DATA WAREHOUSE AND DATA MINING (HDSE22.2F/CO)

ASSESSMENT - 04

Prepared By:

COHDSE222F - 004	O S KUMARASINGHE
COHDSE222F - 065	H A D U MALAKA
COHDSE222F - 064	R S A R A RUKSHAN
COHDSE222F - 062	A A T PABASARA

School of Computing and Engineering
National Institute of Business Management
Colombo-7

Dataset Selection

We have chosen the dataset titled "Data Science Salary (2021 to 2023)" from Kaggle's repository for several compelling reasons. This dataset aligns perfectly with the objectives of our assessment, and its attributes provide a comprehensive view of salary trends within the Data Science field over a three-year period.

Our primary goal is to analyze and understand salary trends in the Data Science sector from 2021 to 2023. This dataset is purpose-built for this objective, offering insights into various factors affecting salaries, such as experience levels, job titles, employment types, and company sizes.
The dataset includes a wide range of attributes that impact salaries, such as experience levels, job titles, company sizes, and employment types. These attributes enable us to conduct a holistic analysis and draw nuanced conclusions.

Overall the "Data Science Salary (2021 to 2023)" dataset from Kaggle which has been created by Harish Kumar DataLab is the perfect fit for our assessment due to its relevance, temporal scope, comprehensive attributes, real-world relevance, accessibility, and potential for valuable insights. It provides a robust foundation for our analysis and conclusions regarding salary trends in the Data Science field over the specified time period.

Link below to the dataset:
https://www.kaggle.com/datasets/harishkumardatalab/data-science-salary-2021-to-2023

This dataset has been curated to provide insights into the evolving salary patterns within the Data Science field, spanning the years 2021 to 2023. It focuses on a range of employment-related factors, including tenure, job roles, and company locations. By offering a comprehensive view of salary distributions, this dataset offers valuable intelligence for understanding the income landscape in this industry.

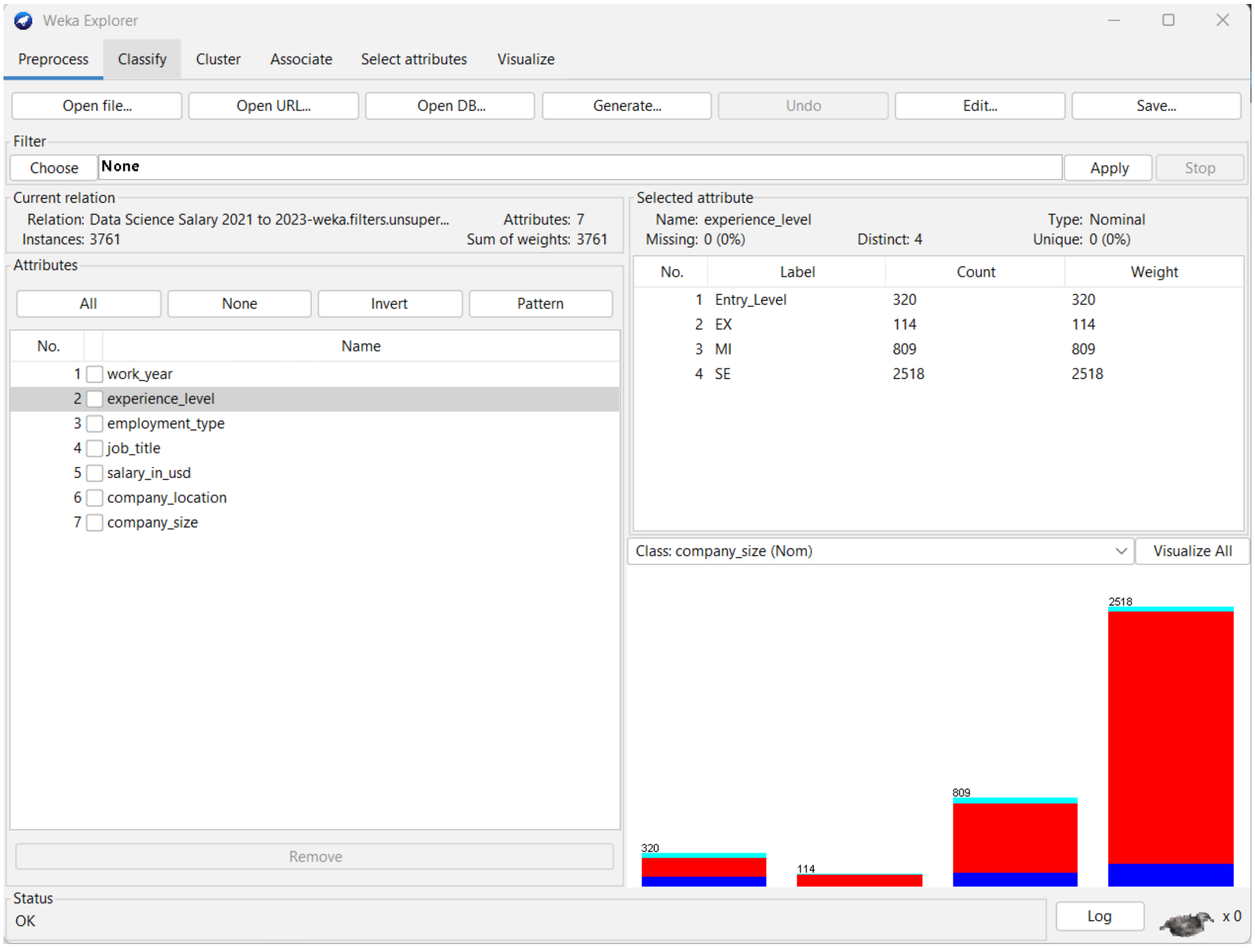
Attributes that are included are work year, Experience Level, Employment type, job title, salary, company location and Company Size.

This dataset serves as a valuable resource for data enthusiasts and analysts seeking to understand the salary dynamics among Data Science professionals in 2023. It enables the identification of trends across various experience levels, job titles, and company sizes. Both job seekers and employers can benefit from this dataset's insights to navigate the economic landscape within the Data Science job market.

Pre-processing

Data preprocessing is an important and crucial initial step before adding a dataset into the software for analysis. During this step, various operations are performed to keep important parts and remove unnecessary information, transform the dataset numeric values into string values, and organize the data. This ensures that the dataset is in a format that Weka software can effectively work with. By performing these actions, the data becomes more consistent, reliable, and ready for meaningful analysis using Weka Software.

After Selecting the dataset we are loading the dataset into the weka software, when we are loading it initially charts look like this.



Filter: None

Relation: Data Science Salary 2021 to 2023-weka.filters.unsuper... Attributes: 7
Instances: 3761 Sum of weights: 3761

No.	Name
1	□ work_year
2	□ experience_level
3	□ employment_type
4	□ job_title
5	□ salary_in_usd
6	□ company_location
7	□ company_size

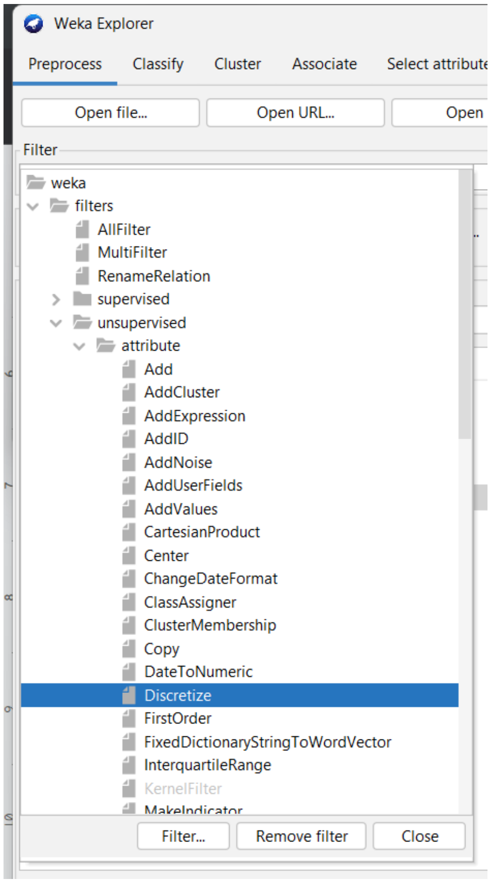
Name: experience_level Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

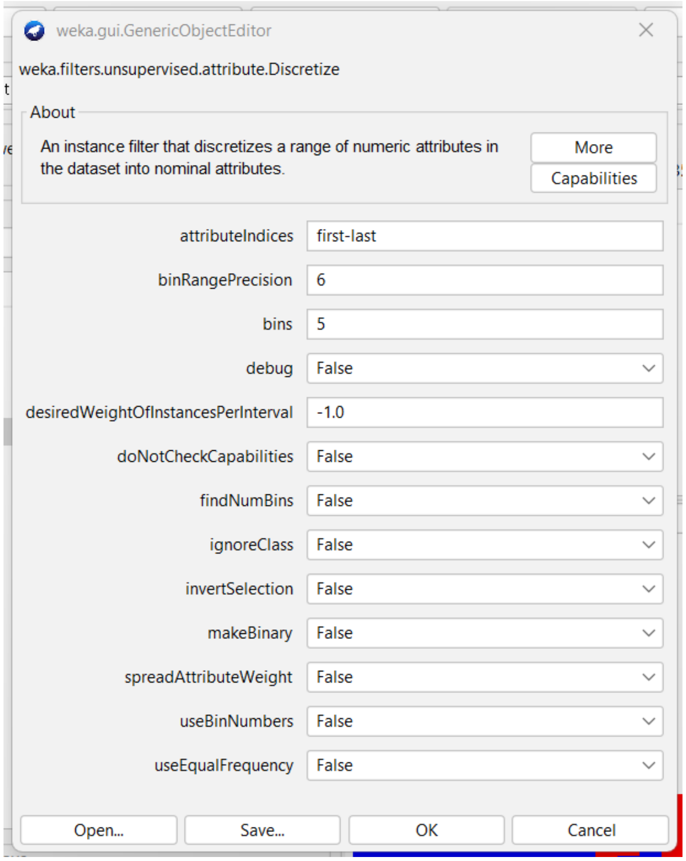
No.	Label	Count	Weight
1	Entry_Level	320	320
2	EX	114	114
3	MI	809	809
4	SE	2518	2518

Class: company_size (Nom)

OK

And as we can see there are numerical values, unwanted attributes and shortened values are available in this dataset. So, first we are removing unwanted values in this dataset. And after removing those we have to search for numerical values and convert them into appropriate characteristics data sets. In order to do that we need to filter data using a discretized filter in an unsupervised attributes folder.





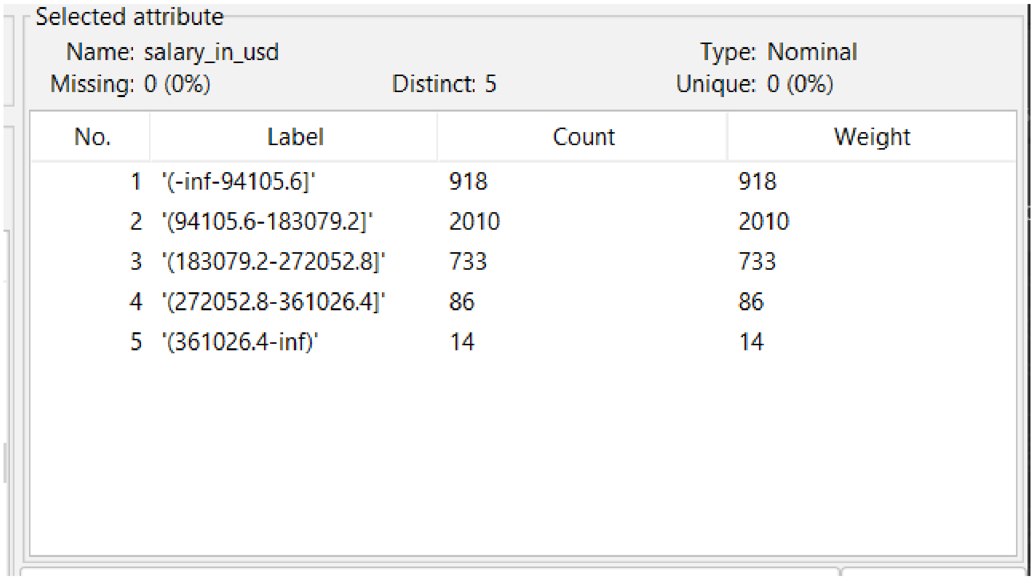
weka.filters.unsupervised.attribute.Discretize

About

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

attributeIndices	first-last
binRangePrecision	6
bins	5
debug	False
desiredWeightOfInstancesPerInterval	-1.0
doNotCheckCapabilities	False
findNumBins	False
ignoreClass	False
invertSelection	False
makeBinary	False
spreadAttributeWeight	False
useBinNumbers	False
useEqualFrequency	False

After adding the filter we need to set bins and divide numerical data into those bins.



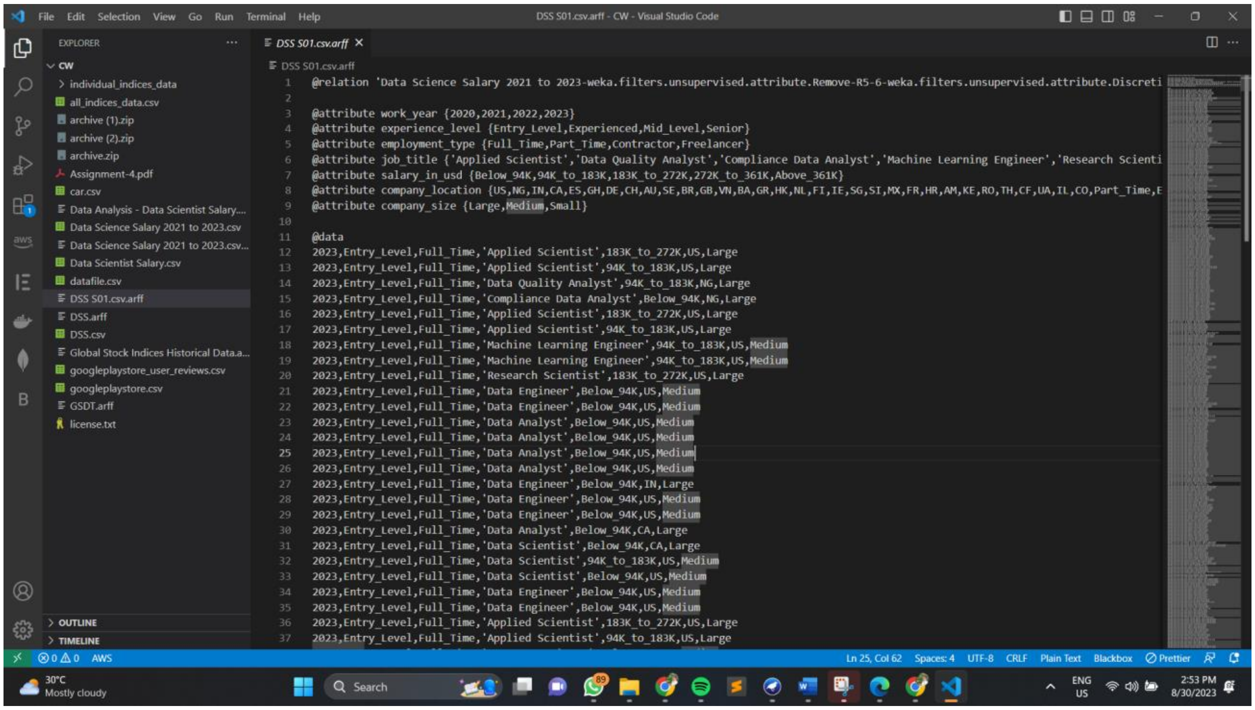
Selected attribute

Name: salary_in_usd Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-94105.6]'	918	918
2	'(94105.6-183079.2]'	2010	2010
3	'(183079.2-272052.8]'	733	733
4	'(272052.8-361026.4]'	86	86
5	'(361026.4-inf)'	14	14

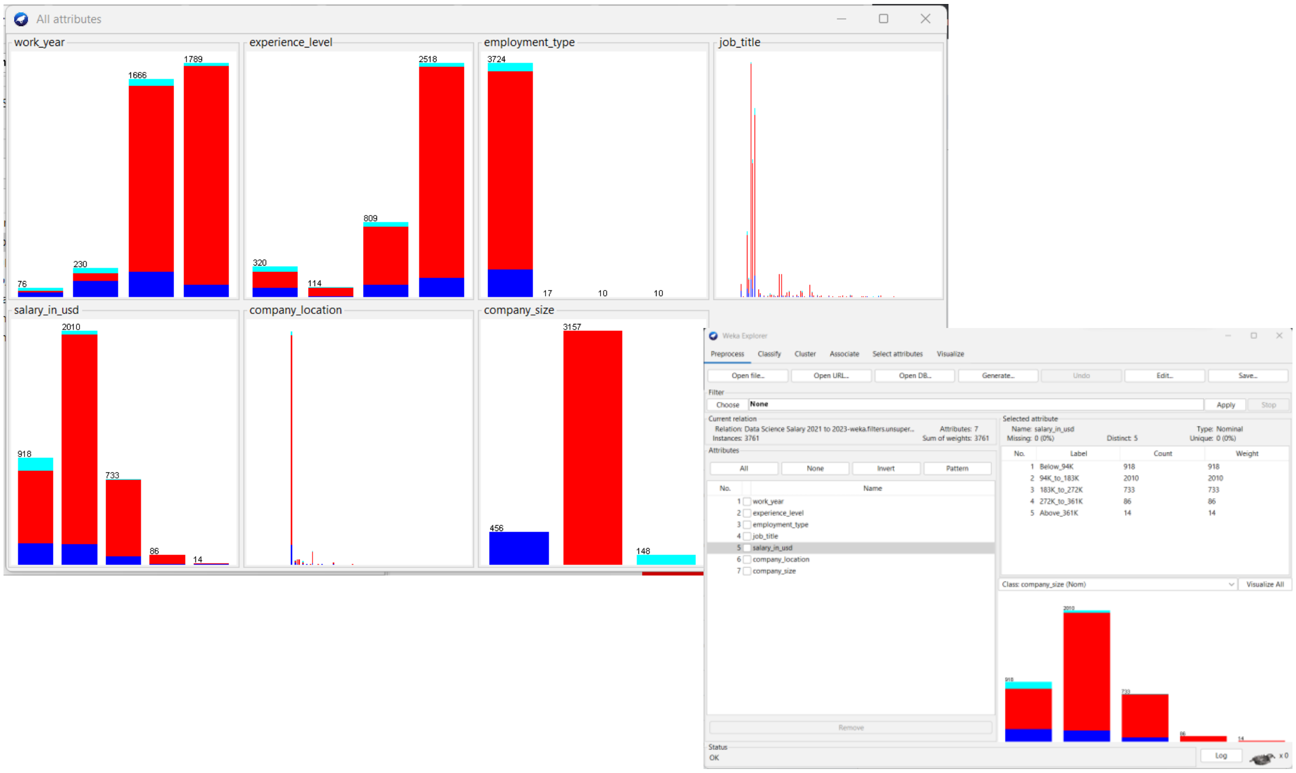
After that process data is looks like above image and after the dividing into the bins we need to save this as arff format and open using txt editor for renaming those bins.

For that we are using Visual Studio code editor and replacing automated texts with more meaningful names. Also in our data set include some short form for words (Ex: SE for Senior) we replaced those names with full words using data set instructions.


File Edit Selection View Go Run Terminal Help
DSS S01.csv.arff - CW - Visual Studio Code
EXPLORER
CW
individual_indices_data
all_indices_data.csv
archive (1).zip
archive (2).zip
archive.zip
Assignment-4.pdf
car.csv
Data Analysis - Data Scientist Salary...
Data Science Salary 2021 to 2023.csv
Data Science Salary 2021 to 2023.csv...
Data Scientist Salary.csv
datafile.csv
DSS S01.csv.arff
DSS.arff
DSS.csv
Global Stock Indices Historical Data.a...
googleplaystore_user_reviews.csv
googleplaystore.csv
GSDT.arff
license.txt
OUTLINE
TIMELINE
DSS S01.csv.arff
Ln 25, Col 62 Spaces: 4 UTF-8 CRLF Plain Text Blackbox Prettier
30°C
Mostly cloudy
Search
ENG
US
2:53 PM
8/30/2023

```
@relation 'Data Science Salary 2021 to 2023-weka.filters.unsupervised.attribute.Remove-R5-6-weka.filters.unsupervised.attribute.Discreti

@attribute work_year {2020,2021,2022,2023}
@attribute experience_level {Entry_Level,Experienced,Mid_Level,Senior}
@attribute employment_type {Full_Time,Part_Time,Contractor,Freelancer}
@attribute job_title {'Applied Scientist','Data Quality Analyst','Compliance Data Analyst','Machine Learning Engineer','Research Scienti
@attribute salary_in_usd {Below_94K,94K_to_183K,183K_to_272K,272K_to_361K,Above_361K}
@attribute company_location {US,NG,IN,CA,ES,GH,DE,CH,AU,SE,BR,GB,VN,BA,GR,HK,NL,FI,IE,SG,SI,MX,FR,HR,AM,KE,RO,TH,CF,UA,IL,CO,Part_Time,E
@attribute company_size {Large,Medium,Small}

@data
2023,Entry_Level,Full_Time,'Applied Scientist',183K_to_272K,US,Large
2023,Entry_Level,Full_Time,'Applied Scientist',94K_to_183K,US,Large
2023,Entry_Level,Full_Time,'Data Quality Analyst',94K_to_183K,NG,Large
2023,Entry_Level,Full_Time,'Compliance Data Analyst',Below_94K,NG,Large
2023,Entry_Level,Full_Time,'Applied Scientist',183K_to_272K,US,Large
2023,Entry_Level,Full_Time,'Applied Scientist',94K_to_183K,US,Large
2023,Entry_Level,Full_Time,'Machine Learning Engineer',94K_to_183K,US,Medium
2023,Entry_Level,Full_Time,'Machine Learning Engineer',94K_to_183K,US,Medium
2023,Entry_Level,Full_Time,'Research Scientist',183K_to_272K,US,Large
2023,Entry_Level,Full_Time,'Data Engineer',Below_94K,US,Medium
2023,Entry_Level,Full_Time,'Data Engineer',Below_94K,US,Medium
2023,Entry_Level,Full_Time,'Data Analyst',Below_94K,US,Medium
2023,Entry_Level,Full_Time,'Data Analyst',Below_94K,US,Medium
2023,Entry_Level,Full_Time,'Data Analyst',Below_94K,US,Medium
2023,Entry_Level,Full_Time,'Data Analyst',Below_94K,US,Medium
2023,Entry_Level,Full_Time,'Data Engineer',Below_94K,IN,Large
2023,Entry_Level,Full_Time,'Data Engineer',Below_94K,US,Medium
2023,Entry_Level,Full_Time,'Data Engineer',Below_94K,US,Medium
2023,Entry_Level,Full_Time,'Data Analyst',Below_94K,CA,Large
2023,Entry_Level,Full_Time,'Data Scientist',Below_94K,CA,Large
2023,Entry_Level,Full_Time,'Data Scientist',94K_to_183K,US,Medium
2023,Entry_Level,Full_Time,'Data Scientist',Below_94K,US,Medium
2023,Entry_Level,Full_Time,'Data Engineer',Below_94K,US,Medium
2023,Entry_Level,Full_Time,'Data Engineer',Below_94K,US,Medium
2023,Entry_Level,Full_Time,'Applied Scientist',183K_to_272K,US,Large
2023,Entry_Level,Full_Time,'Applied Scientist',94K_to_183K,US,Large
```

After the this process we got all attributes like this,



Filter: None

Current relation
Relation: Data Science Salary 2021 to 2023-weka.filters.unsuper... Attributes: 7
Instances: 3761 Sum of weights: 3761

Attributes

No.		Name
1	□	work_year
2	□	experience_level
3	□	employment_type
4	□	job_title
5	□	salary_in_usd
6	□	company_location
7	□	company_size

Selected attribute
Name: salary_in_usd Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	Below_94K	918	918
2	94K_to_183K	2010	2010
3	183K_to_272K	733	733
4	272K_to_361K	86	86
5	Above_361K	14	14

Class: company_size (Nom)

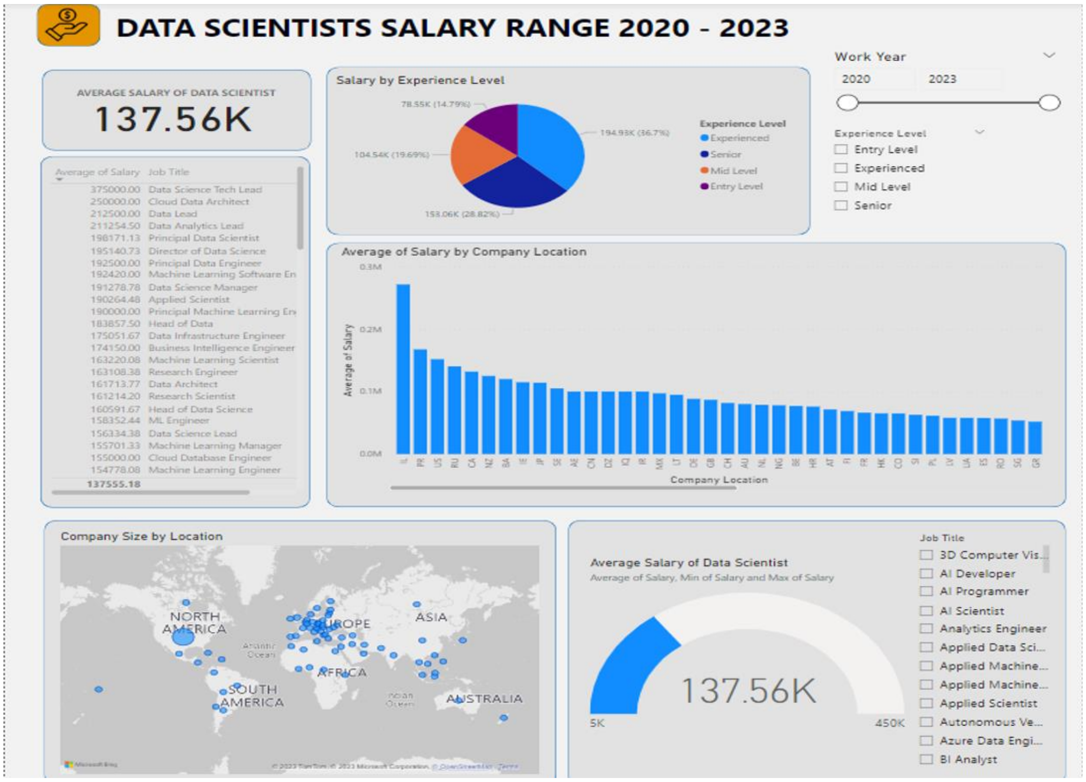
Status
OK

Data Analysis

What is data analysis?

The process of cleaning, modifying, and turning raw data into usable knowledge that allows informed decision-making, lowering risks, and offering helpful insights and statistics in charts, graphics, tables, and graphs is known as data analysis.

Power BI is a data analytics and reporting tool that combines information from many sources to offer you a thorough understanding of the information assets in your organization.



Average of Salary	Job Title
375000.00	Data Science Tech Lead
250000.00	Cloud Data Architect
212500.00	Data Lead
211254.50	Data Analytics Lead
198171.13	Principal Data Scientist
195140.73	Director of Data Science
192500.00	Principal Data Engineer
192420.00	Machine Learning Software En
191278.78	Data Science Manager
190264.48	Applied Scientist
190000.00	Principal Machine Learning En
183857.50	Head of Data
175051.67	Data Infrastructure Engineer
174150.00	Business Intelligence Engineer
163220.08	Machine Learning Scientist
163108.38	Research Engineer
161713.77	Data Architect
161214.20	Research Scientist
160591.67	Head of Data Science
158352.44	ML Engineer
156334.38	Data Science Lead
155701.33	Machine Learning Manager
155000.00	Cloud Database Engineer
154778.08	Machine Learning Engineer
137555.18	

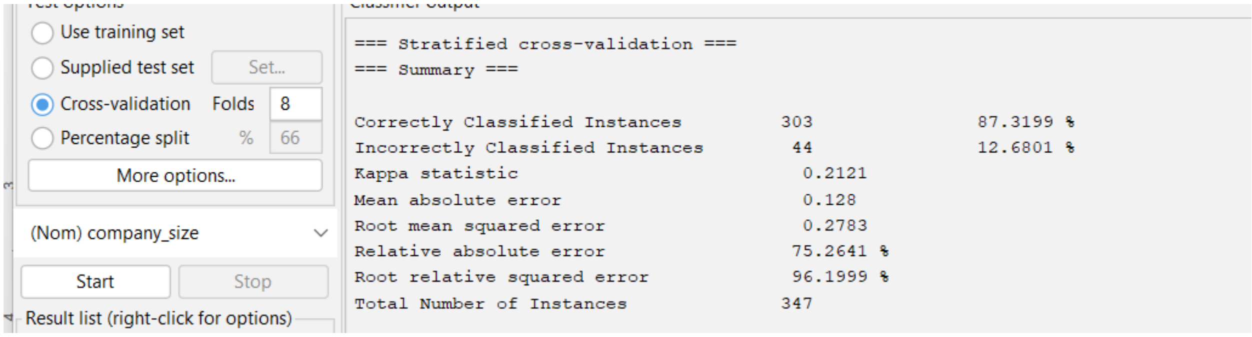
Figure 1 Data visualization done using PowerBi

As the above image illustrates you can see we have arranged the data in a way that we can identify easily.A card is used to mention the average salary on the data scientists.We have used the pie chart to show salary by experience level and as you can see Experienced people has the highest salary which is 36.7% and the lowest salary is taken by the Entry level people which is 14.79% .The Clustered column chart is utilized to present the average salary by Company Location furthermore it represents that the average salary is greater,close to 0.3M in Israel (IL) compared to other countries.Map presents the company size locations moreover gauge shows the average salary of data scientists including the minimum salary along with maximum salary where the min salary would be 5K , max salary 450K and the average salary of 137.56K.

Data Mining Techniques

Classification

J48 Algorithm



○ Use training set
○ Supplied test set [Set...]
● Cross-validation Folds 8
○ Percentage split % 66
[More options...]

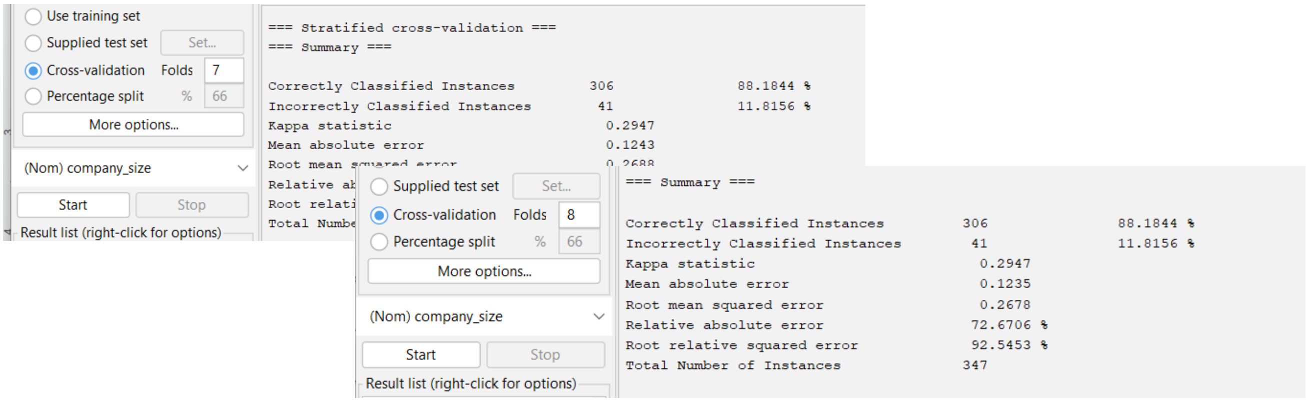
(Nom) company_size

[Start] [Stop]

Result list (right-click for options)

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

Correctly Classified Instances         303               87.3199 %
Incorrectly Classified Instances        44               12.6801 %
Kappa statistic                          0.2121
Mean absolute error                      0.128
Root mean squared error                  0.2783
Relative absolute error                 75.2641 %
Root relative squared error             96.1999 %
Total Number of Instances              347
```

Best Case



○ Use training set
○ Supplied test set [Set...]
● Cross-validation Folds 7
○ Percentage split % 66
[More options...]

(Nom) company_size

[Start] [Stop]

Result list (right-click for options)

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

Correctly Classified Instances         306               88.1844 %
Incorrectly Classified Instances        41               11.8156 %
Kappa statistic                          0.2947
Mean absolute error                      0.1243
Root mean squared error                  0.2688
Relative ab
Root relati
Total Numbe
```

○ Supplied test set [Set...]
● Cross-validation Folds 8
○ Percentage split % 66
[More options...]

(Nom) company_size

[Start] [Stop]

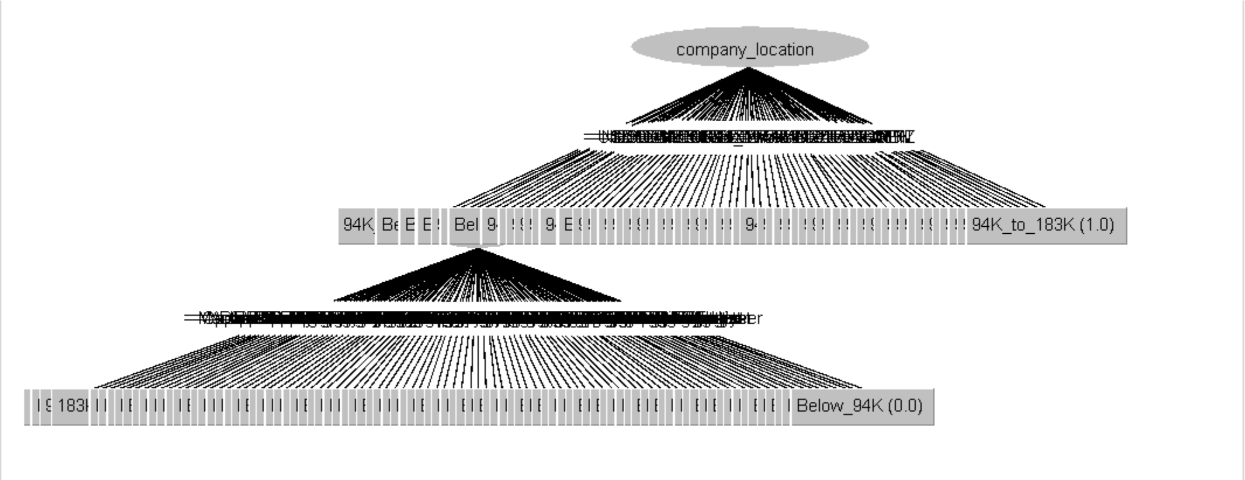
Result list (right-click for options)

```
=== Summary ===

Correctly Classified Instances         306               88.1844 %
Incorrectly Classified Instances        41               11.8156 %
Kappa statistic                          0.2947
Mean absolute error                      0.1235
Root mean squared error                  0.2678
Relative absolute error                 72.6706 %
Root relative squared error             92.5453 %
Total Number of Instances              347
```

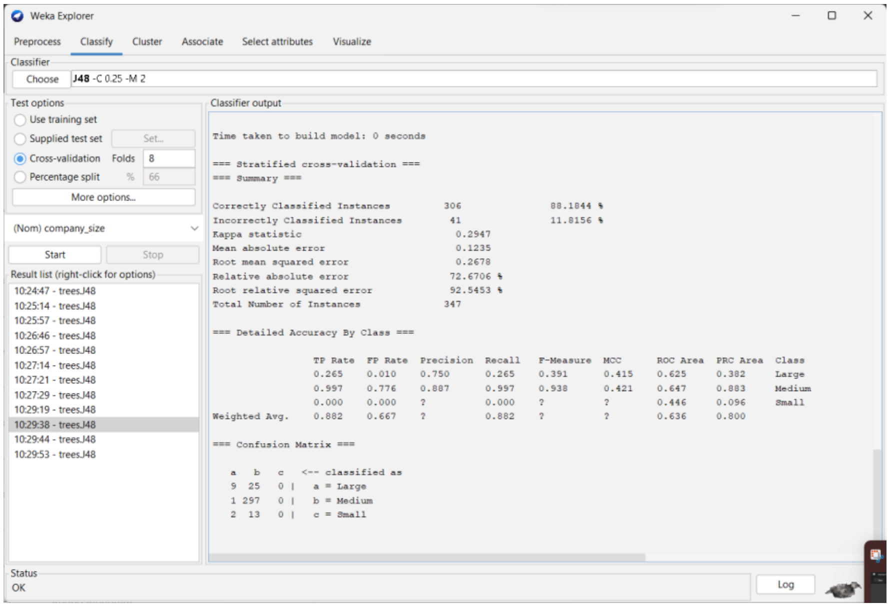
The above figures show how the decision tree algorithm applied with the J48 tree algorithm in weka. To choose the best instance we got an attribute as a Company location and number of folds we took as a 8 for best case scenario.

Best case scenario gives values Correctly Classified Instances, with 87.3199 % Incorrectly Classified Instances with 12.6801 %. Also we got 7 folds one and 9 folds both of those are showing most similar percentage like in range of 83.10%

After selecting the most suitable set as a set with 8 folds. It gives a tree with 75 leaves and a Decision tree with 77 tree size. When we visualize a tree coming like below figure.



And also we got summary reports like this with confusion matrices and stratified cross validations like below figure.



J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set
● Cross-validation Folds 8
○ Percentage split % 66

(Nom) company_size

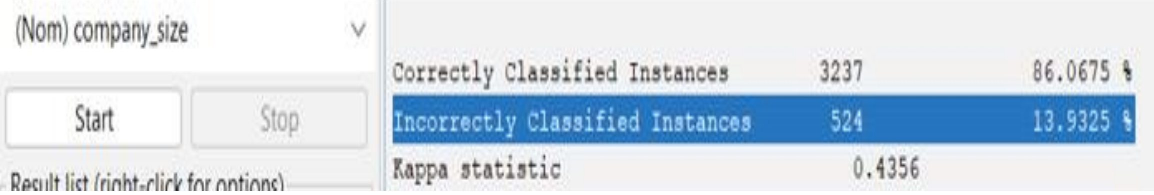
Result list (right-click for options)

```
10:24:47 - trees.J48
10:25:14 - trees.J48
10:25:57 - trees.J48
10:26:46 - trees.J48
10:26:57 - trees.J48
10:27:14 - trees.J48
10:27:21 - trees.J48
10:27:29 - trees.J48
10:29:19 - trees.J48
10:29:38 - trees.J48
10:29:44 - trees.J48
10:29:53 - trees.J48
```

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         306               88.1844 %
Incorrectly Classified Instances        41               11.8156 %
Kappa statistic                          0.2947
Mean absolute error                      0.1235
Root mean squared error                  0.2678
Relative absolute error                 72.6706 %
Root relative squared error             92.5453 %
Total Number of Instances              347

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.265    0.010    0.750      0.265    0.391      0.415    0.625     0.382     Large
                 0.997    0.776    0.887      0.997    0.938      0.421    0.647     0.883     Medium
                 0.000    0.000    ?          0.000    ?          ?        0.446     0.096     Small
Weighted Avg.    0.882    0.667    ?          0.882    ?          ?        0.636     0.800

=== Confusion Matrix ===

   a   b   c   <-- classified as
   9  25   0 |   a = Large
   1 297   0 |   b = Medium
   2  13   0 |   c = Small
```

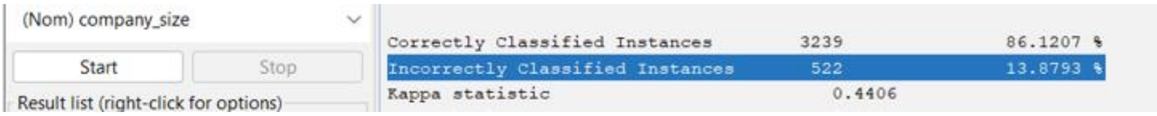
Status
OK

Naive Bayes Algorithm



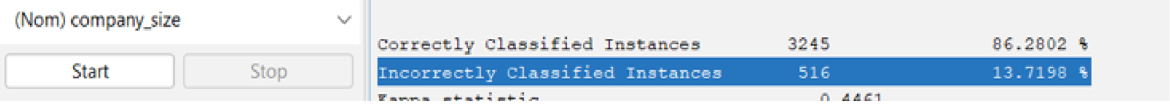
```
Correctly Classified Instances        3237               86.0675 %
Incorrectly Classified Instances       524               13.9325 %
Kappa statistic                          0.4356
```

Figure: Cross validation Fold 8



```
Correctly Classified Instances        3239               86.1207 %
Incorrectly Classified Instances       522               13.8793 %
Kappa statistic                          0.4406
```

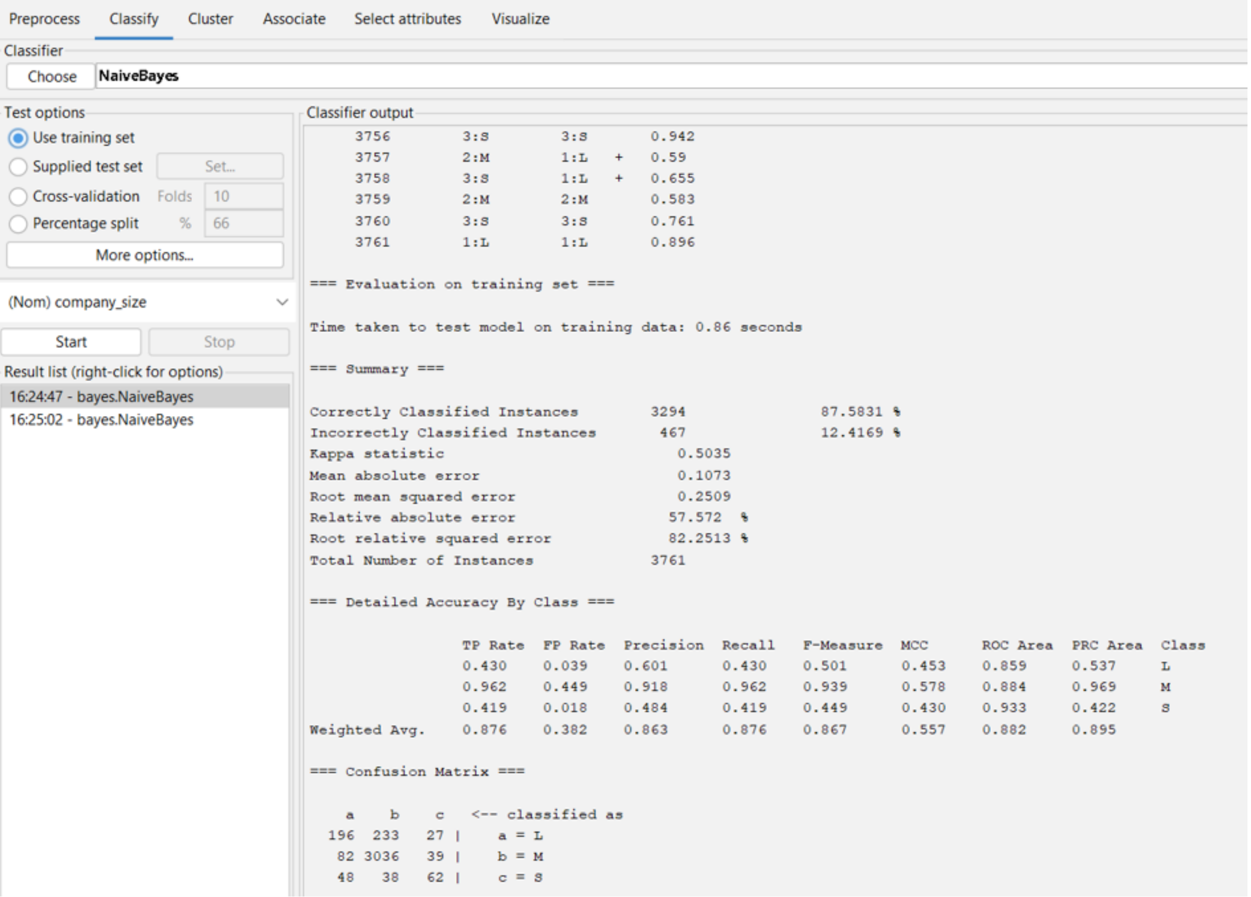
Figure: Cross validation Fold 6



```
Correctly Classified Instances        3245               86.2802 %
Incorrectly Classified Instances       516               13.7198 %
Kappa statistic                          0.4461
```

Figure: Cross validation Fold 7

Over here it is shown that the classification is done with the folds 6,7 and 8 using the Naive Bayes which is an algorithm to classify issues based on the Bayes Theorem. We have chosen the folds with 7 because the incorrect instance of the fold 6 is 13.8793%,fold 7 is 13.7198% and for fold 8 it is 13.9325%.Therefore as fold 7 is accurate we use it to test the dataset.



Classifier: NaiveBayes

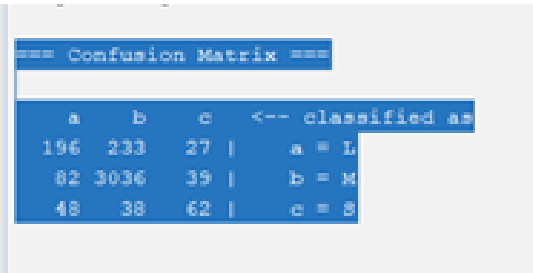
Test options:
- ● Use training set
- ○ Supplied test set
- ○ Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) company_size

Result list (right-click for options):
- 16:24:47 - bayes.NaiveBayes
- 16:25:02 - bayes.NaiveBayes

```
     3756        3:S        3:S        0.942
     3757        2:M        1:L    +   0.59
     3758        3:S        1:L    +   0.655
     3759        2:M        2:M        0.583
     3760        3:S        3:S        0.761
     3761        1:L        1:L        0.896

=== Evaluation on training set ===

Time taken to test model on training data: 0.86 seconds

=== Summary ===

Correctly Classified Instances        3294               87.5831 %
Incorrectly Classified Instances       467               12.4169 %
Kappa statistic                          0.5035
Mean absolute error                      0.1073
Root mean squared error                  0.2509
Relative absolute error                 57.572  %
Root relative squared error             82.2513 %
Total Number of Instances             3761

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.430    0.039    0.601      0.430    0.501      0.453    0.859     0.537     L
                 0.962    0.449    0.918      0.962    0.939      0.578    0.884     0.969     M
                 0.419    0.018    0.484      0.419    0.449      0.430    0.933     0.422     S
Weighted Avg.    0.876    0.382    0.863      0.876    0.867      0.557    0.882     0.895

=== Confusion Matrix ===

    a    b    c   <-- classified as
  196  233   27 |    a = L
   82 3036   39 |    b = M
   48   38   62 |    c = S
```

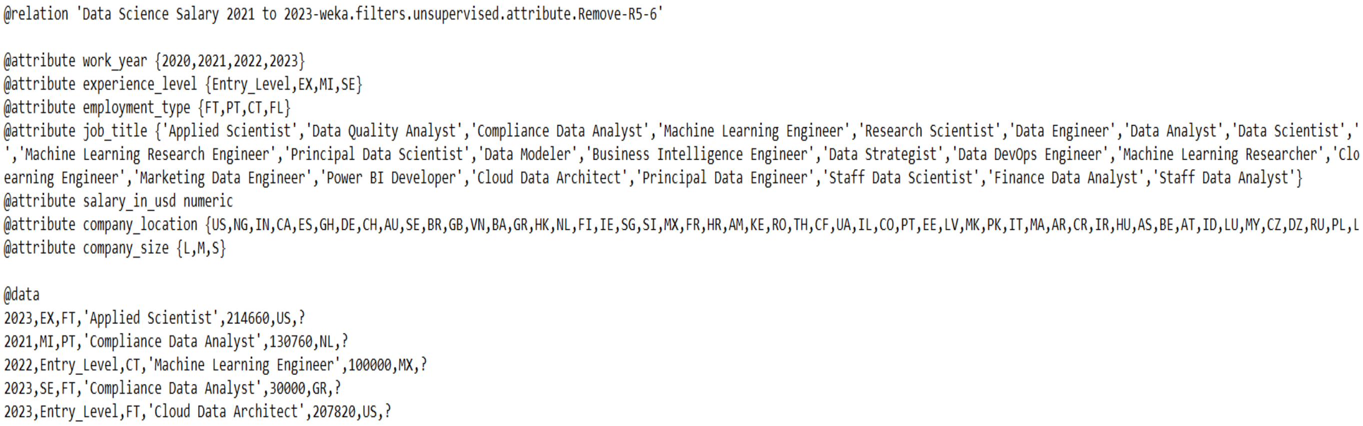
Using the training set we have tested the generalization of the errors. The correctly classified instance is 87.5831% and the Incorrect Classified Instance is 12.4169% hence we can state after viewing the mcc which is the correlation of the predicted class and the ground truth, this is a well connection relationship of accuracy as class L is 0.453 even though there are some utilities.



```
=== Confusion Matrix ===

    a    b    c   <-- classified as
  196  233   27 |    a = L
   82 3036   39 |    b = M
   48   38   62 |    c = S
```

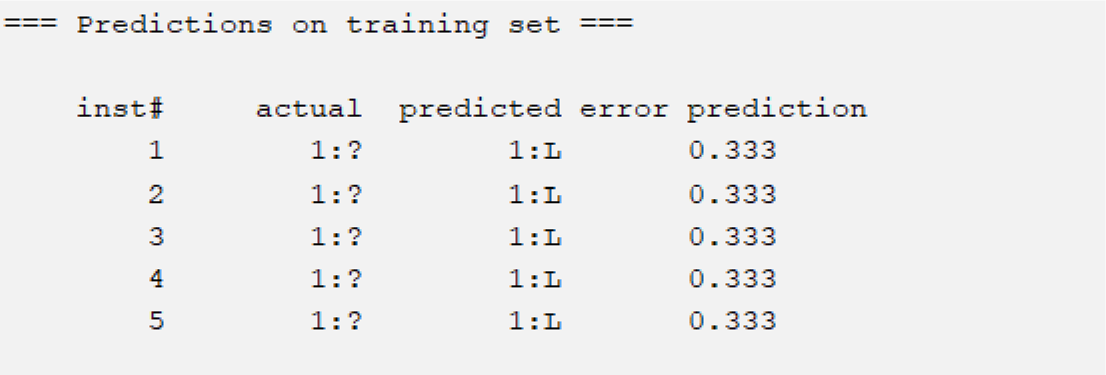
As you can see in the confusion matrix it has been categorized into 3 means the training set of the Naïve works as predicted.

We entered our own values to show the Naive Bayes.



```
@relation 'Data Science Salary 2021 to 2023-weka.filters.unsupervised.attribute.Remove-R5-6'

@attribute work_year {2020,2021,2022,2023}
@attribute experience_level {Entry_Level,EX,MI,SE}
@attribute employment_type {FT,PT,CT,FL}
@attribute job_title {'Applied Scientist','Data Quality Analyst','Compliance Data Analyst','Machine Learning Engineer','Research Scientist','Data Engineer','Data Analyst','Data Scientist','
','Machine Learning Research Engineer','Principal Data Scientist','Data Modeler','Business Intelligence Engineer','Data Strategist','Data DevOps Engineer','Machine Learning Researcher','Clo
earning Engineer','Marketing Data Engineer','Power BI Developer','Cloud Data Architect','Principal Data Engineer','Staff Data Scientist','Finance Data Analyst','Staff Data Analyst'}
@attribute salary_in_usd numeric
@attribute company_location {US,NG,IN,CA,ES,GH,DE,CH,AU,SE,BR,GB,VN,BA,GR,HK,NL,FI,IE,SG,SI,MX,FR,HR,AM,KE,RO,TH,CF,UA,IL,CO,PT,EE,LV,MK,PK,IT,MA,AR,CR,IR,HU,AS,BE,AT,ID,LU,MY,CZ,DZ,RU,PL,L
@attribute company_size {L,M,S}

@data
2023,EX,FT,'Applied Scientist',214660,US,?
2021,MI,PT,'Compliance Data Analyst',130760,NL,?
2022,Entry_Level,CT,'Machine Learning Engineer',100000,MX,?
2023,SE,FT,'Compliance Data Analyst',30000,GR,?
2023,Entry_Level,FT,'Cloud Data Architect',207820,US,?
```

Below are the predicted results.



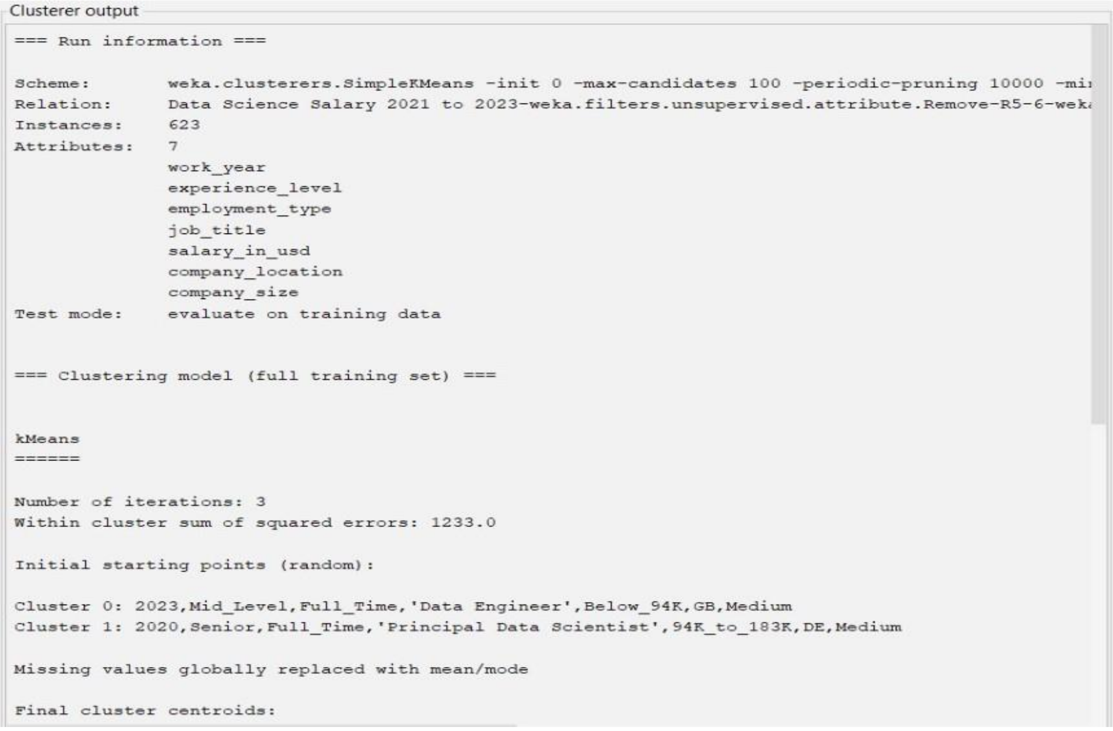
```
=== Predictions on training set ===

    inst#     actual  predicted error prediction
        1        1:?        1:L       0.333
        2        1:?        1:L       0.333
        3        1:?        1:L       0.333
        4        1:?        1:L       0.333
        5        1:?        1:L       0.333
```

Clustering

Clustering is an algorithm used in Data Analyzing to group similar data points together into categories or clusters based on their similarities or patterns within the data. Cluster algorithm the primary goal is to discover hidden structures or natural groupings within a dataset.

Simple K Means Algorithm

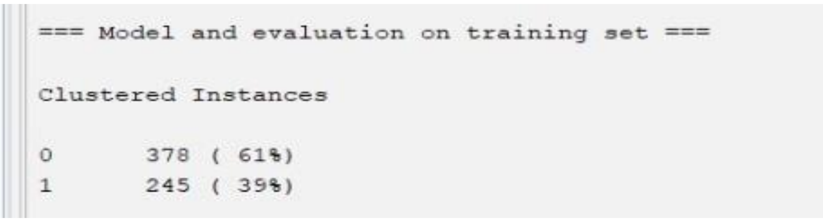
K Means algorithm is a popular clustering technique. It aims to partition a dataset into clusters, where each cluster is represented by its centroid.



Clusterer output

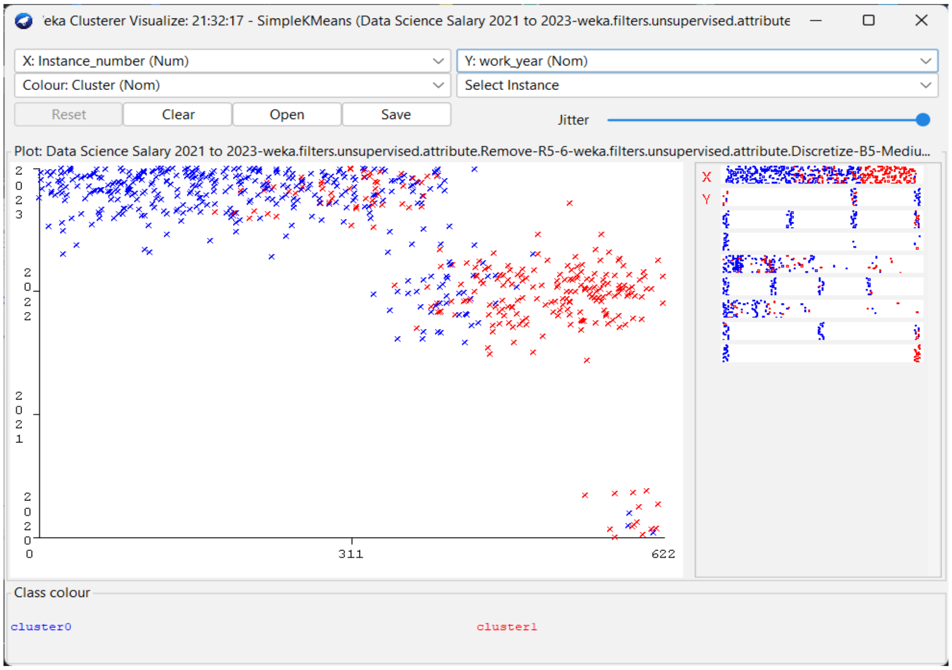
```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -mi
Relation:     Data Science Salary 2021 to 2023-weka.filters.unsupervised.attribute.Remove-R5-6-wek
Instances:    623
Attributes:   7
              work_year
              experience_level
              employment_type
              job_title
              salary_in_usd
              company_location
              company_size
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 1233.0

Initial starting points (random):

Cluster 0: 2023,Mid_Level,Full_Time,'Data Engineer',Below_94K,GB,Medium
Cluster 1: 2020,Senior,Full_Time,'Principal Data Scientist',94K_to_183K,DE,Medium

Missing values globally replaced with mean/mode

Final cluster centroids:
```

This is the output we received after applying the K Means algorithm to the dataset. According to the information shown in the figure there are 623 instances and 7 attributes.There are also 3 iterations while the sum of squared errors is 1233.0. The above scenario is done using 2 clusters and 10 seeds.



```
=== Model and evaluation on training set ===

Clustered Instances

0       378 ( 61%)
1       245 ( 39%)
```

There are 2 clustered instances. Cluster 0 has 378 rows of data which means 61% of the dataset. Cluster 1 has only 245 rows which means only 39% of the dataset.



'eka Clusterer Visualize: 21:32:17 - SimpleKMeans (Data Science Salary 2021 to 2023-weka.filters.unsupervised.attribute

X: Instance_number (Num)

Y: work_year (Nom)

Colour: Cluster (Nom)

Select Instance

Plot: Data Science Salary 2021 to 2023-weka.filters.unsupervised.attribute.Remove-R5-6-weka.filters.unsupervised.attribute.Discretize-B5-Mediu...

Class colour

cluster0

cluster1

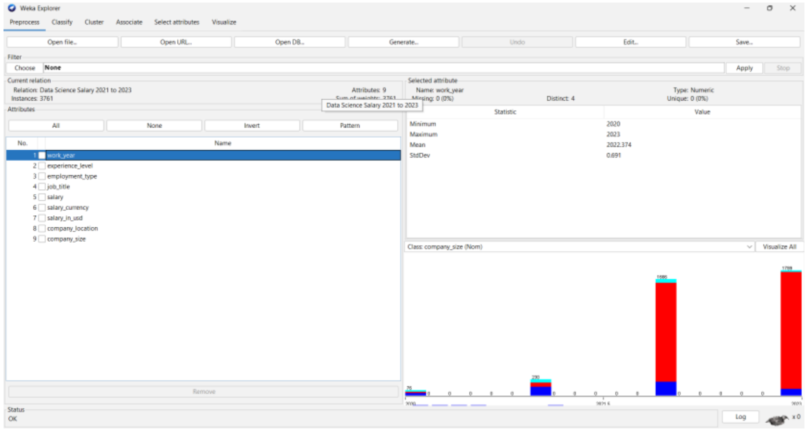
The above figure shows the visualization of the dataset which is clustered using K Means algorithm. It is clustered into two and cluster 0 is shown in Blue while cluster 1 is shown in red. The X axis shows the work_year attribute while the Y axis shows Instance_number attribute.

Association

Apriori Algorithm

The Apriori algorithm is one such algorithm in machine learning that finds out the probable associations and creates association rules. WEKA provides the implementation of the Apriori algorithm. You can define the minimum support and an acceptable confidence level while computing these rules

Select the Data Science Salary 2021-2023.csv database from the installation folder by clicking the Open file... button on the Preprocess tab of the WEKA explorer. After the data is loaded, the screen below will appear.



Filter: None

Current relation
Relation: Data Science Salary 2021 to 2023
Instances: 3761
Attributes: 9

No.	Name
1	work_year
2	experience_level
3	employment_type
4	job_title
5	salary
6	salary_currency
7	salary_in_usd
8	company_location
9	company_size

Selected attribute
Name: work_year Type: Numeric
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

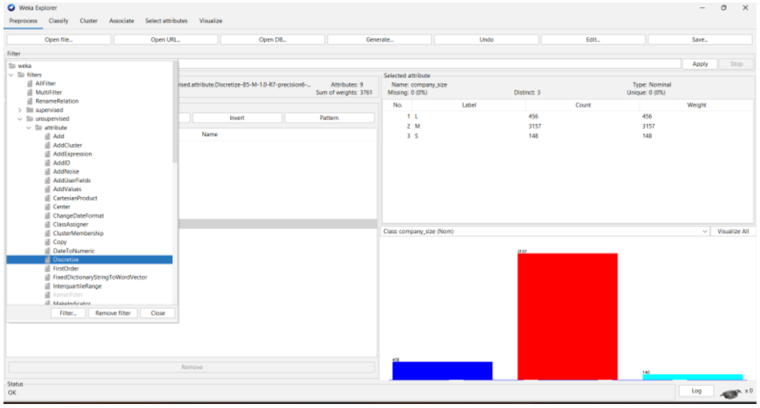
Statistic	Value
Minimum	2020
Maximum	2023
Mean	2022.374
StdDev	0.691

Class: company_size (Nom)

Status
OK

There are 9 attributes and 3761 instances in the database. You can see how challenging it would be to find relationships among this data . Fortunately, the Apriori algorithm automates this process.

Click on the Preprocess TAB and click on the Choose button.click choose-> unsupervised -> attribute -> Discretize



...ised.attribute.Discretize-B5-M-1.0-R7-precision6-... Attributes: 9 Sum of weights: 3761

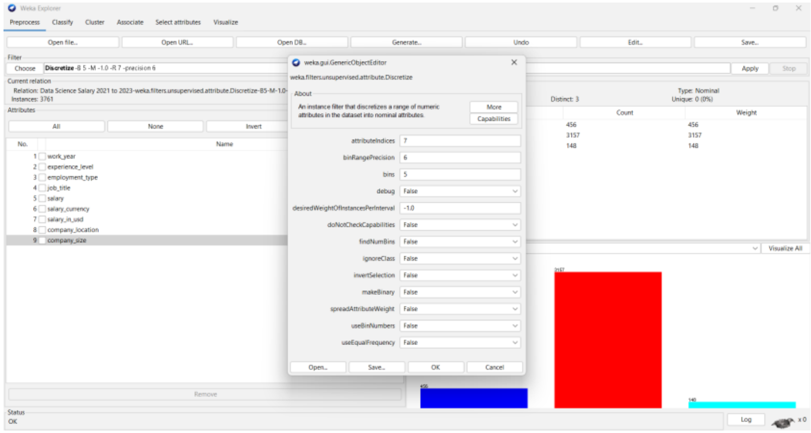
Selected attribute
Name: company_size Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	L	456	456
2	M	3157	3157
3	S	148	148

Class: company_size (Nom)

Status
OK

To set the parameters, click on Discretize name, a window will pop up as shown below that allows you to set the parameters.



Filter: **Discretize** -B 5 -M -1.0 -R 7 -precision 6

Current relation
Relation: Data Science Salary 2021 to 2023-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-
Instances: 3761

Attributes

No.	Name
1	□ work_year
2	□ experience_level
3	□ employment_type
4	□ job_title
5	□ salary
6	□ salary_currency
7	□ salary_in_usd
8	□ company_location
9	□ company_size

Distinct: 3
Type: Nominal
Unique: 0 (0%)

Count	Weight
456	456
3157	3157
148	148

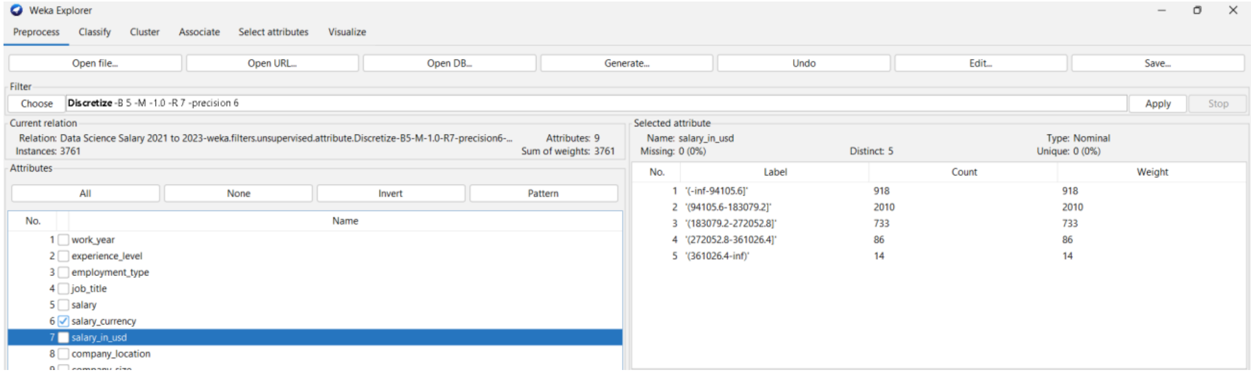
Status
OK

weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.Discretize

About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

attributeIndices	7
binRangePrecision	6
bins	5
debug	False
desiredWeightOfInstancesPerInterval	-1.0
doNotCheckCapabilities	False
findNumBins	False
ignoreClass	False
invertSelection	False
makeBinary	False
spreadAttributeWeight	False
useBinNumbers	False
useEqualFrequency	False

After adding the filter we need to set bins and divide numerical data into those bins.



Filter: **Discretize** -B 5 -M -1.0 -R 7 -precision 6

Current relation
Relation: Data Science Salary 2021 to 2023-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-R7-precision6-...
Instances: 3761
Attributes: 9
Sum of weights: 3761

Attributes

No.	Name
1	□ work_year
2	□ experience_level
3	□ employment_type
4	□ job_title
5	□ salary
6	☑ salary_currency
7	□ salary_in_usd
8	□ company_location
9	□ company_size

Selected attribute
Name: salary_in_usd
Missing: 0 (0%)
Distinct: 5
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-94105.6]'	918	918
2	'(94105.6-183079.2]'	2010	2010
3	'(183079.2-272052.8]'	733	733
4	'(272052.8-361026.4]'	86	86
5	'(361026.4-inf)'	14	14

This step is often necessary to transform continuous numerical data into discrete bins
Binning can simplify complex numerical data, making it easier for certain machine learning algorithms to handle.

Interpretation of Results and Conclusion

The analysis of data scientist salaries based on their job title, work experience, company location, and company size was analyzed by using the WEKA platform. We applied several machine learning algorithms such as, Clustering with J48 algorithm to get decision trees, Naive Bayes algorithm to predict missing values in dataset, Simple K-Means algorithm to identify similar groups and class that associated with our data, and Association when we preprocess the data, were employed to gain insights into the salary trends within the data science field.

Clustering with J48, Naive Bayes, Simple K-Means, and Association are machine learning algorithms that allow us to define dataset patterns and relationships. To understand the salary trends these algorithms were useful in grouping similar data points and discovering hidden associations, which were valuable.

The analysis of data scientist salaries based on their job title, work experience, company location, and company size provides valuable insights for both job seekers and employers in the high demanding data science field. Our analysis key takeaways include the importance of gaining experience and progressing in an individual's career to earn higher salaries in the data scientist fields, the influence of most suitable company locations and company size on compensation, and the utility of machine learning algorithms for exploring data scientists' complex salary data with their skills and company. This information can guide individuals in making informed career choices and choose most suitable organizations in making competitive salary offers to attract and retain top data scientist experiences. Additionally, continuous monitoring and analysis of salary trends in the industry are essential to stay competitive and keep track in the rapidly evolving field of data science.

Google Drive link :- https://drive.google.com/drive/folders/155GSdWREZZVfE-_Tv5FPRSfOvdlIfl5R?usp=sharing