

# ***BABU BANARASI DAS UNIVERSITY***



*Submitted To :  
Mr. Vikas Kumar*

*Submitted By :  
Aman Yadav*

**Agenda/Definition:** The project aims to predict customer churn for a Gym using the CHAID decision tree method. By analyzing customer data, the model identifies key factors influencing churn, helping the bank target retention efforts effectively

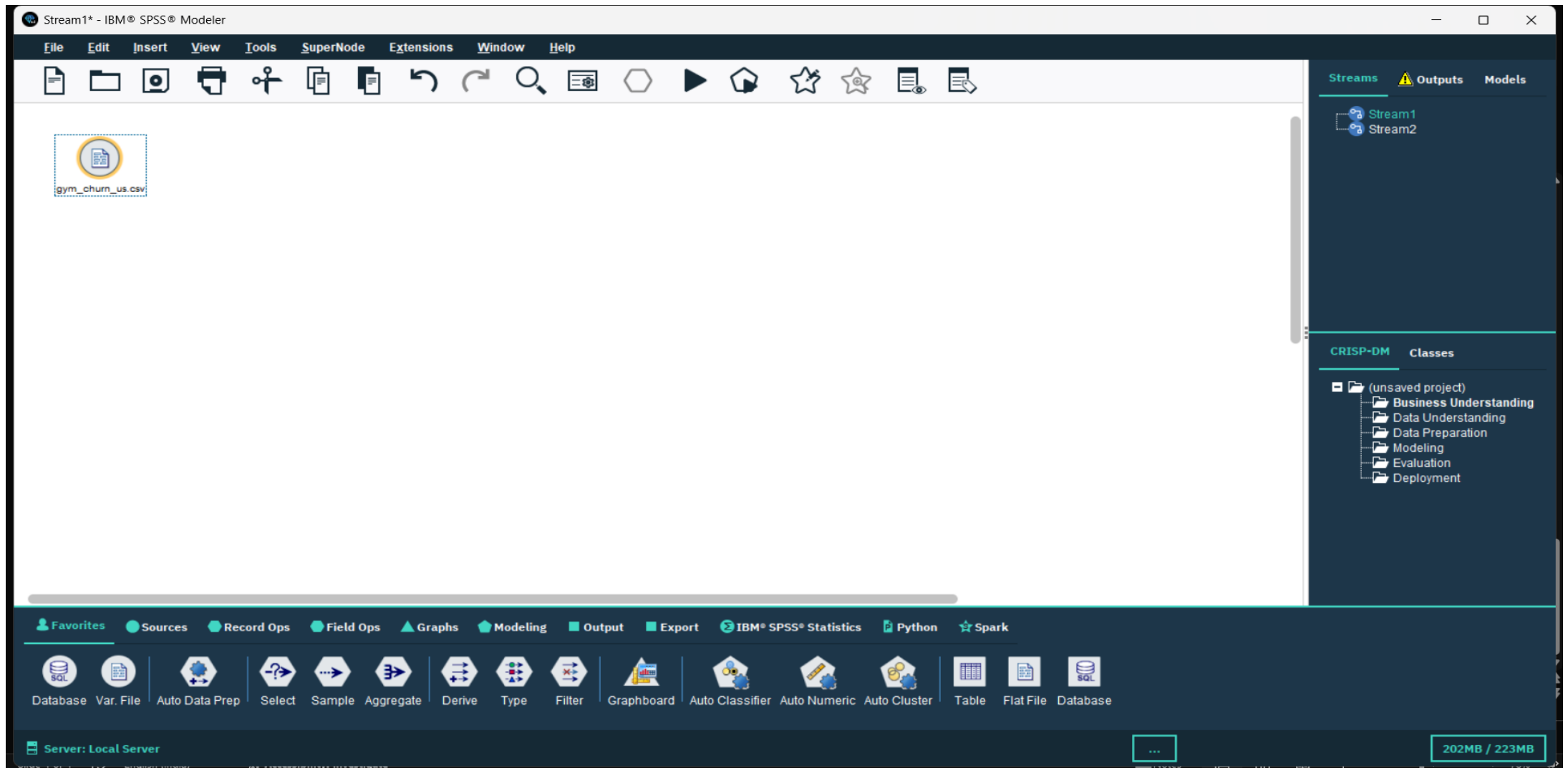
**Outcomes/Learning:** You will learn how to build a classification model to predict customer churn using CHAID in IBM SPSS Modeler. The project demonstrates the process of data preparation, model configuration, execution, and interpretation of results.

**Required Tool:** The tool used for this project is IBM SPSS Modeler.

**Working :** The project involves importing customer data, setting variable roles, configuring the CHAID model node, running the decision tree analysis, and interpreting the decision tree output. This workflow aids in understanding customer segments likely to churn.

## Step 1: Import Data

Loaded the dataset (churn\_prediction.csv) into SPSS Modeler and confirmed all fields were correctly recognized.



## Step 2: Inspect and Prepare Data:

Checked for missing or invalid values and corrected any formatting or data type issues

The screenshot displays the IBM SPSS Modeler Stream2 interface. A central window titled "Table (14 fields, 4,000 records) #4" is open, showing a preview of the data. The table has 14 columns: gender, Near\_Location, Partner, Promo\_friends, Phone, Contract\_period, Group\_visits, Age, and Avg\_additional\_charges. The first 20 rows are visible, showing various values for these attributes. The interface includes a top menu bar (File, Edit, Insert, View, Tools, SuperNode, Extensions, Window, Help) and a toolbar with icons for file operations, data manipulation, and visualization. On the right side, there is a sidebar with tabs for Streams, Outputs, and Models. The Streams tab is active, showing a project named "Churn" with a CRISP-DM process flow diagram. The bottom of the interface features a toolbar with icons for various data processing tasks like Select, Sample, Aggregate, Derive, Type, Filter, Graphboard, Auto Classifier, Auto Numeric, Auto Cluster, Table, Flat File, and Database. The status bar at the bottom indicates "Server: Local Server" and "201MB / 223MB".

	gender	Near_Location	Partner	Promo_friends	Phone	Contract_period	Group_visits	Age	Avg_additional_charges
1	1	1	1	1	0	6	1	29	
2	0	1	0	0	1	12	1	31	
3	0	1	1	0	1	1	0	28	
4	0	1	1	1	1	12	1	33	
5	1	1	1	1	1	1	0	26	
6	1	1	0	0	1	1	1	34	
7	1	1	1	1	0	6	1	32	
8	0	1	0	0	1	1	0	30	
9	1	1	1	1	1	1	1	23	
10	0	1	0	0	1	1	0	31	
11	0	1	0	0	0	6	1	32	
12	1	1	1	1	0	1	0	27	
13	0	1	1	1	1	1	1	33	
14	1	1	0	0	1	1	1	27	
15	0	1	0	0	1	6	0	35	
16	0	1	1	1	1	12	0	29	
17	0	1	1	1	1	1	1	31	
18	0	1	0	1	1	6	1	29	
19	0	1	0	0	1	1	1	30	
20	1	1	1	1	1	12	1	29	

### Step 3: Assign Variable Types/Roles :

Used the **Type node** to assign roles and measurement levels. The churn field was defined as the **target variable**.

The screenshot displays the IBM SPSS Modeler interface. In the main workspace, a flow diagram shows a 'gym\_churn\_us.csv' file connected to a 'Table' node, which is then connected to a 'Type' node. The 'Type' node is currently selected, and its configuration window is open. The window title is 'Table (5 fields, 2 records) #5'. It contains a 'Table' tab and an 'Annotations' tab. The 'Table' tab shows a table with 5 fields and 2 records. The 'Annotations' tab is currently empty. The right sidebar shows the 'Streams' panel with a 'Churn' stream. Below it, the 'CRISP-DM' panel shows a project structure with folders for Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The bottom toolbar includes various nodes like Database, Var. File, Auto Data Prep, Select, Sample, Aggregate, Derive, Type, Filter, Graphboard, Auto Classifier, Auto Numeric, Auto Cluster, Table, Flat File, and Database. The status bar at the bottom indicates 'Server: Local Server' and '189MB / 242MB'.

	Lifetime_Mean	Avg_class_frequency_total_Mean	Churn_Mean	gender_to_m/f	Record_Count
1	3.775	1.893	0.266	M	2041
2	3.673	1.865	0.265	F	1959

## Step 4: Derive Node:

**Derive Node** converted the numeric gender codes (0 and 1) into categorical labels “F” and “M” for better readability.

The screenshot displays the IBM SPSS Modeler Stream2\* interface. A workflow is visible on the left, starting with a 'gym\_churn\_us.csv' file, followed by a 'Table' node, a 'Type' node, a 'gender\_to\_m/f' node, and finally an 'active\_gym\_users' node. A central window titled 'Table (5 fields, 2 records) #6' shows the following data:

	Lifetime_Mean	Avg_class_frequency_total_Mean	Churn_Mean	gender_to_m/f	Record_Count
1	3.775	1.893	0.266	M	2041
2	3.673	1.865	0.265	F	1959

The right sidebar shows the 'Streams' tab with a 'Churn' node. Below it, the 'CRISP-DM' section lists the project structure: (unsaved project) > Business Understanding > Data Understanding > Data Preparation > Modeling > Evaluation > Deployment. The bottom toolbar includes various nodes like Database, Var. File, Auto Data Prep, Select, Sample, Aggregate, Derive, Type, Filter, Graphboard, Auto Classifier, Auto Numeric, Auto Cluster, Table, Flat File, and Database. The status bar at the bottom indicates 'Server: Local Server' and '199MB / 242MB'.

## Step 5: Partitoin node:

A **Partition Node** in IBM SPSS Modeler is used to split the dataset into separate subsets, such as **training** and **testing** samples. It helps in **model validation** by allowing you to test the model's accuracy on unseen data.

The screenshot displays the IBM SPSS Modeler interface. In the background, a workflow diagram is visible, starting with a 'Table' node connected to 'gym\_churn\_us.csv', followed by a 'Type' node, a 'Partition' node, and then a 'gender\_to\_m/f' node, leading to 'active\_gym\_users'. The 'Partition' node is highlighted, and its configuration dialog is open in the foreground.

The 'Partition' dialog box 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:** 60 (Label: Training, Value: "1\_Training")
- Testing partition size:** 40 (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: [Field Selection]

The dialog box has 'OK', 'Cancel', 'Apply', and 'Reset' buttons at the bottom.

The right sidebar shows the 'Streams' tab with a 'Churn' stream. Below it, the 'CRISP-DM' classes are listed: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

The bottom toolbar includes various nodes and tools: Database, Var. File, Auto Data Prep, Select, Sample, Aggregate, Derive, Type, Filter, Graphboard, Auto Classifier, Auto Numeric, Auto Cluster, Table, Flat File, and Database. The status bar at the bottom indicates 'Server: Local Server' and '204MB / 242MB'.

# Step 6: Aggregate Node:

The **Aggregate Node** in IBM SPSS Modeler is used to **summarize data by grouping records** based on key fields. It helps compute statistics like **mean, sum, count, or maximum** for each group to identify overall trends and patterns.

Stream2\* - IBM® SPSS® Modeler

File Edit Insert View Tools SuperNode Extensions Windows

gym\_churn\_us.csv

Table

Type

Partition

gender\_to\_m/f

Aggregate

Table

active\_gym\_users

Database

Var. File

Auto Data Prep

Select

Sample

Aggregate

Derive

Type

Filter

Graphboard

Auto Classifier

Auto Numeric

Auto Cluster

Table

Flat File

Database

Server: Local Server

205MB / 242MB

Aggregate

Preview

Settings Optimization Annotations

Key fields:  
gender\_to\_m/f

Basic Aggregates

Aggregate fields:

Field	Sum	Mean	Min	Max	SDev	Median	Count	Variance	1st Quartile	3rd Quartile
Churn		<input checked="" type="checkbox"/>								
Lifetime		<input checked="" type="checkbox"/>								
Avg_class_freq...		<input checked="" type="checkbox"/>								

Default mode: ☒ Sum ☒ Mean ☐ Min ☐ Max ☐ SDev ☐ Median ☐ Count ☐ Variance ☐ 1st Quartile ☐ 3rd Quartile

New field name extension:

☒ Include record count in field

Add as: ☒ Suffix ☐ Prefix

Aggregate Expressions

Field	Expression

OK Cancel Apply Reset



## Step 7: Train the Model (Run CHAID)

Executed the model stream and generated the CHAID decision tree output.

The screenshot displays the IBM SPSS Modeler interface. The main workspace shows a model stream with the following components: 'gym churn us.csv' (Table) -> 'Type' (Table) -> 'Partition' (Table) -> 'Churn' (Table). A secondary path shows 'active gym users' (Table) -> 'gender to mlt' (Table) -> 'Aggregate' (Table) -> 'Table' (Table). The 'Churn' node is highlighted, indicating it is the active model.

The 'Churn' dialog box is open, showing the 'Model' tab. The 'Model' list contains the following rules:

- Lifetime <= 0 [Mode: 1]
- Lifetime > 0 and Lifetime <= 1 [Mode: 1]
- Lifetime > 1 and Lifetime <= 2 [Mode: 0]
- Lifetime > 2 and Lifetime <= 4 [Mode: 0]
- Lifetime > 4 [Mode: 0]

The 'Predictor Importance' chart shows the relative importance of predictors for the target 'Churn'. The chart is a horizontal bar chart with 'Lifetime' as the most important predictor, followed by 'Age', 'Avg...', 'Contr...', 'Month...', 'Partner', 'gender', and 'Group...'. The x-axis is labeled 'Least Important' to 'Most Important'.

The bottom of the interface shows the 'Server: Local Server' status and a memory usage indicator of 193MB / 300MB.

## Step 8: Filter Node :

A **Filter Node** in IBM SPSS Modeler is used to **include or exclude specific fields** from the dataset. It helps in **removing irrelevant or unwanted variables** before analysis or modeling.

The screenshot displays the IBM SPSS Modeler interface. The main workspace shows a workflow starting with a 'gym\_churn\_us.csv' file, followed by a 'Table' node, a 'Type' node, a 'Partition' node, and a 'Churn' node. A 'Filter' node is also present, connected to the 'Churn' node. The 'Filter' node's configuration dialog is open, showing the 'Filter' tab. The dialog lists 16 fields, with 13 filtered (indicated by red X's) and 3 out (indicated by arrows). The fields are:

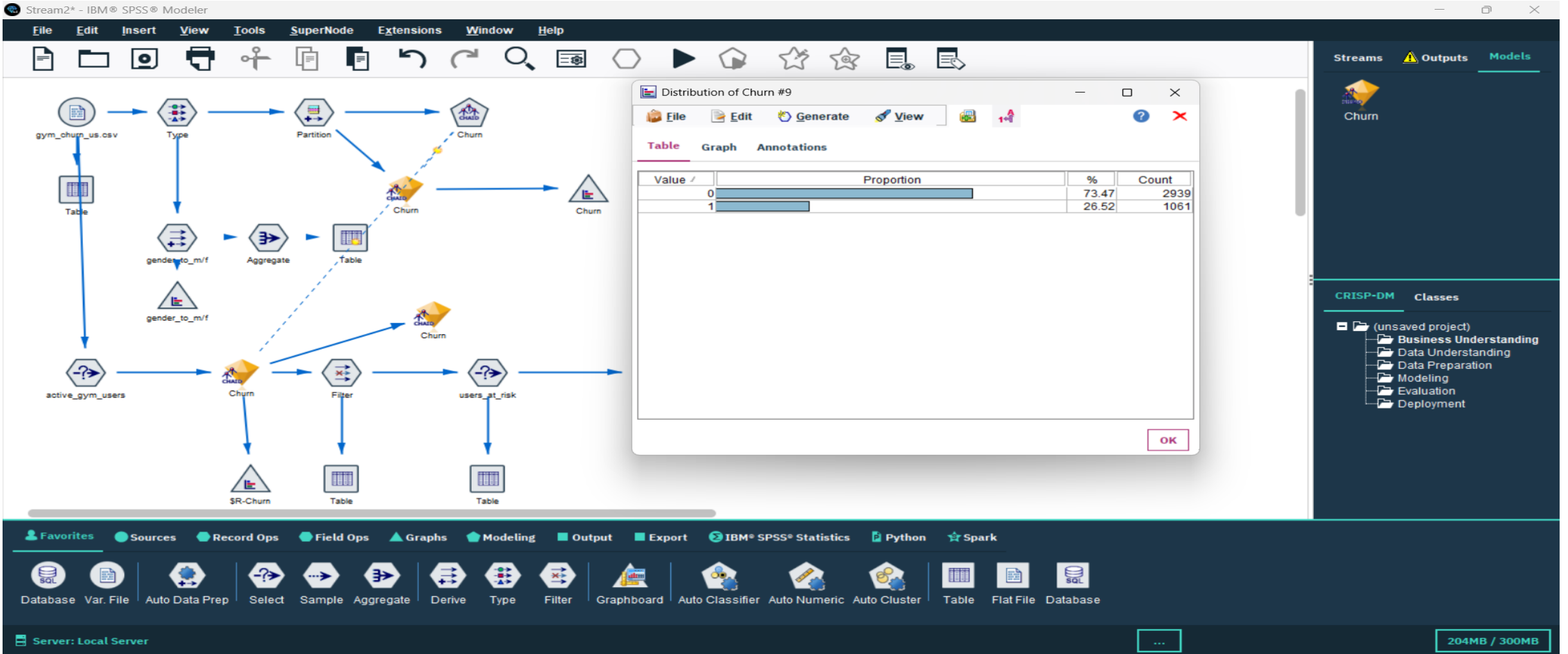
Field	Filter	Field
Group_visits	—X—>	Group_visits
Age	—X—>	Age
Avg_additional_charges_total	—X—>	Avg_additional_charges_total
Month_to_end_contract	—X—>	Month_to_end_contract
Lifetime	—X—>	Lifetime
Avg_class_frequency_total	—X—>	Avg_class_frequency_total
Avg_class_frequency_current...	—X—>	Avg_class_frequency_current...
Churn	—>	Churn
\$R-Churn	—>	\$R-Churn
\$RC-Churn	—>	\$RC-Churn

The dialog also includes a 'Preview' button, a 'Filter' tab, and a 'View current fields' radio button selected. The bottom status bar shows 'Server: Local Server' and '195MB / 300MB'.

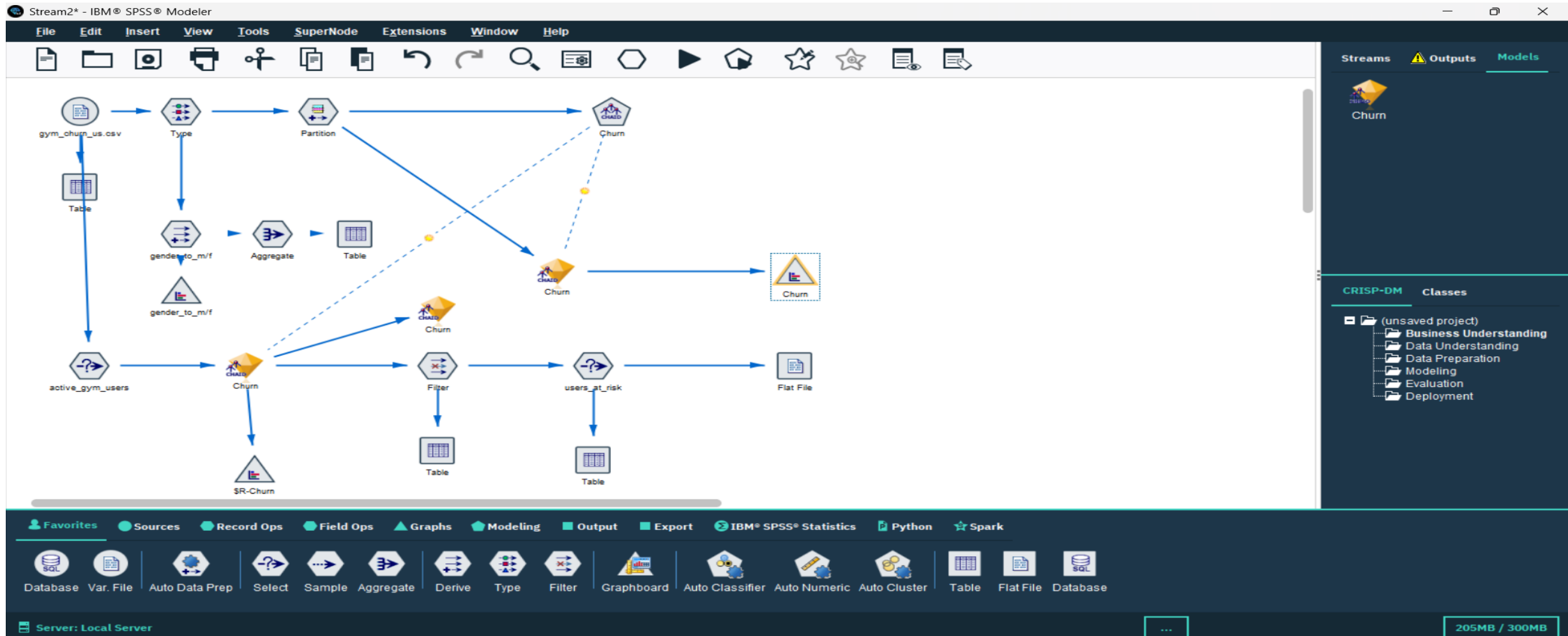
### Step 8: Calculate Churn Rate:

Used Aggregate and Table nodes to compute churn proportions.

- 0  $\rightarrow$  81.47% (Non-churned)
- 1  $\rightarrow$  18.53% (Churned)



Compared actual vs. predicted churn rates to evaluate model performance and interpret findings for actionable retention planning. The complete SPSS Modeler stream (shown below) illustrates the workflow from data import to churn prediction and analysis:





## Conclusion

The churn analysis conducted using **IBM SPSS Modeler** provided valuable insights into customer behavior and retention at the gym. Through systematic data preparation and transformation, key variables such as **gender**, **lifetime**, and **average class frequency** were analyzed to understand their relationship with churn. The **Derive Node** was effectively used to convert numeric gender codes into readable labels (“M” and “F”), improving the interpretability of the results.

Further, by using the **Aggregate Node**, important statistical summaries like mean lifetime, average class frequency, and churn rate were computed for each gender group. The analysis revealed that both male and female customers have similar churn rates, but slight variations in engagement and lifetime values. These findings highlight the importance of personalized engagement strategies to reduce member dropout and improve retention.

Overall, the project demonstrates how **IBM SPSS Modeler** can be leveraged to perform data preparation, transformation, and statistical analysis in a structured way. It also emphasizes the role of data-driven decision-making in understanding customer patterns and supporting effective business strategies.

# *Summary*

In summary, this project successfully applied the CHAID decision tree to uncover actionable insights for customer retention. It highlights how data-driven approaches can help banks anticipate churn, improve engagement, and make informed strategic decisions. The knowledge gained from this workflow strengthens analytical proficiency in SPSS Modeler and lays a foundation for future enhancements using advanced machine learning models or automated churn monitoring systems