



IBM SPSS Modeller

Name - Aradhya Dubey, Amogh Srivastava, Aditya Rai, Abhishek Giri

Class - BCADS - 31

Teacher - Mr. Vikas



Practical 1

Start a new stream.

1. Open SPSS Modeler and create a new Stream

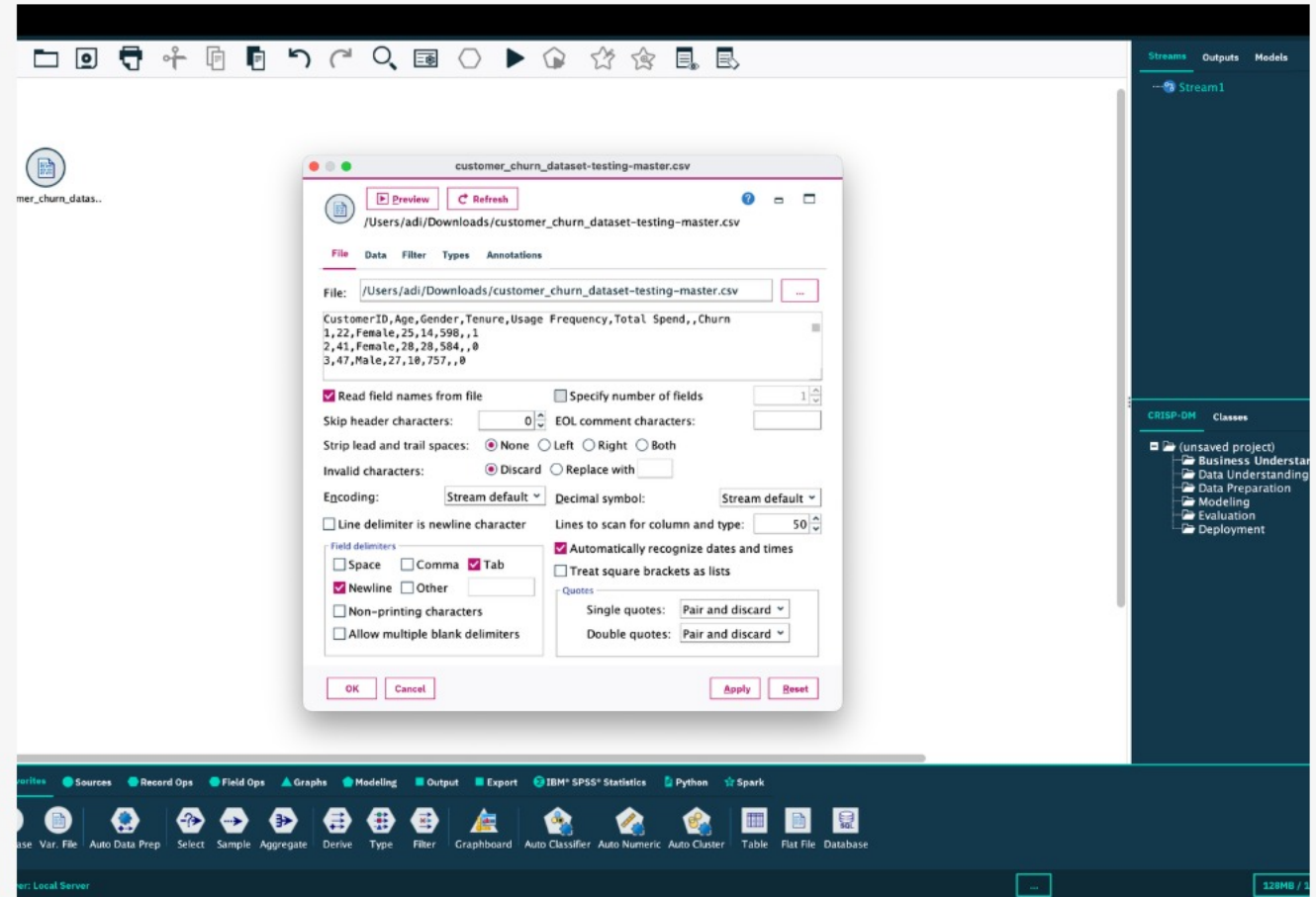
Read the CSV

1. Add a **Var.File** node (or **Source** → **File** → **CSV** depending on your version).

2. Point it to your CSV file
customer_churn_dataset-testing-master.csv.

3. Check “First row contains field names”.

4. Click **Preview** to inspect rows.



Practical 1

Add a Type node (Data type & role correction.
Drag a **Type** node and connect the Var.File → Type.
In the Type node:
Set CustomerID → Role = **Id** (or **Key**),
Measurement = **Nominal** (but role = Id excludes
from modeling).
Set Age → Role = **Input**, Measurement = **Ratio** or
Continuous.
Gender → Role = **Input**, Measurement = **Nominal**.
Tenure → Role = **Input**, Measurement = **Ratio**.
Usage Frequency → Role = **Input**, Measurement = **Ratio**.
Total Spend → Role = **Input**, Measurement = **Ratio**.
Unnamed: 6 → Role = **None** (or remove column).
Churn → Role = **Target**, Measurement = **Nominal**
(convert 0/1 to Nominal if it's numeric).

The screenshot displays the IBM SPSS Modeler interface. A 'Type' node is connected to a 'Var. File' node. The 'Type' node's configuration window is open, showing a table of field settings. The table has columns for Field, Measurement, Values, Missing, Check, and Role. The fields listed are CustomerID, Age, Gender, Tenure, Usage Frequency, Total Spend, Unnamed: 6, and Churn. The roles assigned are: CustomerID (Id), Age (Input), Gender (Input), Tenure (Input), Usage Frequency (Input), Total Spend (Input), Unnamed: 6 (None), and Churn (Target). The measurement for Churn is set to Nominal. The 'View current fields' radio button is selected.

Field	Measurement	Values	Missing	Check	Role
CustomerID	Nominal	<Read>		None	Id
Age	Continuous	<Read>		None	Input
Gender	Nominal	<Read>		None	Input
Tenure	Continuous	<Read>		None	Input
Usage Fre...	Continuous	<Read>		None	Input
Total Spend	Continuous	<Read>		None	Input
Unnamed: 6	Categorical	<Read>		None	None
Churn	Nominal	<Read>		None	Target

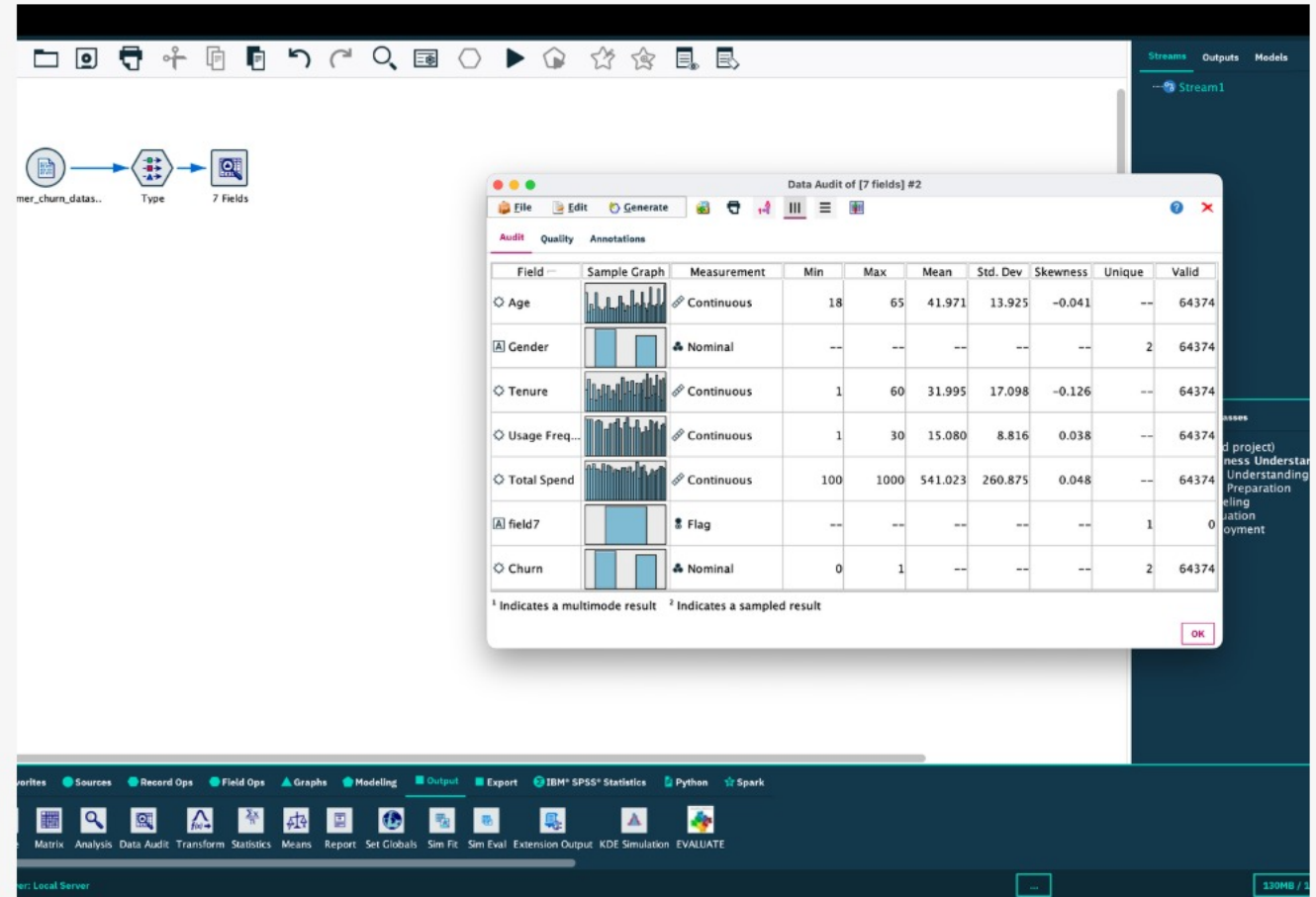
Practical 1

Add a Data Audit node

Connect **Type** → **Audit** (or use **Data Audit**).

Run it to check missing values, distributions, outliers.

Action: You'll see Unnamed: 6 empty – confirm removal (if not already set in Type).



Practical 1

Feature engineering (Derive node)

Add a **Derive** node to create useful predictors.
Connect after Type / Filler.

Examples to create:

$\text{SpendPerUse} = \text{Total Spend} / (\text{Usage Frequency} + 0.001)$ — helps capture intensity.

The screenshot displays the Alteryx software interface. On the left, a workflow diagram shows a data source 'mer_churn_data...' connected to a 'Type' node, which then connects to a 'Derive' node. The 'Derive' node is configured to calculate 'SpendPerUse'. Below the 'Type' node, a '7 Fields' node is visible. The 'Derive' node's output is connected to a 'Table' node. On the right, a data preview window titled 'Table (9 fields, 64,374 records)' is open, showing a table with columns: 'der', 'Tenure', 'Usage Frequency', 'Total Spend', 'field7', 'Chu...', and 'SpendPerUse'. The table contains 20 rows of data. The bottom of the interface shows the Alteryx toolbar with various tools like 'Auto Data Prep', 'Select', 'Sample', 'Aggregate', 'Derive', 'Type', 'Filter', 'Graphboard', 'Auto Classifier', 'Auto Numeric', 'Auto Cluster', 'Table', 'Flat File', and 'Database'.

	der	Tenure	Usage Frequency	Total Spend	field7	Chu...	SpendPerUse
1	le	25	14	598	1	0	42.714
2	le	28	28	584	0	0	20.857
3		27	18	757	0	0	75.699
4		9	12	232	0	0	19.333
5	le	58	24	533	0	0	22.208
6		41	14	500	0	0	35.714
7	le	37	15	574	1	0	38.266
8	le	36	11	323	0	0	29.363
9		20	5	687	0	0	137.397
10		8	4	995	0	0	248.744
11	le	42	27	526	1	0	19.481
12		13	23	187	0	0	8.138
13		2	7	758	0	0	188.284
14		46	27	438	0	0	16.222
15		21	17	663	0	0	39.608
16	le	1	3	677	1	0	225.659
17		54	3	636	0	0	211.993
18		9	38	127	0	0	4.233
19	le	48	2	396	0	0	197.998
20		39	19	282	0	0	18.632

Practical 2

Partition the data

Add a **Partition** node (important for honest evaluation).

Set proportions: **Training = 70%, Test = 30%** (or 60/40). Use random seed for reproducibility (e.g., seed = 12345).

Connect output to modeling nodes (the partition node has ports for training/test — most model nodes will use training, and Evaluation node will use both).

The screenshot displays the Orange3 data mining software interface. In the main workspace, a workflow is visible: a data source node 'mer_churn_data...' is connected to a 'Type' node, which then connects to a 'SpendPerUse' node, and finally to a 'Partition' node. Below the 'Type' node is a '7 Fields' node, and below the 'SpendPerUse' node is a 'Table' node. A 'Partition' node settings dialog is open in the foreground. The dialog has two tabs: 'Settings' and 'Annotations'. The 'Settings' tab is active, showing the following configuration:

- Partition field:** Partition
- Partitions:** ☒ Train and test ☐ Train, test and validation
- Training partition size:** 70 (Label: Training, Value = "1_Training")
- Testing partition size:** 30 (Label: Testing, Value = "2_Testing")
- Validation partition size:** 0 (Label: Validation, Value = "3_Validation")
- Total size:** 100%
- Values:** ☐ Use system-defined values ("1", "2" and "3") ☒ Append labels to system-defined values ☐ Use labels as values
- ☒ Repeatable partition assignment
- Seed:** 1234567 (Generate button)
- ☐ Use unique field to assign partitions: (dropdown menu)

The dialog has 'OK', 'Cancel', 'Apply', and 'Reset' buttons at the bottom. The background interface shows a toolbar with various data processing nodes and a sidebar with project and class lists.

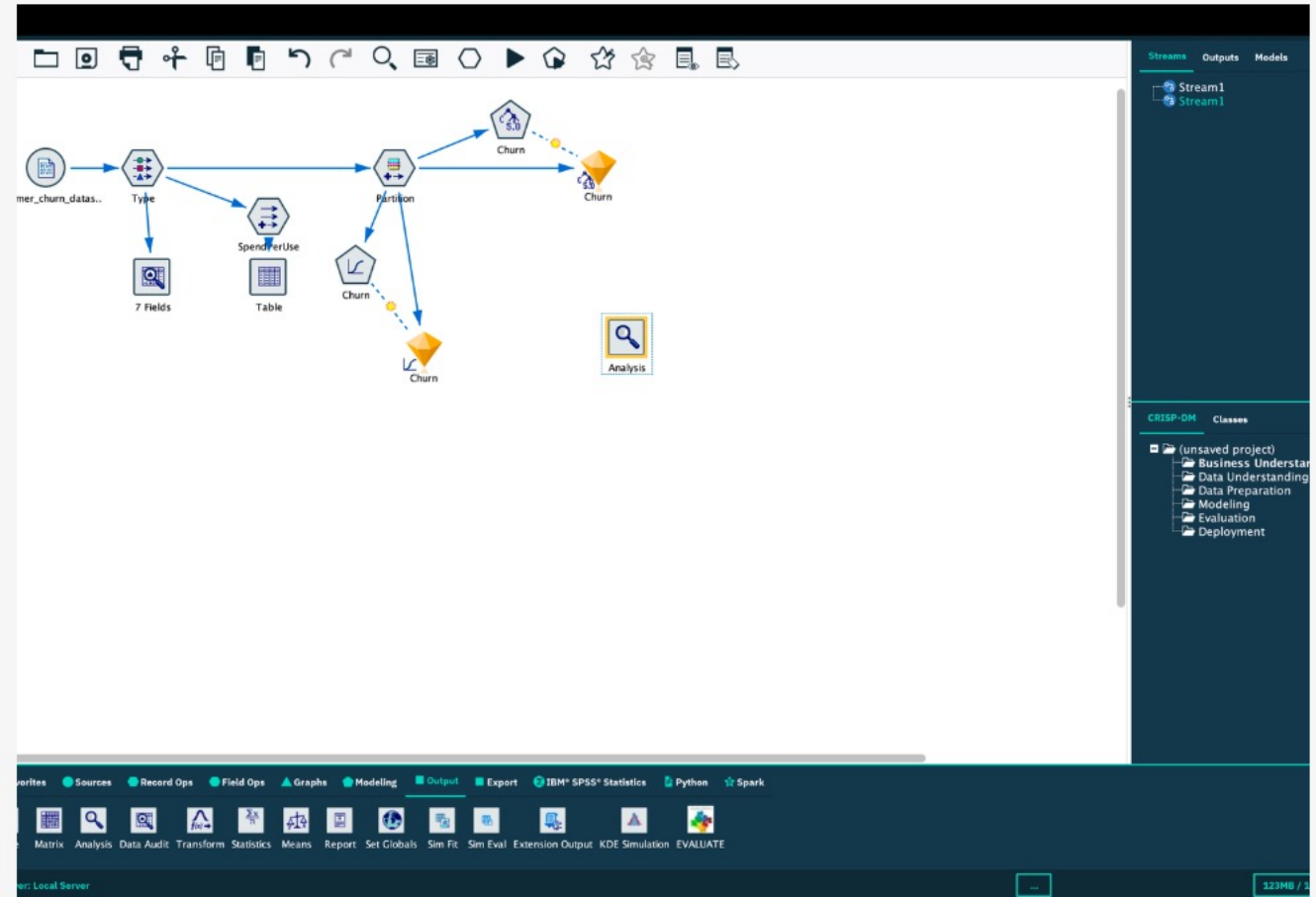
Practical 2

Choose & add models

Add at least two different model nodes so you can compare:

Auto Classifier or **C5.0** (decision tree) — good interpretability.

Logistic Regression (for baseline & coefficients).



Practical 2

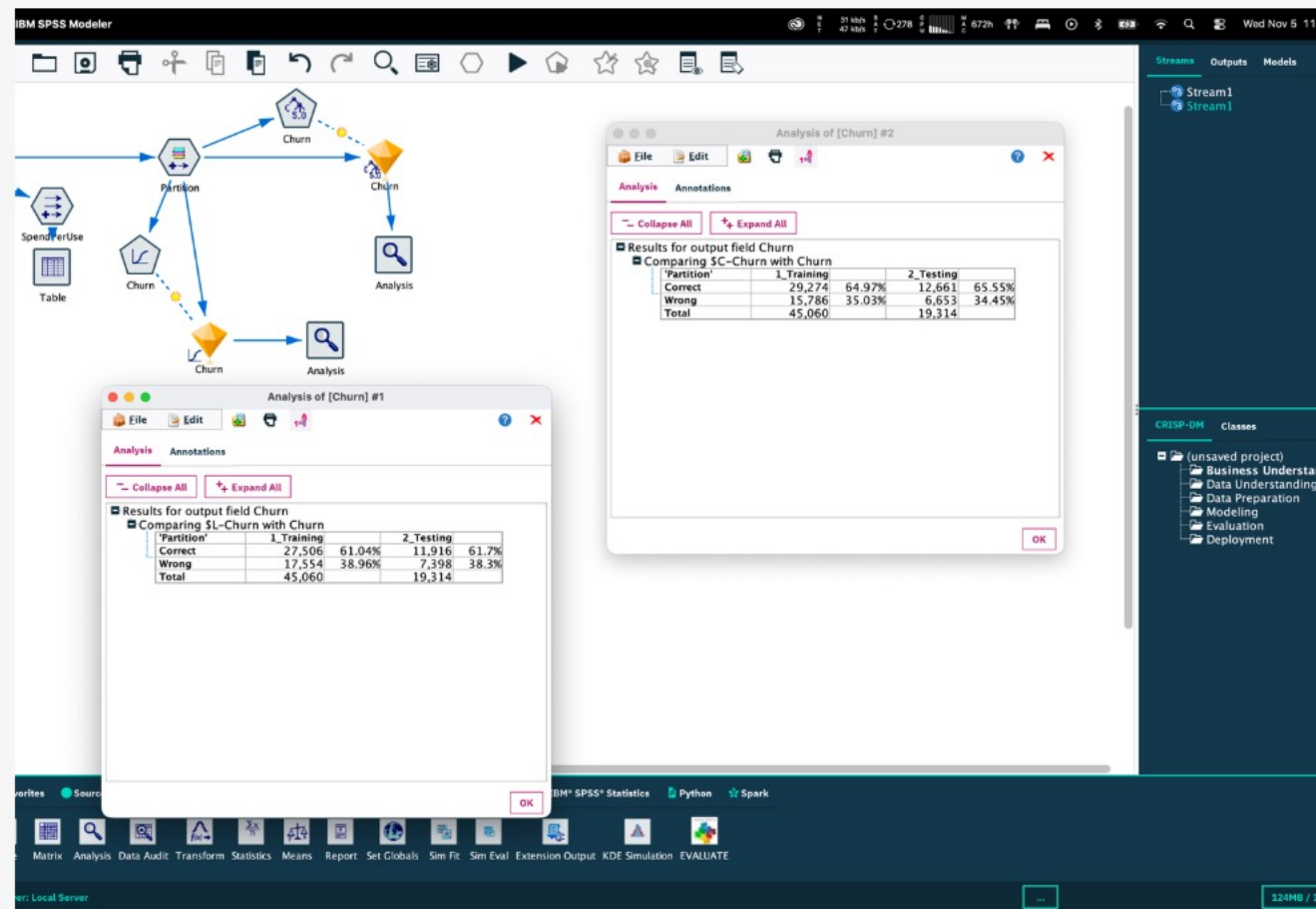
Evaluation node

Add an **Analysis** → **Evaluation** → **Analysis / Evaluation** node (called **Analysis**, **Analysis node**, or **Evaluation** depending on version).

Connect all model outputs into the Evaluation node (drag connecting lines from model nodes to Evaluation).

Run the Evaluation node; it will compare models on Test partition and give you:

Confusion matrix, accuracy, precision, recall, ROC curve, AUC, lift chart.





Thank You!!!

