4 explains the methods. Section 5 presents the results, and Section 6 discusses them. Section 7 gives business recommendations, and Section 8 concludes.

2 | Literature Review

2.1 The AI and Data Science Job Market

Recent studies show strong growth in the demand for AI skills. Alekseeva *et al.* [1] found that AI skills carry a pay premium of about 11% within the same firm and 5% within the same job title.

Acemoglu *et al.* [2] also found fast growth in AI vacancies, but only small effects on the wider labour market so far. Hansen *et al.* [3] showed that remote work offers vary a lot between employers, even between employers in the same industry who hire for the same jobs. These three studies use large sets of US job postings. This is a strength, because the samples are very large. It is also a limit, because postings show the pay that is offered, not the pay that people receive. The dataset in this report is much smaller, and the way it was collected is not documented. Its results should therefore be compared with these studies with care.

2.2 Predicting Salaries with Data Mining

In our search, we found few peer-reviewed studies that predict salaries in data jobs. Martín *et al.* [4] is one of the closest. They used 4,000 IT job offers from Spain and built a classifier for salary ranges with tree-based ensembles. Like this report, they turned salary into ranges instead of predicting an exact value. They also found that experience is rewarded more than education. Their data included the skills asked for in each offer. The dataset used here has no skill information, so the models can only use job title, experience, location, company and work arrangement. This is a clear gap. It also means that the accuracy reached here is likely to be lower than in studies that include skills.

2.3 Classification Techniques

Three classifiers are used in this report. They were chosen because they work in different ways.

J48 is the Weka version of the C4.5 decision tree [9], [10]. A decision tree splits the data on one attribute at a time and ends with a class in each leaf. Its main strength is that people can read it as a set of if-then rules. Its main weakness is that a single tree can change a lot when the data change a little.

Naive Bayes assumes that all attributes are independent of each other once the class is known. This assumption is rarely true. Domingos and Pazzani [11] showed that the method can still be optimal for classification, measured by error rate, even when the assumption is clearly broken. For this dataset the assumption is clearly broken, because employee region and company region are almost the same. It is therefore useful to test how much this matters.

Random Forest builds many trees. Each tree uses a random sample of the data and a random choice of attributes at each split [12]. Breiman showed that the error of the forest settles to a limit as more trees are added. In a large comparison of 179 classifiers on 121 datasets, random forests were the most likely to be the best [13]. The cost is that a forest is much harder to explain than a single tree.

Simple baselines are also needed. ZeroR always predicts the most common class. OneR builds a rule from a single attribute. Holte [14] found that, on most of the datasets he studied, the best one-attribute rules were about as accurate as the rules from most machine learning systems. A complex model is only useful if it clearly beats these baselines.

2.4 Evaluating Classifiers

Kohavi [15] compared ways of estimating accuracy and recommended ten-fold stratified cross-validation for choosing a model. Accuracy alone can be misleading, so other measures are needed. Cohen's kappa [16] measures agreement after removing the agreement expected by chance. Sokolova and Lapalme [17] analysed 24 measures for binary, multi-class, multi-labelled and hierarchical tasks. They showed that the measures react in different ways to changes in the confusion matrix. For this reason, this report uses accuracy, kappa, and precision, recall and F-measure for each class. To compare two models fairly, the test must allow for the fact that the result depends on which data were used for training. Nadeau and Bengio [18] showed that many tests ignore this, and they proposed better ways to estimate the variance. Weka's Experimenter offers a corrected paired t-test for this purpose.

2.5 Association Rule Mining

Agrawal and Srikant [7] introduced the Apriori algorithm. It finds all rules whose support and confidence are above chosen thresholds. Support is the share of records that contain the rule. Confidence is the share of records with the left-hand side that also have the right-hand side. Confidence has a weakness: a rule can look strong only because its right-hand side is already common. Brin *et al.* [19] discussed this problem and proposed conviction, a measure that also takes into account how common the right-hand side is. Geng and Hamilton [20] reviewed many such measures. They argued that good measures help users to rank patterns and to keep only the interesting ones. This matters here, because a first Apriori run on this data gave many rules that said almost the same thing.

2.6 Summary and Research Gap

The studies reviewed agree that AI and data jobs are growing, and that pay and work arrangements vary by skill and by firm. However, the labour studies reviewed here use very large US posting data, and we found few studies that apply data mining to salaries in data jobs. The method literature supports trees, Naive Bayes and Random Forest as strong and well-understood choices, and it shows that results must be checked against simple baselines with proper tests. This report applies these methods to a small, global dataset of data jobs. It compares accuracy with interpretability, and it uses association rules to describe the attribute patterns behind high pay.

3 | Dataset and Preparation

3.1 The Dataset

The Data Science Salaries dataset [5] has 607 records and 11 variables. Each record is one job in AI or data, from 2020, 2021 or 2022. The variables describe the job (title, experience level and employment type), the salary (in local currency and in US dollars), the location of the employee and the company, the company size and the remote ratio. There are no missing values. The data are not balanced: 61% of the jobs are at North American companies, and 97% are full-time.

3.2 Cleaning

The cleaning was done in Python and repeated in Weka as a check, with the same results. First, the file has an empty first column with no name. It is only a saved row number, and Weka cannot load the file while it is there, so it was removed. Second, 42 rows were exact copies of other rows. There is no ID column, so they could be different people with the same job and pay. They were removed to avoid counting the same record twice, which leaves 565 records. This choice changes the median salary only from \$101,570 to \$100,000. Third, the salary and salary currency columns were removed, because the salary in US dollars already gives the same information in one currency.

Ten salaries were above the upper limit of the interquartile range (IQR) rule, which is \$283,864. Weka's InterquartileRange filter found the same ten records. All ten are at US companies, and seven are senior or executive roles. They look like real salaries, so they were kept. They do not affect the classifiers much, because salary is only used to create the class.

3.3 Feature Engineering

Some attributes had too many values for rule-based methods. Job title had 50 values, and 28 of them appeared fewer than five times. Titles were grouped by keywords into six job families. Countries were grouped into four regions, both for the employee and for the company. A new attribute,

Table 1: Attributes in the Final Weka File (565 Records)

Attribute	Values (records)
work_year	2020 (72), 2021 (215), 2022 (278)
experience_level	Entry (88), Mid (208), Senior (243), Executive (26)
employment_type	Full-time (546), other three types (19)
job_family	Data Engineering (155), Data Science (153), Data Analysis (102), ML/AI (80), Leadership (58), Research (17)
employee_region	North America (323), Europe (160), Asia (60), Rest of World (22)
company_region	North America (346), Europe (154), Asia (47), Rest of World (18)
works_abroad	No (514), Yes (51)
remote_ratio	On-site (121), Hybrid (98), Remote (346)
company_size	Small (82), Medium (290), Large (193)
salary_band (class)	Low (192), Medium (189), High (184)

works_abroad, shows whether the employee lives in a different country from the company. Codes such as “SE” were replaced with readable labels, and the year and remote ratio were made nominal.

The class attribute, *salary_band*, was made by splitting salary in US dollars into three groups of almost equal size: Low (up to \$75,000; 192 records), Medium (up to \$130,000; 189 records) and High (above \$130,000; 184 records). Equal groups mean that no class dominates, and that always guessing the largest class is right only 34% of the time. The numeric salary was then removed, because the models could otherwise read the answer from it. Table 1 lists the final attributes. The same file is used for both classification and association rules.

3.4 Findings from the Exploratory Analysis

In Component 1, company region and experience level were by far the strongest factors linked to salary, with Kruskal–Wallis effect sizes of 0.39 and 0.30. Every other attribute was 0.07 or less. Some other gaps, such as the low pay of hybrid jobs, were caused by the mix of regions in the data. The models are therefore expected to rely mainly on region and experience.

4 | Methodology

4.1 Process and Tools

The work follows the CRISP-DM phases [8]. Component 1 covered business understanding, data understanding and most of the data preparation. This report covers modelling and evaluation, and it ends with recommendations for use. The data were prepared in Python. All models were built in Weka 3.8.7 [6], using the Explorer for single models and the Experimenter for the comparison of models.

4.2 Classification Models

Five classifiers were used, all with Weka’s default settings:

- **ZeroR** always predicts the largest class. It is the lowest baseline.
- **OneR** builds rules from the single best attribute [14]. It is a stronger baseline.
- **J48**, the C4.5 decision tree [9], with a pruning confidence of 0.25 and at least two records per leaf.

- **Naive Bayes** [11], which estimates the probability of each class from the attribute values.
- **Random Forest** [12] with 100 trees.

J48 was chosen as the main model because it is easy to read. Naive Bayes and Random Forest were chosen because they work in different ways: one is a simple probability model, and the other is an ensemble of many trees. Default settings were kept, because the dataset is small and tuning many settings on 565 records could easily overfit.

4.3 Evaluation

Each model was first tested in the Explorer with stratified ten-fold cross-validation [15]. This gave the confusion matrix and the precision, recall and F-measure of each class. The models were then compared in the Experimenter, using ten runs of ten-fold cross-validation, which gives 100 results for each model. The corrected paired t-test was used at the 5% level, with J48 as the base model. This test allows for the overlap between training sets [18].

Accuracy was the main measure, because the three classes have almost the same size. Cohen's kappa [16] was also reported. It shows how much better the model is than chance, which is useful when comparing with the 34% baseline. Results for each class were also reported, because a model can be good on average but weak on one class [17].

Some attributes in the data are almost copies of each other. Employee region and company region are the same for 94% of records, works_abroad is "No" for 91%, and employment type is "Full-time" for 97%. To test whether these attributes help or harm, the comparison was repeated on a reduced set of six attributes. This set removes employment type, company region and works_abroad.

4.4 Association Rules

Apriori [7] was run on the same file in class-rule mode, so that every rule has the salary band on its right-hand side. The minimum support was 0.05 (28 records) and the minimum confidence was 0.70. A first run on all nine attributes gave many rules that said almost the same thing. For example, 12 of the top 15 rules differed only by adding "Full-time", "works_abroad = No" or the second region attribute (Fig. 5). Apriori was therefore run on the reduced set of six attributes. Confidence alone can overrate a rule when its class is already common [19], [20], so lift was also calculated. Lift is the confidence of the rule divided by the share of records in that class. A lift above 1 means the rule makes the class more likely than normal.

5 | Results

5.1 Model Comparison

Table 2 shows the Experimenter results over ten runs of ten-fold cross-validation (Fig. 1 in Appendix A). All three main models beat both baselines. Random Forest had the highest mean accuracy (64.5%), followed by J48 (63.0%) and Naive Bayes (62.3%). However, the corrected paired t-test found no significant difference between J48 and either of the other two models ($p = 0.43$ for Random Forest and $p = 0.70$ for Naive Bayes). All three were significantly better than OneR (58.2%) and ZeroR (34.0%). Kappa values of 0.44 to 0.47 show moderate agreement beyond chance. The standard deviations of about 5 points show that accuracy changed quite a lot from fold to fold, which is expected with only 56 or 57 records in each test fold.

5.2 Results by Class

Table 3 shows the F-measure for each class and the J48 confusion matrix, from the Explorer (Figs. 2–4). All models found Low and High salaries much more easily than Medium ones. The

Table 2: Model Comparison (Experimenter, 10 Runs of 10-Fold CV)

Model	Acc. % (SD)	Kappa	W. F	vs J48
ZeroR	34.0 (0.7)	0.000	–	worse
OneR	58.2 (3.1)	0.375	–	worse
J48	63.0 (5.1)	0.445	0.620	base
Naive Bayes	62.3 (5.0)	0.435	0.592	n.s.
Random Forest	64.5 (5.9)	0.468	0.640	n.s.

n.s.: no significant difference (corrected paired t-test, $\alpha = 0.05$).

Table 3: F-Measure by Class and J48 Confusion Matrix (10-Fold CV)

F-measure	Low	Medium	High
J48	0.733	0.456	0.693
Naive Bayes	0.760	0.326	0.695
Random Forest	0.727	0.500	0.701
J48: actual ↓ / predicted →	Low	Medium	High
Low (192)	147	38	7
Medium (189)	53	80	56
High (184)	9	44	131

F-measure for Medium was only 0.33 to 0.50, compared with about 0.70 for the other two classes. Naive Bayes had the weakest Medium result: it found only 25% of Medium records and placed most of the others in Low or High. Random Forest was the most balanced model, with the best Medium F-measure (0.50).

The confusion matrix shows that almost all errors are between neighbouring bands. J48 confused Low with High in only 16 of 565 records (2.8%). Most errors involve the Medium band, which lies between the other two.

5.3 What the Models Learned

The OneR baseline chose employee region as its single attribute. Its rule is simple: North America predicts High, and all other regions predict Low. This rule alone was right 58% of the time. The J48 tree also splits on employee region first. Asia leads straight to Low (60 records, 13 errors). For Europe and North America, the next split is experience level. For example, Executive leads to High in both regions, and Mid-level in Europe leads to Low. Deeper branches use year, remote ratio, job family and company size, but they cover few records. The pruned tree has 36 leaves. These results agree with Component 1, where region and experience were the strongest factors.

5.4 Effect of Removing Redundant Attributes

With the reduced set of six attributes, J48 (63.1%) and Random Forest (64.8%) gave almost the same accuracy as before. Naive Bayes fell from 62.3% to 60.4%. None of these changes made any model significantly different from J48. The three removed attributes therefore add little information, and the simpler set can be used without loss.

5.5 Association Rules

Apriori found 14 rules with a confidence of at least 0.70 on the reduced set. All of them predict Low or High. No rule for Medium reached the threshold. The best Medium rule (Data Analysis, North America and a medium-sized company) had a confidence of only 0.64. Table 4 shows eight rules that

Table 4: Selected Class Association Rules (Reduced Attribute Set)

Rule (left-hand side $\Rightarrow$ band)	<i>n</i>	Conf.	Lift
Senior, N. America, Large $\Rightarrow$ High	48	0.81	2.49
2022, Senior, Data Science, N. America $\Rightarrow$ High	37	0.81	2.49
Senior, Data Science, N. America $\Rightarrow$ High	47	0.74	2.29
Data Eng., N. America, Remote, Medium co. $\Rightarrow$ High	48	0.73	2.24
2021, Asia $\Rightarrow$ Low	38	0.82	2.40
Asia $\Rightarrow$ Low	60	0.78	2.31
Europe, Hybrid $\Rightarrow$ Low	53	0.72	2.11
Mid, Hybrid $\Rightarrow$ Low	41	0.71	2.08

n: records matching the left-hand side. Region = employee region.

are not simple variants of each other (all 14 are in Fig. 6). Every rule has a lift above 2, which means the rule makes its salary band more than twice as likely as normal.

The High rules all include North America, and most also include Senior level. Senior jobs at large North American companies gave the strongest High rule (confidence 0.81, lift 2.49). Data Engineering jobs that are fully remote at medium-sized North American companies were also often High. The Low rules are led by Asia: 78% of jobs based in Asia are in the Low band. Hybrid work appears in two Low rules, together with Europe or with Mid-level experience.

6 | Discussion

6.1 Answer to the Research Question

The first part of the research question asked how well job-market attributes can classify salary bands. The answer is: moderately well. The best models were right about 63–65% of the time, almost twice the ZeroR baseline. However, the gain over OneR was only about 5 to 6 points. A single rule based on employee region was right 58% of the time. This agrees with Holte [14], who found that very simple rules are often close to more complex models. In this dataset, most of the useful information sits in one or two attributes.

Random Forest had the best mean accuracy, kappa and F-measure, but the difference from J48 was not significant. J48 can be read as a set of rules, while a forest of 100 trees cannot. When accuracy is the same, the readable model is more useful for the people who must act on it. For this reason, J48 is the preferred model here.

6.2 Why the Medium Band Is Hard

All models found Medium salaries much harder than Low or High ones. There are three likely reasons. First, Medium lies between the other two bands, so a record near \$75,000 or \$130,000 can easily fall on the wrong side. The confusion matrix supports this, because almost all errors were between neighbouring bands. Second, the band limits come from percentiles of this sample, not from any natural break in pay. Third, the data have no information on skills. Martín *et al.* [4] used the skills in each job offer. Without them, two jobs with the same title, level and location can still have very different pay.

Naive Bayes had the weakest Medium result, even though its overall accuracy was close to J48. Its independence assumption is broken in this data, yet its accuracy stayed competitive. This fits the finding of Domingos and Pazzani [11]. The cost appears in the hardest class, not in the overall figure, which shows why per-class measures are needed [17].

6.3 Agreement Between the Two Parts

The classifiers and the association rules tell the same story. Both point to region and experience level as the main factors. The J48 tree split first on employee region, and every High rule included North America. No Medium rule reached the confidence threshold, which matches the weak Medium results of the classifiers. Lift was useful, because confidence alone does not show how common each band already is [19], [20]. All selected rules had a lift above 2, so they are much stronger than the base rate.

Remote work appeared in some rules but not near the top of the tree. Fully remote Data Engineering jobs were linked to High pay, while hybrid work was linked to Low pay. However, 76% of hybrid jobs are outside North America, so this link may mostly reflect region. Hansen *et al.* [3] also found that remote work offers differ a lot between employers. The effect of remote work alone cannot be separated in this data.

6.4 Limitations

The results have several limits. The dataset is small, with 565 records, and 57% of employees are in North America. Only 60 records come from Asia and 22 from the rest of the world. How the data were collected is not documented, so the sample may not represent the real market. Salaries are in US dollars and are not adjusted for living costs, so part of the regional gap may reflect prices rather than the value of the work. The association rules show links, not causes. Finally, removing duplicate rows assumes that they are repeated entries. If some are real separate jobs, the counts would change slightly.

7 | Business Recommendations

The results suggest several practical actions. Because the data are small and not fully representative, these should be treated as starting points to check with larger, local data, not as final answers.

7.1 For Employers and Recruiters

- *Set pay by location and level first.* Region and experience level were the strongest factors in every model and in the rules. A pay framework should start from these two factors before adding others such as job family or company size.
- *Use the J48 rules as a simple check, not as a pay tool.* The tree gives clear rules, for example that Executive roles in North America are mostly High. Recruiters can use such rules to spot offers that look unusual. However, with about 63% accuracy, the model should not set individual salaries.
- *Review Medium-band roles with extra care.* The models often confused Medium with the bands next to it. Pay for these roles depends on details that the data do not capture, such as skills. Employers should compare these offers with market data for the specific skills they need.

7.2 For Job Seekers

- *Expect large gaps by location.* Jobs based in Asia were mostly Low, while senior jobs at large North American companies were mostly High. People who work for companies in another region should compare offers in US dollars and also consider their living costs.
- *Moving up a level matters.* Senior level appeared in three of the four High rules in Table 4. Gaining the experience to move from Mid to Senior level seems to be one of the clearest routes to higher pay.

7.3 For Training Providers

- *Link courses to career progression.* Since experience level drives pay, training that helps people move into senior roles may add more value than entry-level training alone.
- *Collect better data.* The lack of skill data was the main gap in this study. Providers and employers could collect salary data together with the skills used in each job. This would allow a clearer test of the AI skill premium found in job postings [1].

7.4 For Future Analysis

The same Weka process should be repeated on newer and larger salary data. Adding skills, industry and cost of living would likely improve the Medium band, which was the weakest result in this study.

8 | Conclusion

This report used Weka to classify AI and Data Science salaries into Low, Medium and High bands, and to find association rules linked to each band. The work continued the exploratory analysis of Component 1 and followed the CRISP-DM process.

The answer to the first part of the research question is that job-market attributes classify salary bands moderately well. J48, Naive Bayes and Random Forest reached 62–65% accuracy, against 34% for ZeroR. The differences between them were not significant. J48 is recommended, because its rules are easy to read. The models predicted Low and High well, but Medium poorly, because it lies between the other bands and the data have no skill information.

The answer to the second part is that High salaries are most strongly linked to North America and Senior level, especially at large companies. Low salaries are most strongly linked to Asia and to hybrid work in Europe or at Mid level. The rules and the classifiers agree that region and experience level are the main factors.

The main limits are the small and unbalanced sample, the unknown collection method and the lack of skill and living-cost data. Future work should use larger and newer data with skills, and could test ordinal or regression models that use the order of the salary bands.

References

- [1] L. Alekseeva, J. Azar, M. Giné, S. Samila, and B. Taska, “The demand for AI skills in the labor market,” *Labour Economics*, vol. 71, p. 102002, 2021.
- [2] D. Acemoglu, D. Autor, J. Hazell, and P. Restrepo, “Artificial intelligence and jobs: Evidence from online vacancies,” *Journal of Labor Economics*, vol. 40, no. S1, pp. S293–S340, 2022.
- [3] S. Hansen, P. J. Lambert, N. Bloom, S. J. Davis, R. Sadun, and B. Taska, “Remote work across jobs, companies, and space,” National Bureau of Economic Research, Cambridge, MA, USA, NBER Working Paper 31007, 2023, revised Jan. 2026.
- [4] I. Martín, A. Mariello, R. Battiti, and J. A. Hernández, “Salary prediction in the IT job market with few high-dimensional samples: A Spanish case study,” *International Journal of Computational Intelligence Systems*, vol. 11, no. 1, pp. 1192–1209, 2018.
- [5] ruchi798, “Data science job salaries,” Kaggle dataset, 2022, accessed: Oct. 17, 2026. [Online]. Available: <https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries>
- [6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: An update,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
- [7] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in *Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*. Santiago, Chile: Morgan Kaufmann, 1994, pp. 487–499.
- [8] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, *CRISP-DM 1.0: Step-by-Step Data Mining Guide*, SPSS Inc., 2000.
- [9] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.
- [10] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Cambridge, MA, USA: Morgan Kaufmann, 2016.
- [11] P. Domingos and M. Pazzani, “On the optimality of the simple Bayesian classifier under zero-one loss,” *Machine Learning*, vol. 29, no. 2–3, pp. 103–130, 1997.
- [12] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [13] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, “Do we need hundreds of classifiers to solve real world classification problems?” *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.
- [14] R. C. Holte, “Very simple classification rules perform well on most commonly used datasets,” *Machine Learning*, vol. 11, no. 1, pp. 63–90, 1993.
- [15] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, Montreal, QC, Canada, 1995, pp. 1137–1145.
- [16] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [17] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

- [18] C. Nadeau and Y. Bengio, “Inference for the generalization error,” *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003.
- [19] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, “Dynamic itemset counting and implication rules for market basket data,” in *Proc. 1997 ACM SIGMOD Int. Conf. Management of Data*, Tucson, AZ, USA, 1997, pp. 255–264.
- [20] L. Geng and H. J. Hamilton, “Interestingness measures for data mining: A survey,” *ACM Computing Surveys*, vol. 38, no. 3, p. 9, 2006.

A | Weka Evidence

The screenshots below were taken in Weka 3.8.7. The Windows taskbar is kept so that the date and time are visible. The images keep their full resolution, so the text can be read by zooming in on the PDF.

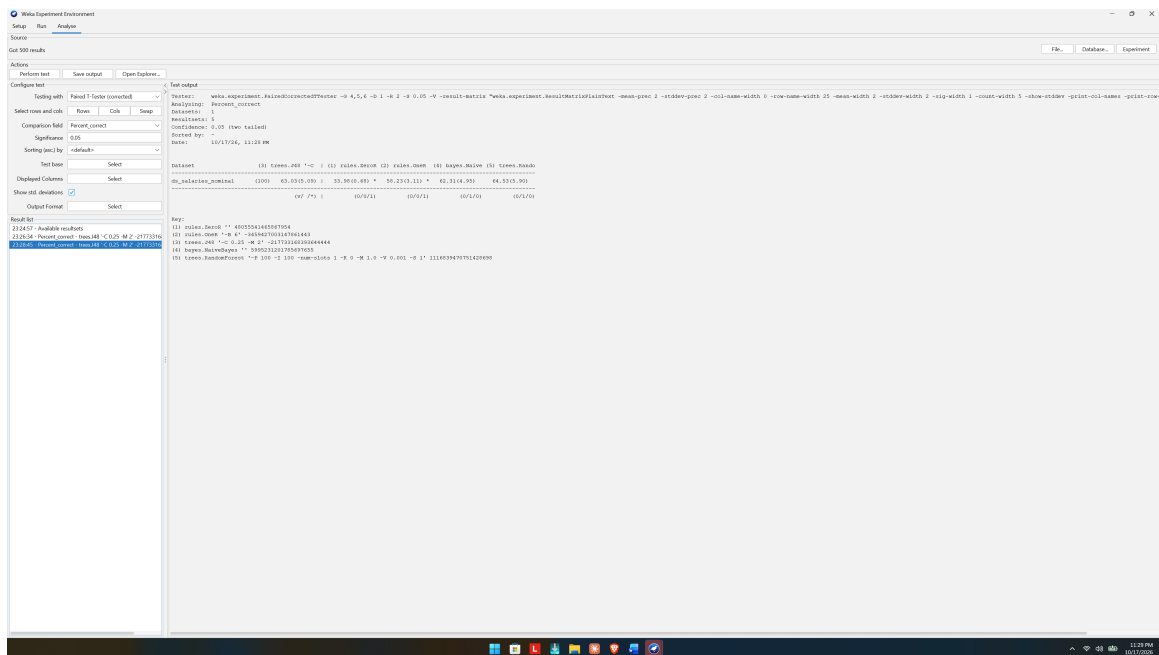


Figure 1: Experimenter: ten runs of ten-fold cross-validation with the corrected paired t-test (J48 as the base model, standard deviations in brackets). Source of Table 2.

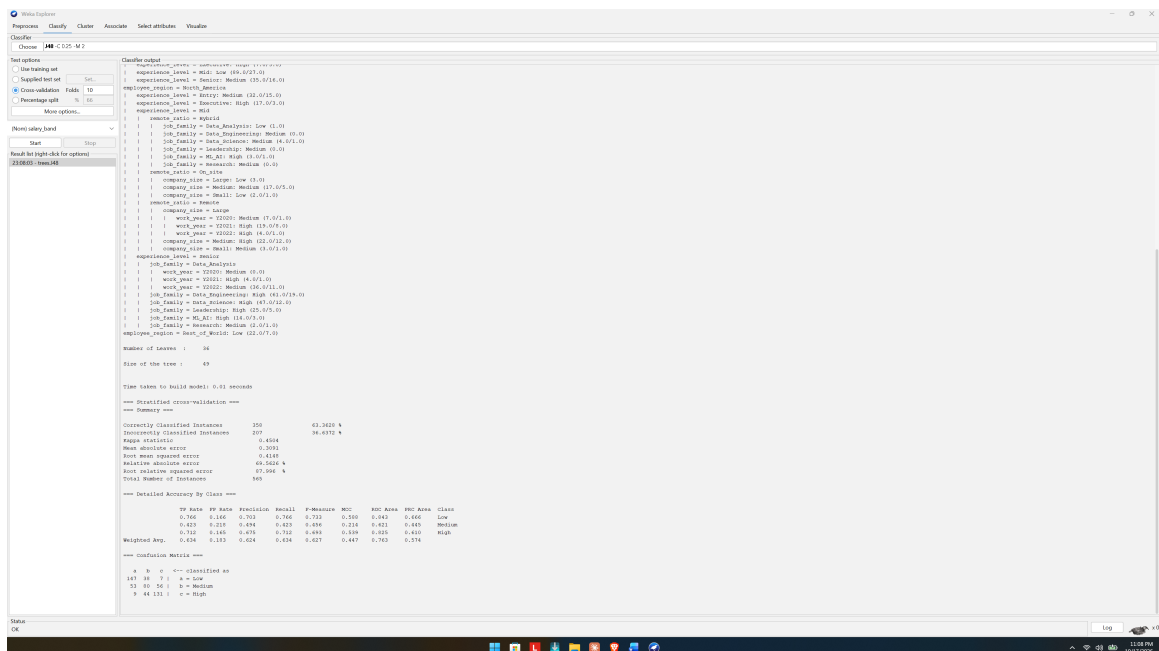


Figure 2: Explorer: J48 output with ten-fold cross-validation, showing the per-class results and the confusion matrix. Source of Table 3.

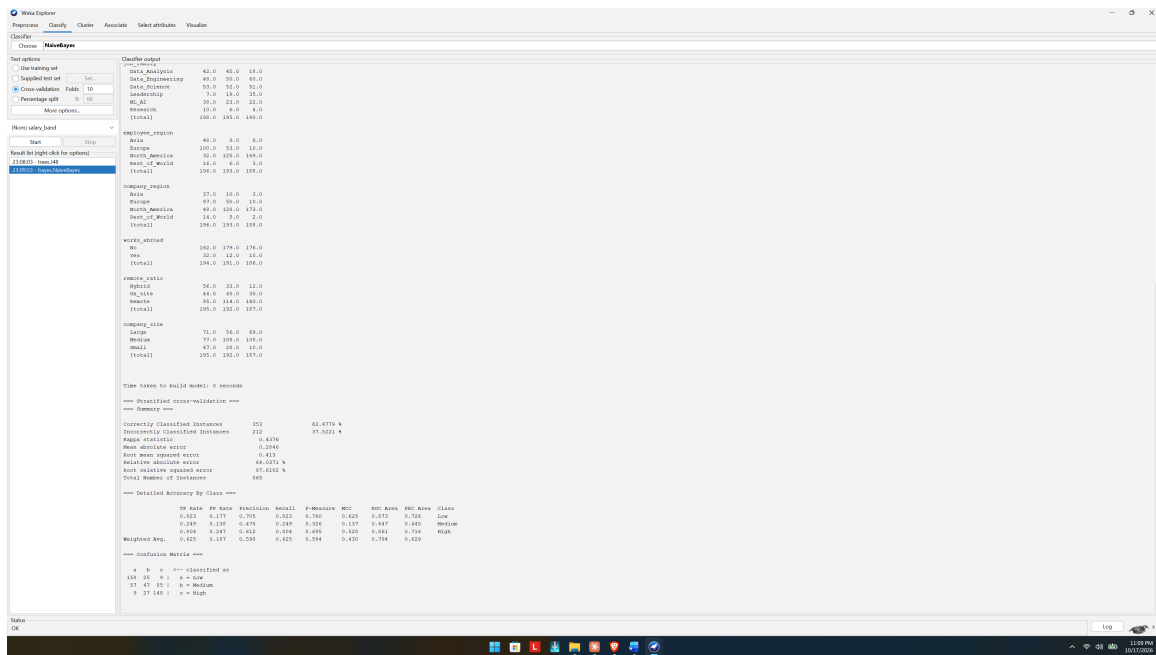


Figure 3: Explorer: Naive Bayes output with ten-fold cross-validation, showing the per-class results and the confusion matrix.

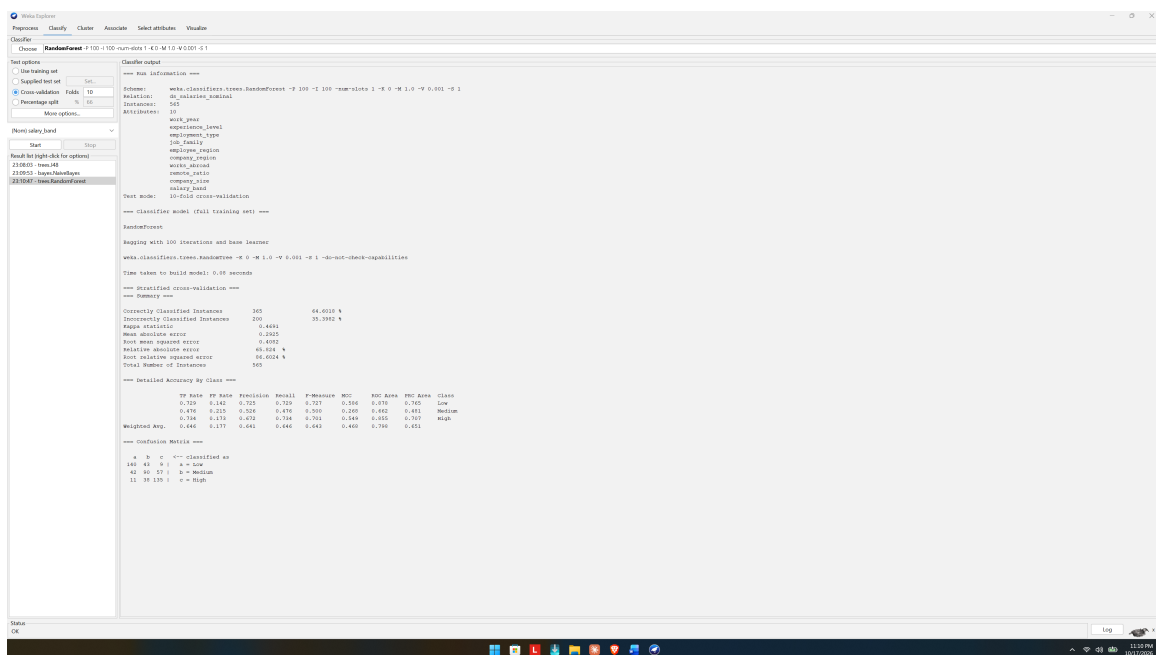
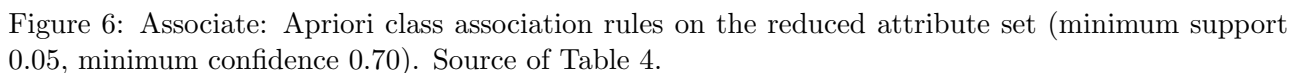
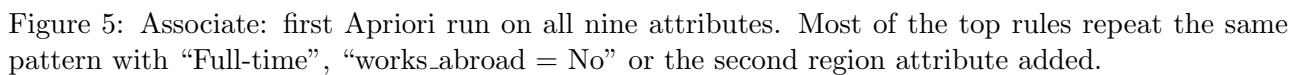


Figure 4: Explorer: Random Forest (100 trees) output with ten-fold cross-validation, showing the per-class results and the confusion matrix.



List of Figures

1	Experimenter: ten runs of ten-fold cross-validation with the corrected paired t-test (J48 as the base model, standard deviations in brackets). Source of Table 2.	14
2	Explorer: J48 output with ten-fold cross-validation, showing the per-class results and the confusion matrix. Source of Table 3.	14
3	Explorer: Naive Bayes output with ten-fold cross-validation, showing the per-class results and the confusion matrix.	15
4	Explorer: Random Forest (100 trees) output with ten-fold cross-validation, showing the per-class results and the confusion matrix.	15
5	Associate: first Apriori run on all nine attributes. Most of the top rules repeat the same pattern with “Full-time”, “works_abroad = No” or the second region attribute added.	16
6	Associate: Apriori class association rules on the reduced attribute set (minimum support 0.05, minimum confidence 0.70). Source of Table 4.	16

List of Tables

1	Attributes in the Final Weka File (565 Records)	6
2	Model Comparison (Experimenter, 10 Runs of 10-Fold CV)	8
3	F-Measure by Class and J48 Confusion Matrix (10-Fold CV)	8
4	Selected Class Association Rules (Reduced Attribute Set)	9

List of Abbreviations

AI	Artificial intelligence
CRISP-DM	Cross-Industry Standard Process for Data Mining
CV	Cross-validation
IQR	Interquartile range
n.s.	Not significant
SD	Standard deviation
W. F	Weighted F-measure