Ahsanullah University of Science & Technology

Department of Computer Science & Engineering

Course No : **CSE4142**

Course Title : **Data Warehousing and Mining Lab**

Assignment No : **02**

Date of Submission : **26.05.2024**

Submitted To : **Mr. Saha Reno & Mr. Raiyan Jahangir**

Submitted By-

Name: Nabila Rahman

ID: 20200204065

Section: B1

(i) Create a Custom Dataset Which Will Have 5 Attributes: 2 Numeric, 2 Nominal & 1 Class (3 Class Values)

Ans:

```
@relation EmployeeReviews

@attribute years_experience real
@attribute salary real
@attribute department {HR, IT, Sales}
@attribute education_level {High_School, Bachelors, Masters}
@attribute performance {Excellent, Good, Poor}
```

(ii) Create 20 Instances of That Dataset Which Should Have Some Missing Values inside Any 2 Attributes + Make 10 Instances of 1st Class Value, 6 Instances of 2nd Class Value & Rest of the Instances Should be of 3rd Class Value

Ans:

```
@data
5.2, 70.5, IT, Bachelors, Good
3.1, 50.0, HR, High_School, Excellent
7.5, 85.3, Sales, Masters, Excellent
4.0, ?, IT, Bachelors, Good
8.3, 95.2, Sales, Masters, Excellent
2.7, 45.1, HR, High_School, Poor
10.0, 110.0, IT, Masters, Excellent
6.5, 75.8, Sales, Bachelors, Good
4.8, ?, HR, Bachelors, Poor
?, 80.2, IT, Masters, Excellent
5.7, 65.3, Sales, High_School, Good
9.2, 90.5, HR, Masters, Excellent
3.9, 55.0, IT, Bachelors, Excellent
6.3, ?, Sales, Masters, Good
7.1, 78.6, HR, Bachelors, Excellent
4.5, 62.0, IT, High_School, Poor
?, 68.4, Sales, Bachelors, Good
8.8, 88.0, HR, Masters, Excellent
2.5, 40.7, IT, High_School, Poor
7.0, 82.1, Sales, Masters, Excellent
```

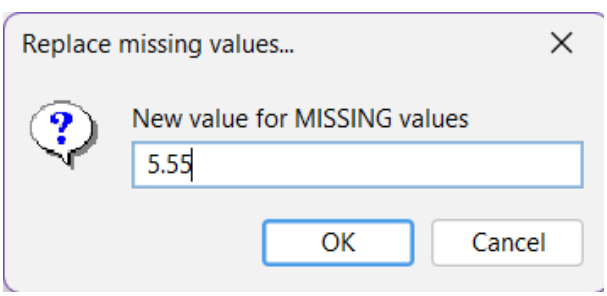
(iii) Using Preprocessing Tab, Fill-Out Those Missing Values using Your Preferred Values

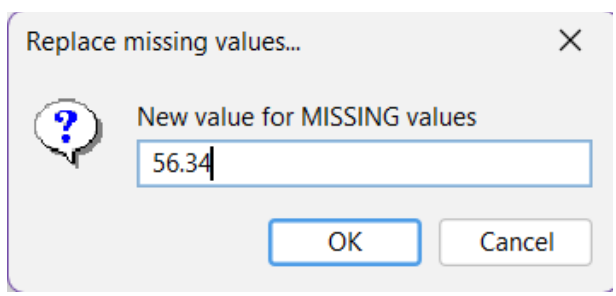
Ans: Before filling out missing values-

Relation: EmployeeReviews

No.	1: years_experience Numeric	2: salary Numeric	3: department Nominal	4: education_level Nominal	5: **performance** Nominal
1	5.2	70.5	IT	Bachelors	Good
2	3.1	50.0	HR	High_School	Excellent
3	7.5	85.3	Sales	Masters	Excellent
4	4.0		IT	Bachelors	Good
5	8.3	95.2	Sales	Masters	Excellent
6	2.7	45.1	HR	High_School	Poor
7	10.0	110.0	IT	Masters	Excellent
8	6.5	75.8	Sales	Bachelors	Good
9	4.8		HR	Bachelors	Poor
10		80.2	IT	Masters	Excellent
11	5.7	65.3	Sales	High_School	Good
12	9.2	90.5	HR	Masters	Excellent
13	3.9	55.0	IT	Bachelors	Excellent
14	6.3		Sales	Masters	Good
15	7.1	78.6	HR	Bachelors	Excellent
16	4.5	62.0	IT	High_School	Poor
17		68.4	Sales	Bachelors	Good
18	8.8	88.0	HR	Masters	Excellent
19	2.5	40.7	IT	High_School	Poor
20	7.0	82.1	Sales	Masters	Excellent

The missing values in the "years_experience" column have been replaced with the mean value of 5.55, and the missing values in the "salary" column have been replaced with the mean value of 56.34. After filling out missing values-



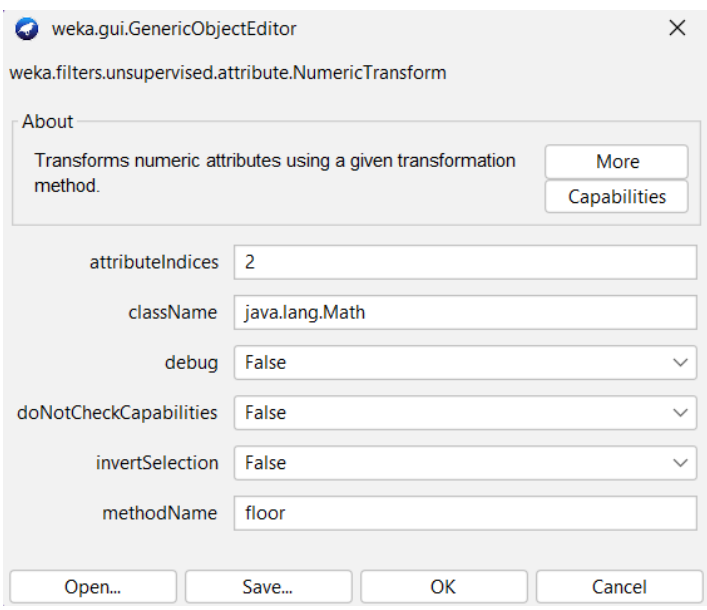


Relation: EmployeeReviews

No.	1: years_experience Numeric	2: salary Numeric	3: department Nominal	4: education_level Nominal	5: **performance** Nominal
1	5.2	70.5	IT	Bachelors	Good
2	3.1	50.0	HR	High_School	Poor
3	7.5	85.3	Sales	Masters	Excellent
4	4.0	5.55	IT	Bachelors	Good
5	8.3	95.2	Sales	Masters	Excellent
6	2.7	45.1	HR	High_School	Poor
7	10.0	110.0	IT	Masters	Excellent
8	6.5	75.8	Sales	Bachelors	Good
9	4.8	5.55	HR	Bachelors	Poor
10	56.34	80.2	IT	Masters	Excellent
11	5.7	65.3	Sales	High_School	Good
12	9.2	90.5	HR	Masters	Excellent
13	3.9	55.0	IT	Bachelors	Poor
14	6.3	5.55	Sales	Masters	Good
15	7.1	78.6	HR	Bachelors	Excellent
16	4.5	62.0	IT	High_School	Poor
17	56.34	68.4	Sales	Bachelors	Good
18	8.8	88.0	HR	Masters	Excellent
19	2.5	40.7	IT	High_School	Poor
20	7.0	82.1	Sales	Masters	Excellent

(iv) Convert Any 1 Real Attribute's Values from Float to Integers (which is less than or equal to the original value)

Ans:



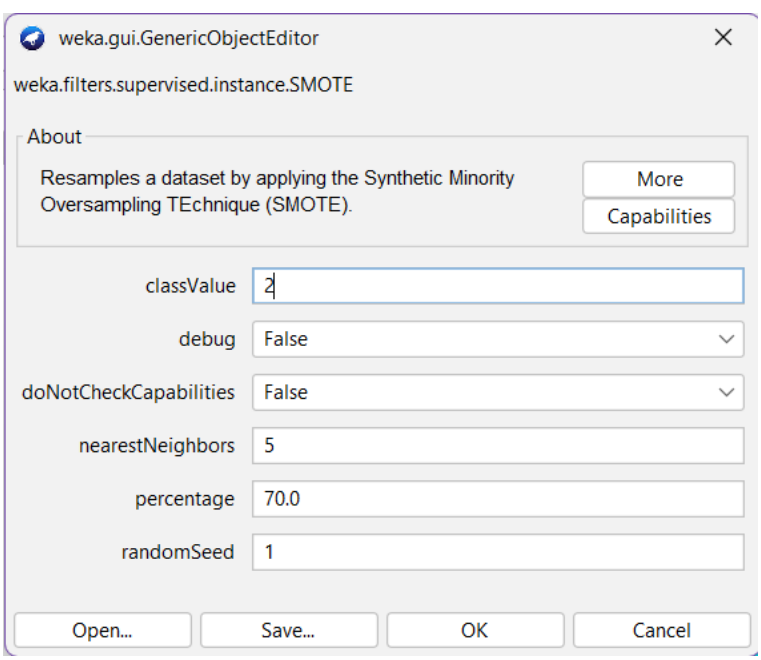
Transforming the ‘salary’ column to integer values by flooring them to get less than or equal to the original value. After converting-

Relation: EmployeeReviews-weka.filters.unsupervised.attribute.NumericTransform-R2-Cjava.lang.Math-Mfloor

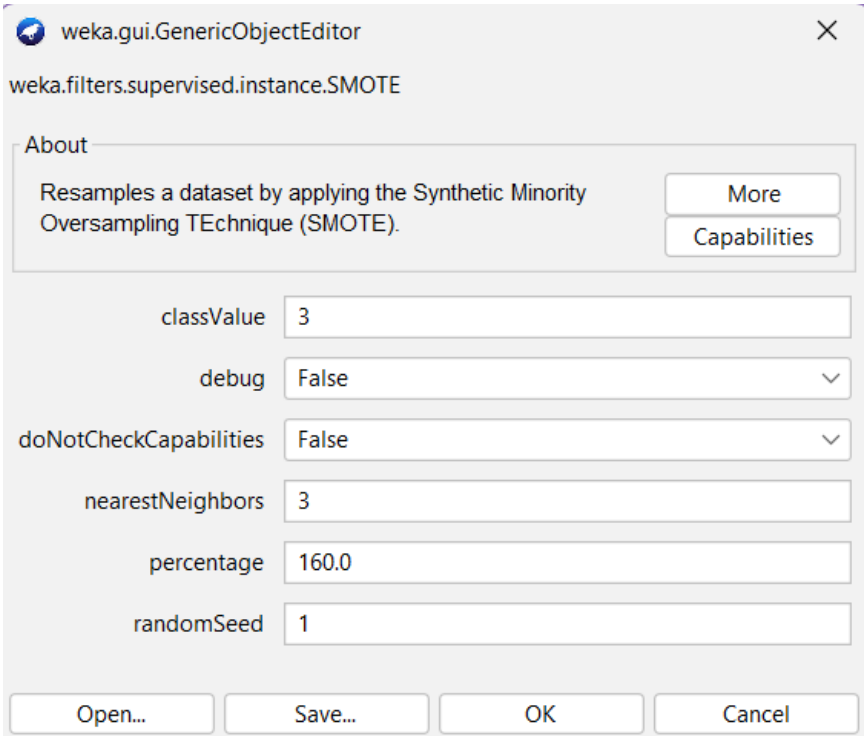
No.	1: years_experience Numeric	2: salary Numeric	3: department Nominal	4: education_level Nominal	5: **performance** Nominal
1	5.2	70.0	IT	Bachelors	Good
2	3.1	50.0	HR	High_School	Excellent
3	7.5	85.0	Sales	Masters	Excellent
4	4.0	56.0	IT	Bachelors	Good
5	8.3	95.0	Sales	Masters	Excellent
6	2.7	45.0	HR	High_School	Poor
7	10.0	110.0	IT	Masters	Excellent
8	6.5	75.0	Sales	Bachelors	Good
9	4.8	56.0	HR	Bachelors	Poor
10	5.55	80.0	IT	Masters	Excellent
11	5.7	65.0	Sales	High_School	Good
12	9.2	90.0	HR	Masters	Excellent
13	3.9	55.0	IT	Bachelors	Excellent
14	6.3	56.0	Sales	Masters	Good
15	7.1	78.0	HR	Bachelors	Excellent
16	4.5	62.0	IT	High_School	Poor
17	5.55	68.0	Sales	Bachelors	Good
18	8.8	88.0	HR	Masters	Excellent
19	2.5	40.0	IT	High_School	Poor
20	7.0	82.0	Sales	Masters	Excellent

(v) Fix the Class Imbalance Problem for the 2nd and 3rd Class by Making the Number of Instances for 2nd Class and 3rd Class Equal as the Number of Instances for 1st Class (10)

Ans:

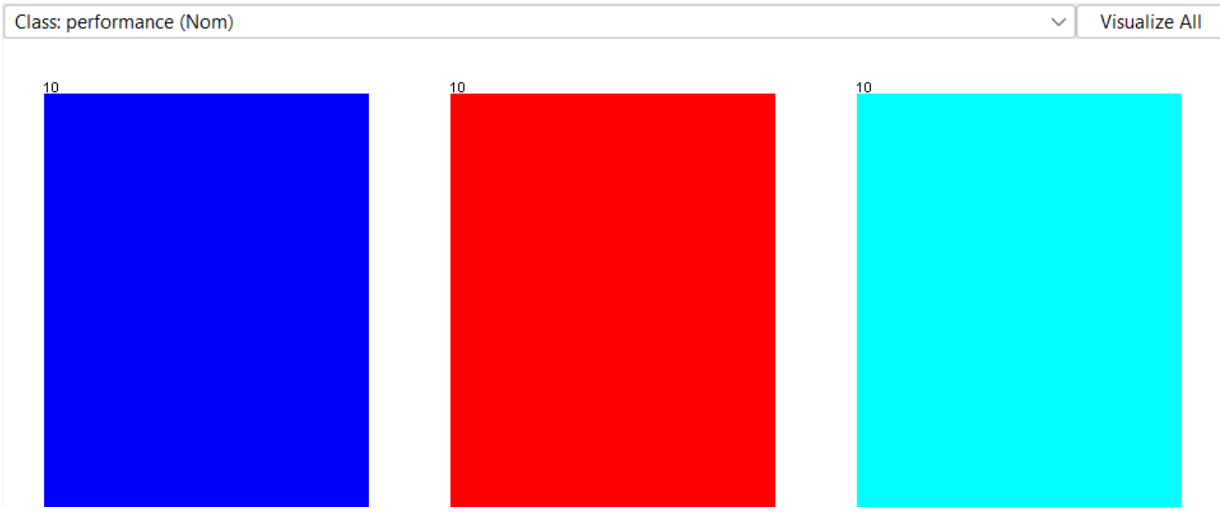


11	5.2	70.0	IT	Bachelors	Good
12	4.0	56.0	IT	Bachelors	Good
13	6.5	75.0	Sales	Bachelors	Good
14	5.7	65.0	Sales	High_School	Good
15	6.3	56.0	Sales	Masters	Good
16	5.55	68.0	Sales	Bachelors	Good
17	5.565988761555462	56.4514...	Sales	Bachelors	Good
18	4.381840503353804	59.0319...	Sales	Bachelors	Good
19	4.132012705842015...	57.1954...	Sales	Bachelors	Good
20	5.751622274391192	57.4340...	Sales	Bachelors	Good



21	2.7	45.0	Sales	High_School	Poor
22	4.8	56.0	HR	Bachelors	Poor
23	4.5	62.0	IT	High_School	Poor
24	2.5	40.0	IT	High_School	Poor
25	2.658457031738057	43.3364...	IT	High_School	Poor
26	4.192553217900486	46.6800...	IT	High_School	Poor
27	4.518040383365427...	61.6831...	IT	High_School	Poor
28	4.646126212465052	56.5162...	IT	High_School	Poor
29	3.088114064008073...	51.1426...	IT	High_School	Poor
30	2.607851933749127	42.9167...	IT	High_School	Poor

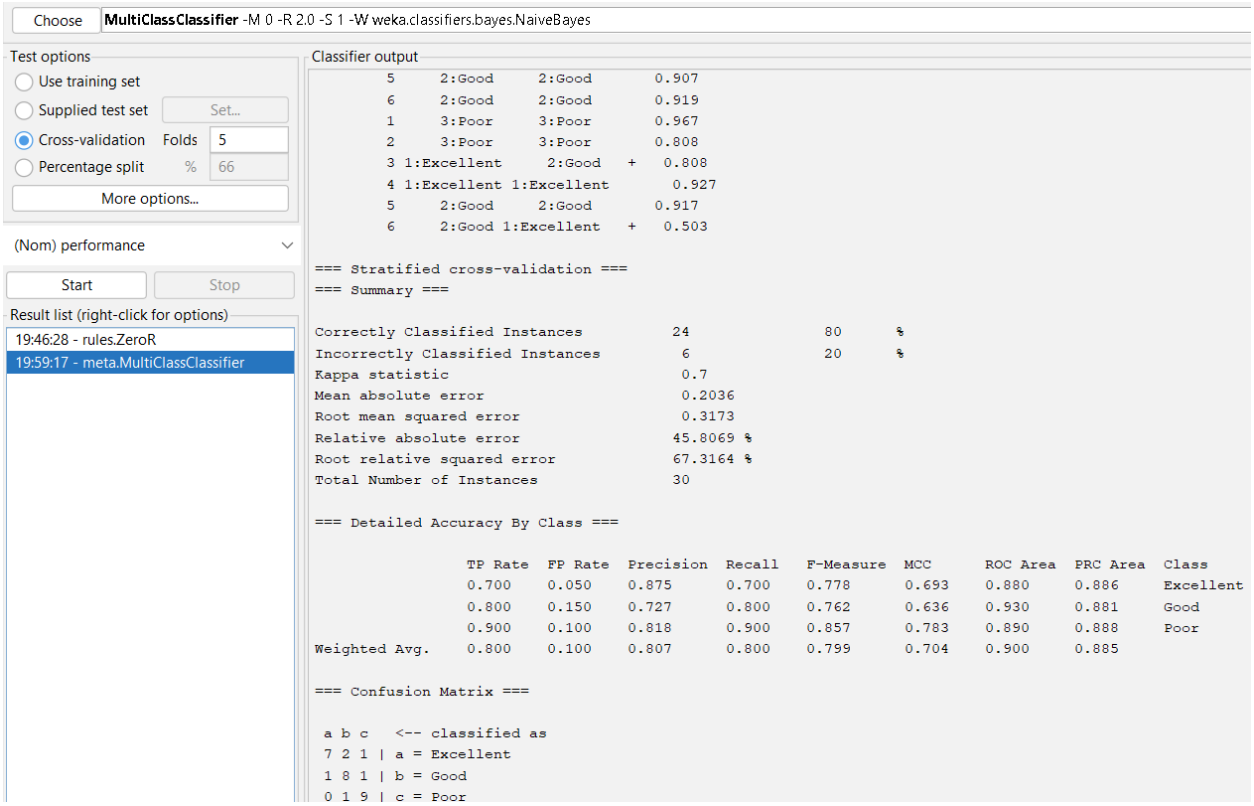
After class balancing, the 2nd and 3rd classes have 10 instances each just like the 1st class.



(vi) Apply Any Classification Algorithm on the Modified Dataset (Use 5-Fold Cross Validation)

Ans: For classification first, the dataset has been normalized(discretized)-

No.	1: years_experience Nominal	2: salary Nominal	3: department Nominal	4: education_level Nominal	5: **performance** Nominal
1	'(-inf-3.25]'	'(-inf-4...	IT	High_School	Poor
2	'(-inf-3.25]'	'(-inf-4...	IT	High_School	Poor
3	'(-inf-3.25]'	'(-inf-4...	IT	High_School	Poor
4	'(-inf-3.25]'	'(-inf-4...	Sales	High_School	Poor
5	'(4-4.75]'	'(-inf-4...	IT	High_School	Poor
6	'(-inf-3.25]'	'(47-54]'	HR	High_School	Excellent
7	'(-inf-3.25]'	'(47-54]'	IT	High_School	Poor
8	'(3.25-4]'	'(54-61]'	IT	Bachelors	Excellent
9	'(3.25-4]'	'(54-61]'	IT	Bachelors	Good
10	'(4-4.75]'	'(54-61]'	IT	High_School	Poor
11	'(4-4.75]'	'(54-61]'	Sales	Bachelors	Good
12	'(4-4.75]'	'(54-61]'	Sales	Bachelors	Good



Here, a multiclass classifier using the Naive Bayes model has been used to classify the datasets using 5-fold cross validation. It correctly classified 80% data and misclassified 20% data.