Lab 1
Data Preprocessing

Objective

Apply different kind of preprocessing techniques on given dataset.

Input Files-

Data file	=>	iris_data.csv
Metadata file	=>	iris_metadata.txt

Reading Data

1. Read the given metadata file and find out about the number of fields and their types (i.e. continuous, nominal, etc.), also have a look at input data file. Notice, field names are not mentioned.
2. Drag-n-drop a *Variable File Node* (from Sources tab in Nodes palette) on to *Modeler Canvas*.
3. Double-click on *Variable File Node* and go to "*File"* tab. Browse and select the given input data file.
4. Uncheck "*Read field names from file"*, and check "*Specify number of fields"* and set the field count value according to information given in metadata file.
5. Go to "*Data"* tab and make sure that you are satisfied with the "*Storage*" type of each field. Otherwise, check *"Override"* and change the storage type.
6. Go to *"Filter"* tab and change the output field names according to the names mentioned in metadata file.
7. Go to *"Types"* and make sure *"Measurement"* of each field is correct (i.e. continuous, nominal, etc.)
8. [Optional] To know about range of values that each field have in this data, click on *"Values"* cell in front of the field, select *"<Read>"* and click on *"Read Values"* button.
9. To mark a field as Classifier field (on which classification has to be performed), mark it as *"Target"* in *"Role"* cell.
10. Click *"Apply"* and *"OK"*. Now, a source node with your input file name must appear on the modeler canvas.
11. [Optional] To view the output of this source node, add a *"Table node"*
12. Connect it to source node and *"Run"* the stream.

Running a stream

To run whole stream, use the “Run Stream” button in toolbar

To run down-stream from a particular node, right click on it and select “Run From Here”

Connecting Node N_1 to N_2

Right-click on **N_1,** select *"Connect"* and then select **N_2, or**
Select **N_1 ,**press *"F2"* and then select **N_2.**

Analyzing Input

1. To understand the input data, use *"Data Audit Node"* from *"Output"* tab in node palette and connect it to the source node.
2. Double-click on data audit node and go to *"Settings"* tab.
3. Either select *"Default"* for auditing all fields or select *"Use custom fields"* and select required fields.
4. Check all options in *"Display"*.
5. Go to *"Quality"* tab and select *"Outlier and Extreme Detection method"*. Specify the values for outliers and extremes.
6. Click *"Apply"* and *"OK"*. Now, a data audit node will appear on the modeler canvas.*"Run"* the stream. (Ctrl+E)
7. Observe output (*"Audit"* and *"Quality"* tabs) carefully and try to determine necessity of preprocessing.

   From *"Audit"* tab, determine answers of following questions.

   a. Is cardinality of nominal attributes correct?
      E.g.- cardinality of Boolean attributes should be
      2 See under column labeled as *"Unique"*
   b. Are Min, Max, Std. deviation, variance, etc. are under permissible limits?
   c. Does input data needs Sampling (Is it too large)?
   d. How values are distributed for each attribute?
      See the graphs

   From *"Quality"* tab, determine answers of following questions.

   a. Are there any *"Outliers"* and/or *"Extreme* "values?
   b. Are there any *"Missing/Null"* values?

Handling Missing Values

Remove records having a NULL value

1. Select a *"Select Node"* from *"Record Ops"* tab in nodes palette.
2. Connect source node to the select node.
3. Double click on select node and go to *"Settings"* tab.
4. In order to discard records with missing values select *"Discard"* and write the expression which selects such records. Expressions can be easily written using Expression Builder Tool.

a. Click on *"Expression Builder"* button.Left hand side lists all functions with their return values and right hand side lists fields in the input dataset. Connectors specified in middle are used to connect two expressions.

b. To select all records with missing Field1 value, select *"@NULL"* function from function list and insert it into expression. Select Field1 from field list and insert it as argument of above mentioned function. *"@NULL"* is used for numeric values only, for string values use *"STRING1 matches STRING2"* and specify STRING2 as "".

c. To select all records with any one of the missing field values, write expressions as explained above, joined by *"OR"* operation.

d. Verify the expression by clicking *"Check"*. Remove syntax error, if any. Click *"OK"*.

5. Click *"Apply"* and *"OK"*. Now, a select node will appear on the modeler canvas.
6. [Optional] To view the output of this select node, add a *"Table node"*. Connect select node to the table node and *"Run"* the stream from the select node.

Replace Numeric attributes by mean value

1. To obtain the mean of each field, connect a *"Set Globals"* from "*Output*" to the stream. Select required fields and operations (Mean, Sum, Min, Max and SDev). Click *Run*.
2. Select a "Filler Node" from "Field Ops" tab in nodes palette.
3. Connect source node to this filler node.
4. Double click on filler node and go to "Settings" tab.
5. Select all numeric fields in "Fill in fields".
6. Select "Replace" condition as "Blank and null values".
7. In "Replace with"box, open the Expression Builder window. Select Globals in right-hand side panel and select Global Mean of appropriate fields. In case of multiple fields selection use @FIELD keyword. Write following expression in replace with box -
8. @GLOBAL_MEAN(@FIELD)
9. Click "Apply" and "OK". Now, a filler node will be appear on the modeler canvas.

Remove Nominal attributes having null value

1. In order to remove nodes with missing nominal values, select a *"Select Node"* from *"Record Ops"* tab in nodes palette.
2. Connect source node to this select node. Go to *"Settings"* tab.
3. Select *"Discard"* and write the expression which selects nominal attributes with null values using *"STRING1 matches STRING2"* and specify STRING2 as "".
4. Click *"Apply"* and *"OK"*. Now, a select node will appear on the modeler canvas.
5. [Optional] To view the output of this select node, add a*"Table node"*. Connect this node to the table node and *"Run"* the stream from the select node.

Discretization (Binning)

1. Select a *"Binning Node"* from *"Field Ops"* tab in nodes palette and connect select node of previous step to this node. Go to *"Settings"* tab.
2. Select all the numeric fields with *"Continuous"* values as *"Bin Fields".*
3. As per the requirement, you may choose among many Binning methods like *"Fixed-width"* or *"Tiles"* or *"Ranks"*or *"Mean/Std Deviation based".* For now, select *"Fixed-width"* method. Read *Help* to know what these term mean.
4. Select *"Number of bins"* as appropriate.
5. [Optional] Go to *"Bin Values"* tab, click on *"Read Values"* button. You can see the lower and upper range of each of the bins for each of the binned field.
6. Click *"Apply"* and *"OK"*. Now, a binning node will appear on the modeler canvas.
7. New Binned values are appended as new fields in the dataset. To remove old fields, select a *"Filter Node"* from *"Field Ops"* tab in nodes palette.
8. Connect binning node to this filter node and double click on it.
9. Click on *"Filter"* column of old fields to filter them out. Click *"Apply"* and *"OK"*.
10. [Optional] To view the output of this node, add a *"Table node".* Connect this node to the table node and *"Run"* the stream from this node.

Sampling

1. Select a *"Sample Node"* from *"Record Ops"* tab in nodes palette. Connect select node of Step 1 to this sample node. Double click on sample node and go to *"Settings"* tab.
2. As per the requirement, you may choose among the *"Simple"* or *"Complex"* sample method. For now, select *"Simple"* method.
3. You can select the sampling criterion as either *"First n"* or *"1-in-n"* or *"Random %"*. Try to understand difference between all the operations.
4. Click *"Apply"* and *"OK"*. Now, a sample node will appear on the modeler canvas.
5. [Optional] To view the output of this sample node, add a *"Table node"*. Connect this node to the table node and *"Run"* the stream.

Normalization

1. Select a "Filler Node" from "Field Ops" tab in nodes palette. Connect select node to this filler node accordingly. Double click on Filler node and go to "Settings" tab. Select all numeric fields in "Fill in fields". Select "Replace" condition as "Always".

2. In *"Replace with"* box, write your normalization expression. For example, for 0-1 normalization you should write

   (@FIELD - @GLOBAL_MIN(@FIELD))/(@GLOBAL_MAX(@FIELD)-
   @GLOBAL_MIN(@FIELD))

3. Click *"Apply"* and *"OK"*. Now, a filler node will appear on the modeler canvas.

4. [Optional] To view the output of this filler node, add a *"Table node"*. Connect this node to the table node and *"Run"* the stream.

Correlation Determination

1. Select a *"Statistics Node"* from *"Output"* tab in nodes palette.
2. Connect select node of Step 1 to this node.
3. Double click on the node and go to *"Settings"* tab.
4. Under *"Examine"* box, select all numeric fields. You may check all statistics parameter under *"Statistics"* box.
5. Under *"Correlate"* box, again select all numeric fields.
6. You may adjust correlation strength parameters for weak, medium and strong correlation. Click *"OK"*.
7. Click *"Apply"* and *"OK"*. Now, a statistics node will appear on the modeler canvas.
8. Run the stream from this node and observe the correlations between different fields. It will help you in selecting features to be consider for further processing.