KNOWLEDGE ENGINEERING LAB (CSE 4.1.7)

5. Performing data preprocessing in Weka – Part1

Study Unsupervised Attribute Filters such as "ReplaceMissingValues" to replace missing values in the given dataset, "Add" to add the new attribute Average, 'Discretize' to discretize the attributes into bins. Explore Normalize and Standardize options on a dataset with numerical attributes.

Finding missing values in the dataset:

- 1. Launch Weka-> click on the tab Explorer
- 2. Load a dataset. (Click on "Open File" & locate the datafile)
- Click on PreProcess tab & then look at your lower R.H.S. bottom window click on drop down arrow and choose "No Class"
- 4. Click on "Edit" tab, a new window opens up that will show you the loaded datafile. By looking at your dataset you can also find out if there are missing values in it or not. Also please note the attribute types on the column header. It would either be 'nominal' or 'numeric'.

If your data has missing values then its best to clean it first before you apply any forms of algorithm to it. Please look below at Figure, you will see the highlighted fields are blank that means the data at hand is dirty and it first needs to be cleaned.

	-	Travalational Kirt.	307*	ACT/A	1.0)	
Vertor	REGAT	-latence Dy	DT*	DT4A	1.0	1.0(CDRAEC7
Equation	REBAT	mission	abries .	354	1.0	2.0 CONKECT
Explan Further	HENT_HEGG			FLUEDS14	1.0	2.0H9NT
Espation	RES.A.T	Vectors	160*	VECTR	1.0	1.0 CORRECT
Esplation	REDAT	Rotatorial King	30.*	10134	1.0	2.0 INCORRECT
Equation	RESIAT	Translational His.	87*	KT10A	1.0	1.6 CDRAECY
Equation	REBAT	rectors	WDC*	VECID	1.0	1.0 CORFICT
Weiter	RESULT	Translational Kirv.	307*	KTDA	1.0	12:0 0044807
Explain-Further	HENT, MOG	Circular Motors	ROTS*	ADTS3A	1.0	2.094047

O Weis Epilerer		- ш х
Proprietes Daniel Chatter Annual Great Articles Weater		
Open Re. Dever URL Dever DR. Deve	ene Vola	DA Dec
film:		
Chonse Stars		(Asity) =
Carriers reliation	Selected attribute	
Teamer astase Athans 32 Interve 401 Service apprendix	Name prod Colour 1	Figure Hardware Managere D (29%)
Antibases M None Issue Parason No. None Issue Parason 1 Inform Issue Issue 2 Statistical Issue Issue 4 Statistical Issue Issue	No. Land Dave + Heart H + Seart H - grown dip	Wege 148 110.0 amo
N 200 mm T Among the property of	(Gaus class (Row)	

Replace Missing Values:

To clean the data, you apply "Filters" to it. Generally the data will be missing with values, so the filter to apply is "ReplaceMissingWithUserConstant" (the filter choice may vary according to our need). Click on Choose button below

- Filters
 - o Unsupervised
 - Attribute
 - ReplaceMissingWithUserConstant

Open Ne Open DR Open DR Open Ne Unso Multic Control Torre Cheese ReplaceMissingWith/AserConstant Africal PLP You have Apple Cheese ReplaceMissingWith/AserConstant Africal PLP You have Apple Withoute Apple Apple Apple Apple Withoute Apple Apple Apple Name Torre Puttorn Torre Torre Name Torre Puttorn Apple Apple Name Torre Puttorn Apple Apple Name Torre Puttorn Apple Apple Name Torre Apple Apple </th <th>9 Weisz Explorer Pharmicken Gannely: Disoler Associate Select attributes Vination</th> <th></th> <th></th> <th>dis al</th>	9 Weisz Explorer Pharmicken Gannely: Disoler Associate Select attributes Vination			dis al
Intermet telefor Separate statistical Affensional Affensiona	Open Ne. Open UNL Open Dit. Dan	eraki Und	Nich	Kiert.
Crease Replace/Ministry/With/Aser/Constant: A frame bail // MALL - 01 P. Y. Yuyu/ MR cont/11th formatal Control Association Response staffinger Marchane: 36 Marchane: 36 Marchane: 36 Marchane: 683 Marchane: 7 Marchane: 36 Marchane: 36 Marchane: 7 Marchane: 36 Marchane: 36 Marchane: 36 Marchane: 8 Marchane: 96 Marchane: 36 Marchane: 36 Marchane: 8 Marchane: 96 Marchane: 36 Marchane: 36 Marchane: 8 Marchane: 96 Marchane: 36 Marchane: 36 Marchane: 8 Marchane: 36 <t< th=""><th>Chev</th><th></th><th>/ ···· Ye</th><th></th></t<>	Chev		/ ···· Ye	
Arrent relation Selected attribute Relation: styphest Instance: 683 Arridues: 36 Them of weights 685 Arridue: 683 Them of weights 685 Arridue: 7 Usage 9 (2%) Arridue: 9 (2%) The of weight Arridue: 9 (2%)	Choose ReplaceMissingWithUserConstant Afronist ATMUL # D-P Sympleter	/T\th-firmercan*	John Street	10 Acces 11
Name Antiduces (0) Name Inter of weights (0) Name Inter of	arrent relation	Selected attribute		
None Pattern No Lobal Court Woogh All All 20.0 70.0 20.0 70.0 20.0	Relation soldean Attribute 35 Instances 483 Team of weights 480	Name mas Menory 1 (0%)	Distant 7	Type Normal Grouper 0 (D%)
All Name Hweet Pattern No Name Hweet Hallen No Name Hweet Hallen Name Hweet Hallen Hallen Name Hallen Hallen Hallen Hallen	Arthules	No Lider	Court	Weight
All Name Have Pattern 3 mage 74 75.0 Yes Hare Hare </td <td></td> <td>t ipri</td> <td>.28</td> <td>28.0</td>		t ipri	.28	28.0
Re Norm File State State <t< td=""><td>and the second s</td><td>3 mag</td><td>78</td><td>75/0</td></t<>	and the second s	3 mag	78	75/0
Tex Name 4 M 100 110	A6 Note I Hadwin	3 (29	81	90.0
New 6 Agent 131 191.0 2 paint manifi 148.0 148.0 148.0 3 park 0 148.0 148.0 148.0 4 park 0 148.0 148.0 148.0 5 park 0 148.0 148.0 148.0 4 park 0 148.0 148.0 148.0 5 paint park 0 100 park 100 100 <td></td> <td>-4. p/y</td> <td>110</td> <td>118:0</td>		-4. p/y	110	118:0
2 jorder mang 347 148.0 2 jorder mang 7 oct.dar 30 3 jorder mang 1000 4 strap tan 1000 5 tan 1000 4 strap tan 1000 5 tan 1000 6 strap tan 1000 6 strap tan 1000 7 strap tan 1000 8 strap tan 1000 8 strap tan 1000 8 strap tan 1000 8 strap tan 1000 9 strap tan 1000 10 strap tan 1000 11 strap tan 1000 12 strap tan 1000 13 strap tan 1000 14 strap tan 1000 15 strap tan 1000 16 strap tan 1000 17 strap tan 1000 16 strap tan 1000 17 strap tan 1000 18 strap tan 1000 19 strap tan 1000	Ne Nere	6 august	231	131.0
2 percent 3 percent 4 percent 5 fail 6 fail 7 max-stampad 8 serverity 9 mechanism 10 pervention 11 pixel-person 12 bewain 13 sethpoint-indep 14 bestippet-indep 15 kedi-persod 16 kedi-persod 17 kedi-persod 18 sethpoint-indep 19 sethpoint-indep 10 persod 11 pixel-persod 12 bewain 13 sethpoint-indep 14 bestippet-indep 15 kedi-persod 16 kedi-persod 17 kedi-persod 18 sethpoint indep 19 sethpoint indep 10 sethpoint indep 11 pixet-persod 12 bewain 13 bestippet-indep 14 bestippet-indep 15 kedi-persod 16 kedi-persod 17 bestipet	A DESCRIPTION OF A DESC	6 september	149	148.0
3 presp 3 presp 4 presp 5 twill 7 servedrunged 8 servedrunged 8 servedrunged 9 secol hard 10 gemmatkolo 11 gilder secol hard 12 kankpesis nalo 14 wakpesis nalo 15 ked skenod 16 ked skenod 17 gilder sec 18 secol mark	2 plant more	7 octuber	95	90.0
4 Strip 6 top-trail 7 sea-damaged 8 seetly 9 seetly 9 seetly 10 periodion 11 plane periodion 12 seetly 13 self-periodic 14 self-periodic 15 self-periodic 16 bodi denaid 17 settered	3 piecio			
b set b set b set c set d set	4 three			
1 Second	5 tet			
1 usis damagai 1 <t< td=""><td>5 stopted</td><td></td><td></td><td>100</td></t<>	5 stopted			100
# secondly #	1 ana-damaped	China: class (Nem)		Wikielde
P press/arr 10 primit granth 11 primit granth 12 kannan 13 kalisponith 14 kalisponith 15 kalisponith 16 kalisponith 17 kalisponith 18 kalisponith 19 kalisponith 10 kalisponith 11 kalisponith 12 kalisponith 13 kalisponith 14 kalisponith 15 kalisponith 16 kalisponith 17 kalisponith 18 kalisponith 19 kalisponith 10 kalisponith 11 kalisponith 12 kalisponith 13 kalisponith 14 kalisponith 15 kalisponith 16 kalisponith 17 kalisponith 18 kalisponith 19 kalisponith 10 kalisponith 11 kalisponith 12 kalisponith 13 kalisponith 14 kalisponith 15 k	B severity			
11 previouslow 12 second 13 second 14 setupcet-tala 15 setupcet-tala 16 setupcet-tala 17 setupcet-tala 18 setupcet-tala 19 setupcet-tala 10 setupcet-tala 11 setupcet-tala 12 setupcet-tala 13 setupcet-tala 14 setupcet-tala 15 setupcet-tala 16 setupcet-tala 17 ket/radi 18 setupcet-tala 19 setupcet-tala 10 setupcet-tala 11 setupcet-tala 12 setupcet-tala 13 setupcet-tala 14 setupcet-tala 15 setupcet-tala 16 setupcet-tala 17 ket/radi 18 setupcet-tala 19 setupcet-tala 10 setupcet-tala 11 setupcet-tala 12 setupcet-tala 13 setupcet-tala 14 setupcet-tala 15 setupcet-tala 16 setupceta	P is seed and	1		
11 Files granth 12 News 13 Kentopolo faite 14 Kentopolo faite 15 Kentopolo faite 16 Kentopolo faite 17 Kentopolo faite 18 Kentopolo faite 19 Kentopolo faite 10 Kentopolo faite 11 Kentopolo faite 12 Kentopolo faite 13 Kentopolo faite 14 Kentopolo faite 15 Kentopolo faite 16 Kentopolo faite 17 Kentopolo faite 18 Kentopolo faite 19 Kentopolo faite 10 Kentopolo faite 11 Kentopolo faite 12 Kentopolo faite 13 Kentopolo faite 14 Kentopolo faite 15 Kentopolo faite 16 Kentopolo faite 17 Kentopolo faite 18 Kentopolo faite 19 Kentopolo faite 10 Kentopolo faite 11 Kentopolo faite 12 Kentopolo faite 14 Kentopolo faite 15 Kentopolo faite 16<	10 germination			145
13 water too 14 water too 15 water too 16 water too 17 water too 18 water too 19 water too	11 statet-profit			10
16 hetpost-sko 16 hetpost-sko 16 hetpost-sko 16 hetpost-sko 16 hetpost-sko 16 hetpost-sko 17 hetpost-sko 18 hetpost-sko 19 hetpost-sko 10 hetpost-sko 11 hetpost-sko 12 hetpost-sko	12 heres		118	
	13 ketrpote-taio			
	14 estapots-marp		17	
	15 🔜 Kalsput-size			
	16 kal skread	the second se		
	17 is hul-mat			
	48. 1 + 4 + 11			
	Comment of Comment			
Male				

A good choice for replacing missing numeric values is to give it values like -1 or 0 and for string values it could be NULL.

1		
Replaces all missing values for	r nominal, string, numeric and	More
date duributes in the uplaser w	nu use suppled constant value	Capabilities
attributes	first-last	
dateFormat	yyyy-MM-dd'T'HH mm ss	
dateReplacementValue		
debug	False	1
doNotCheckGapabilities	False	1
ignoreClass	False	1
nominalStringReplacementValue	NULL	
numeric Depleter emergi (sh. e		

Click on Ok and then Apply

Now, all nominal missing values are replaced by NULL and numeric values with 0.

"Add" to add the new attribute Average:

Let us assume we have dataset features X and Y in the given dataset. Requirement is to add another feature that is the average of X and Y.

- Open the CSV file
- Here X is considered as a1 and Y is considered as a2
- Click on choose button select Add Expression Filter (Weka->Filters->Unsupervised->Attribute)
- Click on the text box next to the choose button where AddExpression is appearing.
- Type the expression (a1+a2)/2 in expression text box
- Click on ok and then apply after the choose.

Weka Explorer	- 0 ×
Proposess Clansify Chester Associate Select attributes Visualize	
Open Na. Open URL Open DB. Ser	wrate Undo Edit Save
Pites	
Groom AddExpression E (s1+x2)? MZ	Apply
Current relation	Selected attribute
Relation: data-webs/filems unsupervised attitude AMExpression . Attitudes: 3 Instances: 7 Sum of weights: 7	Narine II Type Nutreens Ministry 3 (0%) Cedawit 6 Linegue 5 (7%)
Attributes	25dem: Value
Ab None Invest Patern	Manifranto 1 Manifranto 6 Mani 3152 Definio 1527
2 L Y 5 C 2	Chanse X (Marti)
, Names and State Stat	
lina.	
OK .	Log 🛷 x

'Discretize' to discretize the attributes into bins:

Discrete attributes are those that describe a category, called nominal attributes. Those attributes that describe a category that where there is a meaning in the order for the categories are called ordinal attributes. The process of converting a real-valued attribute into an ordinal attribute or bins is called discretization. You can discretize your real valued attributes in Weka using the Discretize filter.

1. Open the Weka Explorer.

2. Load the dataset.



Weka Explorer Loaded Diabetes Dataset

3. Click the "Choose" button for the Filter and select Discretize, it is under unsupervised.attribute.Discretize.



Weka Select Discretize Data Filter

4. Click on the filter to configure it. You can select the indices of the attributes to discretize, the default is to discretize all attributes, which is what we will do in this case. Click the "OK" button.

5. Click the "Apply" button to apply the filter.

You can click on each attribute and review the details in the "Selected attribute" window to confirm that the filter was applied successfully.

Missing	0 (0%) D	istinct: 10	Unique: 1 (0%)	62.0
No.	Label	Count	Weight	4
1	'(-inf-1.7]'	246	246.0	1
2	'(1.7-3.4]'	178	178.0	
3	'(3.4-5.1]'	125	125.0	
4	'(5.1-6.8]'	50	50.0	
5	'(6.8-8.5]'	83	83.0	1
5				
6				
178				
178	125			

Normalize Your Numeric Attributes:

Data normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0. Normalization is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (a bell curve).

You can normalize all of the attributes in your dataset with Weka by choosing the Normalize filter and applying it to your dataset.

You can use the following recipe to normalize your dataset:

1. Open the Weka Explorer.

2. Load your dataset.

Knowledge Engineering Lab Weika Explorer Preprocess Classify Cluster Associate Select attributes Visualize Open URL. Open DB.. Generate... Edit Open file. Undo Save Filter Choose None Apply Selected attribute Current relation Attributes: 9 Sum of weights: 768 Relation: pima_diabetes instances: 768 Name preg Missing 0 (0%) Type: Numeric Inique: 2 (0%) Distinct: 17 Un Statistic Value Attributes Minimum 0 17 Maximum At None Invert Pattern 3.845 Mean StdDev Name 2 plas 3 pres 4 skin Visualize Ali Class: class (Noni) 5 insu 6 mass 7 pedi 8 age 9 class , 0.1 Status 0K Log 0

Weka Explorer Loaded Diabetes Dataset

3. Click the "Choose" button to select a Filter and select *unsupervised.attribute.Normalize*.

lter					1000
	Discretize FirstOrder	*	followed with the		Apply
	FixedDictionaryStringToWordVe InterquartileRange	utes: 9 lights: 768	Name: preg Missing: 0 (0%) Distin	Tyr ct 17 Unig	pe: Numeric ue: 2 (0%)
	MakeIndicator MathExpression		Statistic Minimum	Value	
	MergelinfrequentNominalValues MergeManyValues MergeTwoValues NominalToBinary	Pattern	Maximum Mean StdDev	17 3.845 3.37	
	NominalToString		Class: class (Nom)		Visualize
	NumericCleaner NumericToSinary NumericToNominal NumericTransform Obfuscate PartitionedMultiFilter PKIDiscretize		103 73 50 45	ананананананананананананананананананан	<u> </u>

Weka Select Normalize Data Filter

- 4. Click the "Apply" button to normalize your dataset.
- 5. Click the "Save" button and type a filename to save the normalized copy of your dataset.

Reviewing the details of each attribute in the "Selected attribute" window will give you confidence that the filter was successful and that each attribute was rescaled to the range of 0 to 1.

Name: preg Missing: 0 (0%)	Distinct:	17	Type: Numeric Unique: 2 (0%)
Statistic		Value	
Minimum		0	
Maximum		1	
Mean		0.226	
StdDev		0.198	

Weka Normalized Data Distribution

Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data, such as k-nearest neighbors and artificial neural networks

Standardize Your Numeric Attributes

Data standardization is the process of rescaling one or more attributes so that they have a mean value of 0 and a standard deviation of 1. Standardization assumes that your data has a Gaussian (bell curve) distribution. This does not strictly have to be true, but the technique is more effective if your attribute distribution is Gaussian.

You can standardize all of the attributes in your dataset with Weka by choosing the Standardize filter and applying it your dataset.

You can use the following recipe to standardize your dataset:

- 1. Open the Weka Explorer
- 2. Load your dataset.
- 3. Click the "Choose" button to select a Filter and select unsupervised.attribute.Standardize.



Weka Select Standardize Data Filter

4. Click the "Apply" button to normalize your dataset.

5. Click the "Save" button and type a filename to save the standardized copy of your dataset.

Reviewing the details of each attribute in the "Selected attribute" window will give you confidence that the filter was successful and that each attribute has a mean of 0 and a standard deviation of 1.

Name: preg		Type: Numeric
Missing: 0 (0%)	Distinct: 17	Unique: 2 (0%)
Statistic	Valu	ie
Minimum	-1.	141
Maximum	3.9	04
Mean	-0	
StdDev	1	

Weka Standardized Data Distribution

Standardization is useful when your data has varying scales and the algorithm you are using does make assumptions about your data having a Gaussian distribution, such as linear regression, logistic regression and linear discriminant analysis.