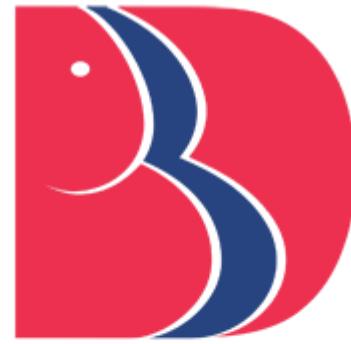


BABU BANARASI DAS UNIVERSITY



**BBD
UNIVERSITY**

Session- 2025-26

Submitted To :

Mr. Vikas Kumar

Submitted By :

Amitesh Singh

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

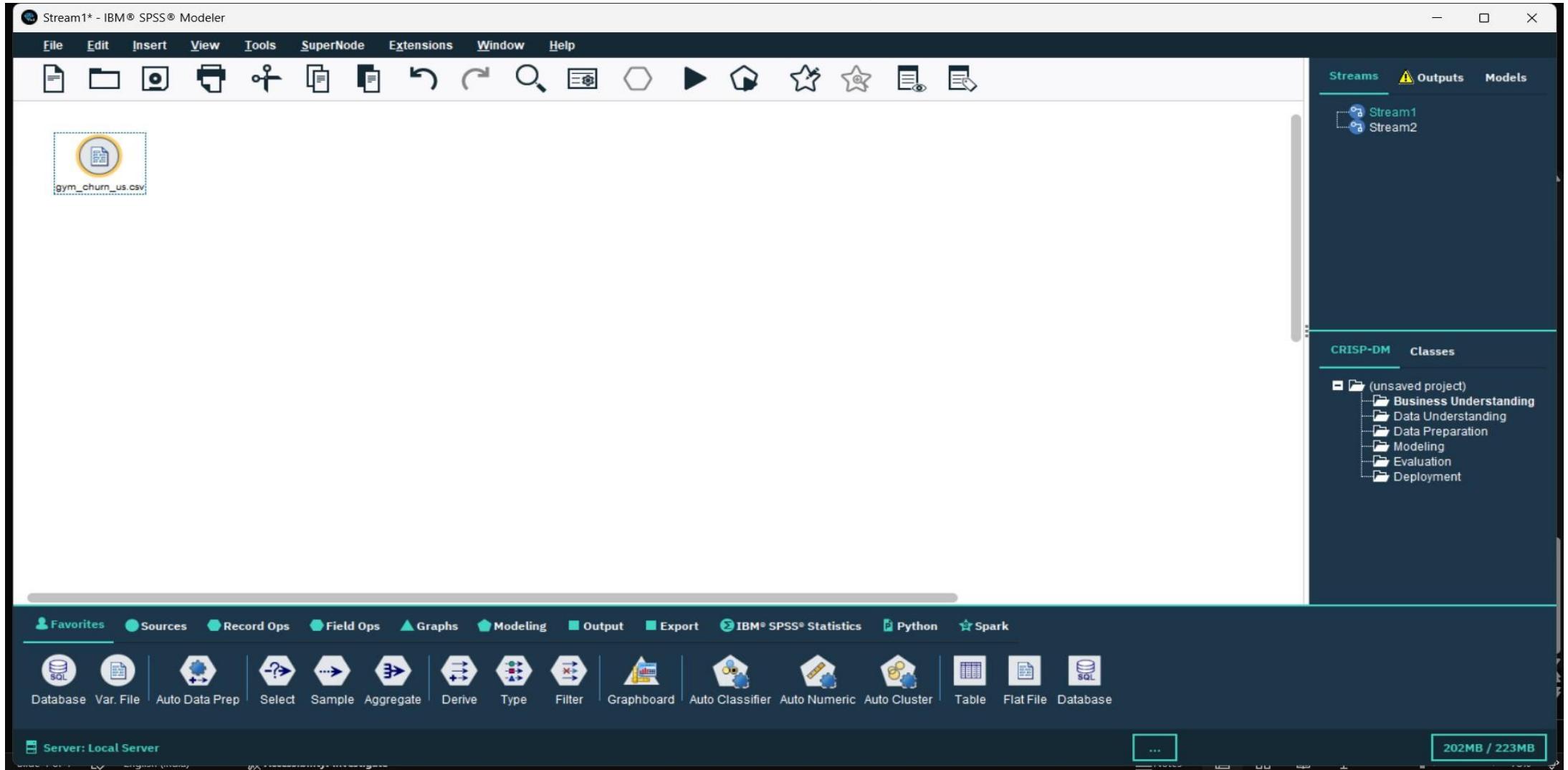
Outcomes/Learning: You will learn how to build a classification model to predict customer churn using CHAID in IBM SPSS Modeler. The project demonstrates the process of data preparation, model configuration, execution, and interpretation of results.

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

Stream2* - IBM® SPSS® Modeler

File Edit Insert View Tools SuperNode Extensions Window Help

Streams Outputs Models

gym_churn_us.csv

Table (14 fields, 4,000 records) #4

Table Annotations

	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	0	1	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	

OK

Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics Python Spark

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 ... 201MB / 223MB

The screenshot shows the IBM SPSS Modeler interface. A central window displays a table titled 'Table (14 fields, 4,000 records) #4' with data from 'gym_churn_us.csv'. The table has 14 columns and 4,000 rows. The columns are: gender, Near_Location, Partner, Promo_friends, Phone, Contract_period, Group_visits, Age, and Avg_additional_charges_. The data includes binary values (0 or 1) for most columns, with some numerical values like age and average additional charges. Below the table is an 'OK' button. The interface also features a toolbar at the top with various icons for file operations, and a bottom navigation bar with tabs for Favorites, Sources, Record Ops, etc. On the right side, there's a sidebar for 'CRISP-DM Classes' which is currently expanded to show the project structure.

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 shows the IBM SPSS Modeler interface with a stream titled "Stream2* - IBM® SPSS® Modeler". The stream consists of a "gym_churn_us.csv" source node connected to a "Type" node. A blue line highlights the connection between the source and the type node. A second blue line highlights the connection from the type node to a "Table" output node. A preview window titled "Table (5 fields, 2 records) #5" is open, displaying 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 interface includes a toolbar at the top, a menu bar with File, Edit, Insert, View, Tools, SuperNode, Extensions, Window, and Help. The bottom navigation bar includes links for Favorites, Sources, Record Ops, Field Ops, Graphs, Modeling, Output, Export, IBM® SPSS® Statistics, Python, and Spark. A bottom status bar shows "Server: Local Server" and memory usage "189MB / 242MB". On the right side, there are tabs for Streams, Outputs, and Models, and a CRISP-DM Classes section listing Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

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 shows the IBM SPSS Modeler interface with a stream titled "Stream2* - IBM® SPSS® Modeler".

Stream Overview:

- Inputs:** A "Table" node connected to "gym_churn_us.csv".
- Process:** The "Table" node connects to a "Type" node, which then connects to a "gender_to_m/f" node.
- Outputs:** The "gender_to_m/f" node connects to "active_gym_users".

Table Node Content:

A "Table (5 fields, 2 records) #6" window is open, showing 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

Right Panel:

- Streams:** Shows a "Churn" stream icon.
- Models:** Shows a "CRISP-DM Classes" section with an "unsaved project" folder containing "Business Understanding", "Data Understanding", "Data Preparation", "Modeling", "Evaluation", and "Deployment".

Bottom Navigation:

- Favorites, Sources, Record Ops, Field Ops, Graphs, Modeling, Output, Export, IBM® SPSS® Statistics, Python, Spark
- 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
- 199MB / 242MB

Step 5: Partition 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.

Stream2* - IBM® SPSS® Modeler

File Edit Insert View Tools SuperNode Extensions Window Help

Partition

Generate Preview

Settings Annotations

Partition field: Partition

Partitions: Train and test (radio button selected) 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") (radio button selected) Append labels to system-defined values Use labels as values

Repeatable partition assignment (checkbox checked)

Seed: 1234567 Generate

Use unique field to assign partitions: [dropdown menu]

OK Cancel Apply Reset

Streams Outputs Models

Churn

CRISP-DM Classes

(unsaved project) Business Understanding Data Understanding Data Preparation Modeling Evaluation Deployment

Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics Python Spark

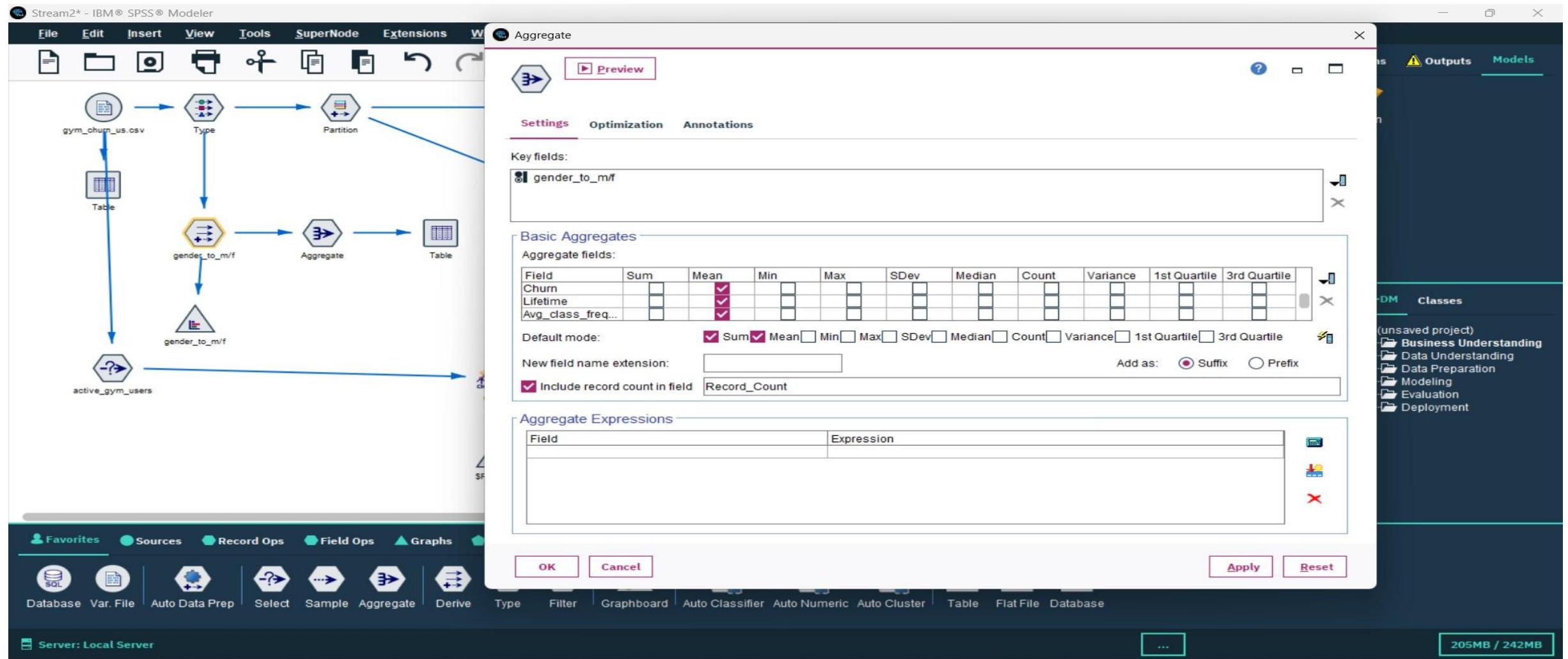
Database Var. File Auto Data Prep Select Sample Aggregate Derive Type Filter Graphboard Auto Classifier Auto Numeric Auto Cluster Table Flat File Database

Server: Local Server ... 204MB / 242MB

```
graph LR; gym_churn_us[CSV File] --> Type[Type]; Type --> Aggregate[Aggregate]; Aggregate --> Partition[Partition]; Partition --> Training[Training]; Partition --> Testing[Testing]; Partition --> Validation[Validation]; Validation --> active_gym_users[active_gym_users]
```

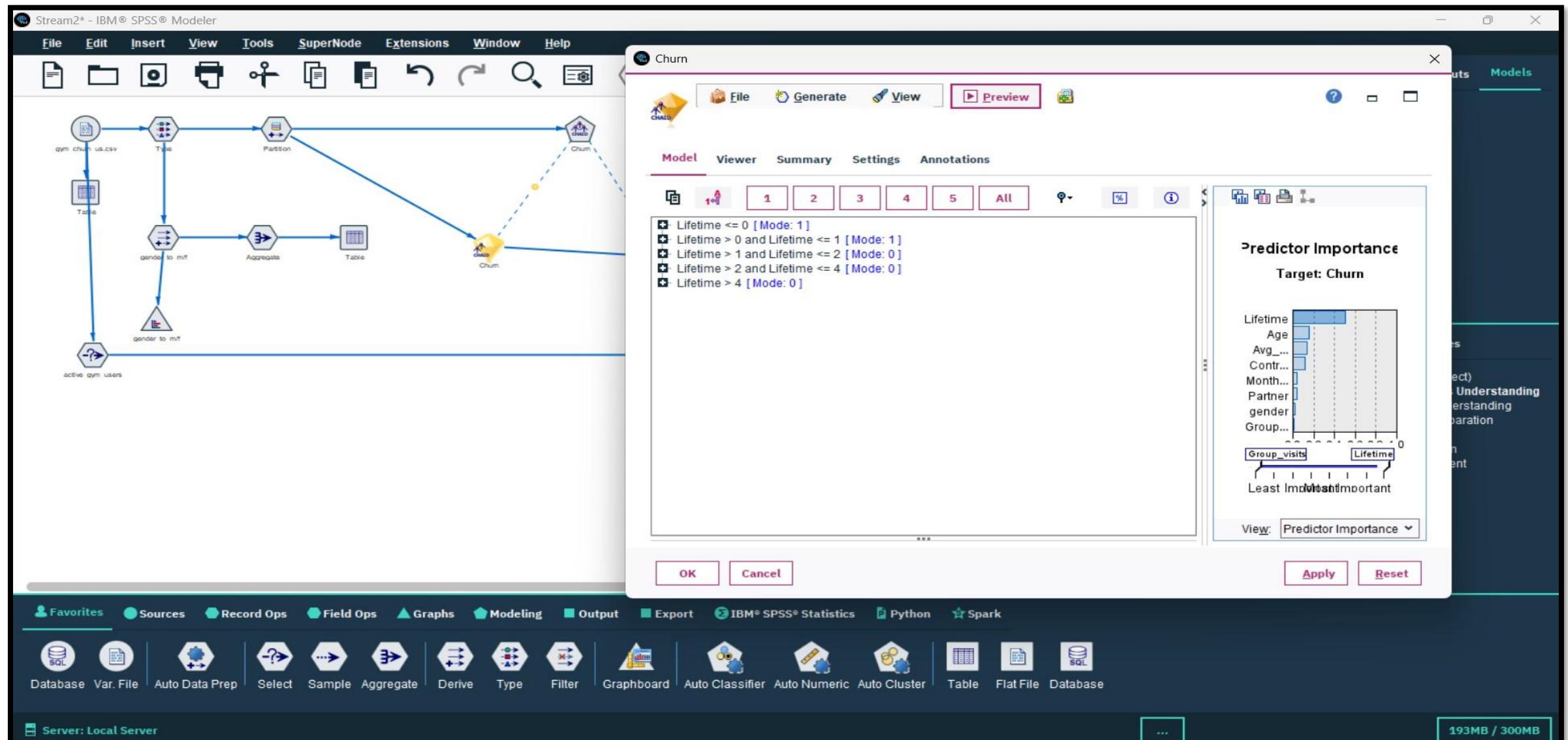
Step 6: Aggregate Node:

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.



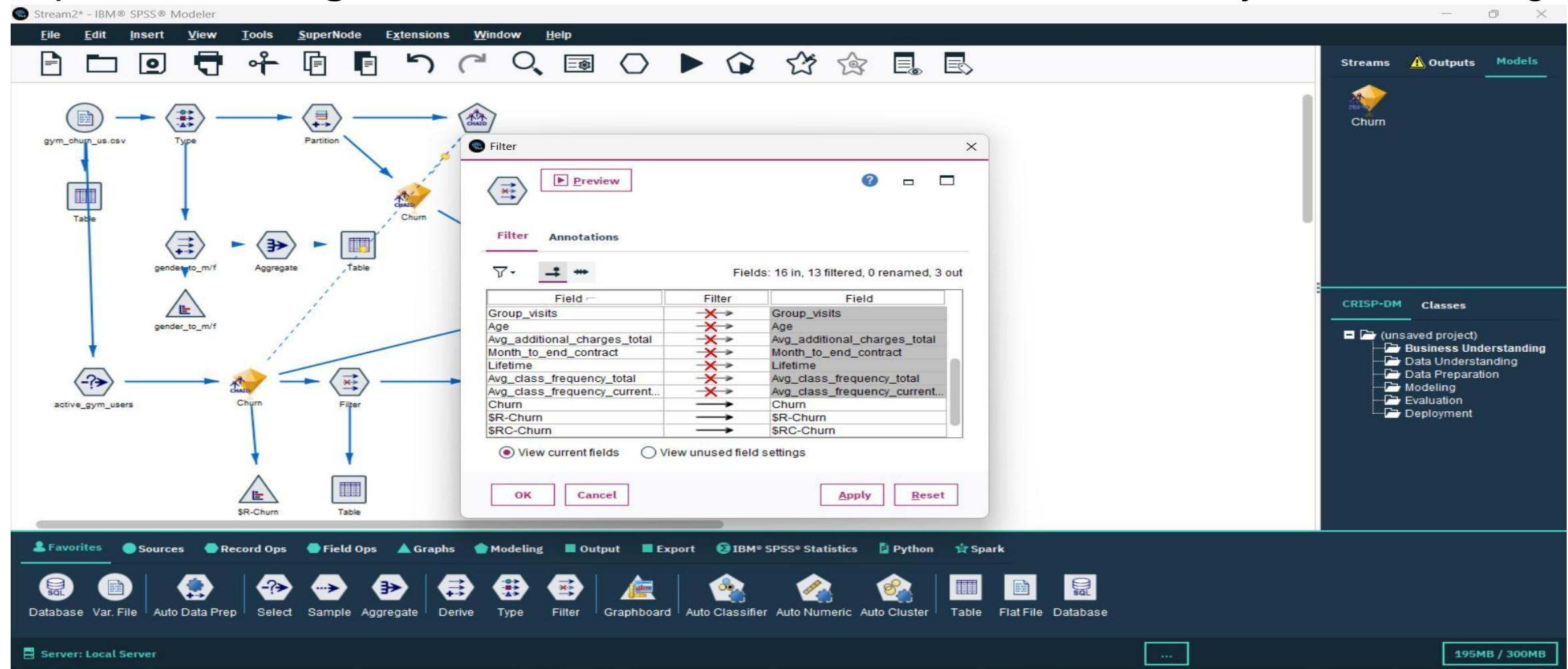
Step 7: Train the Model (Run CHAID)

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



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.



Step 8: Calculate Churn Rate:

Used Aggregate and Table nodes to compute churn proportions.

- 0 → 81.47% (Non-churned)
- 1 → 18.53% (Churned)

Stream2* - IBM® SPSS® Modeler

Streams **Outputs** **Models**

CRISP-DM Classes

- (unsaved project)
- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

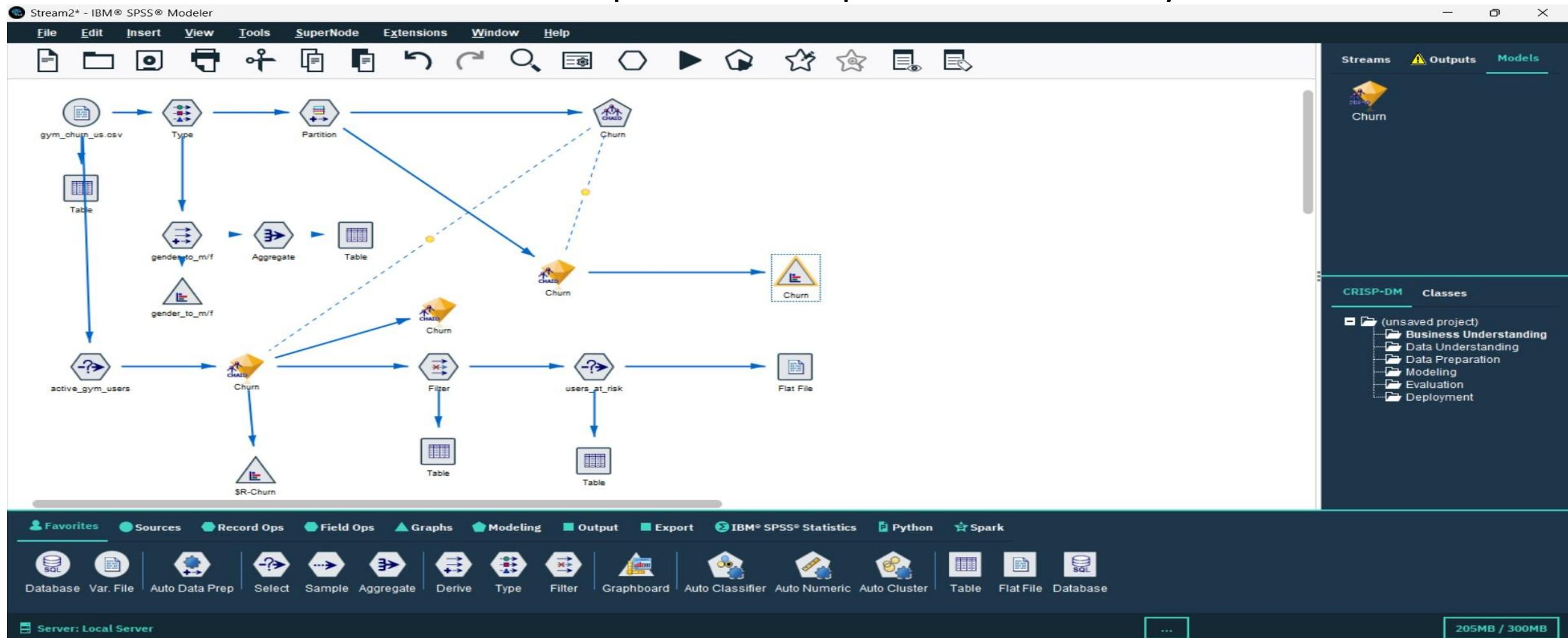
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics Python Spark

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 ... 204MB / 300MB

Step 13: Model Evaluation & Summary

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:



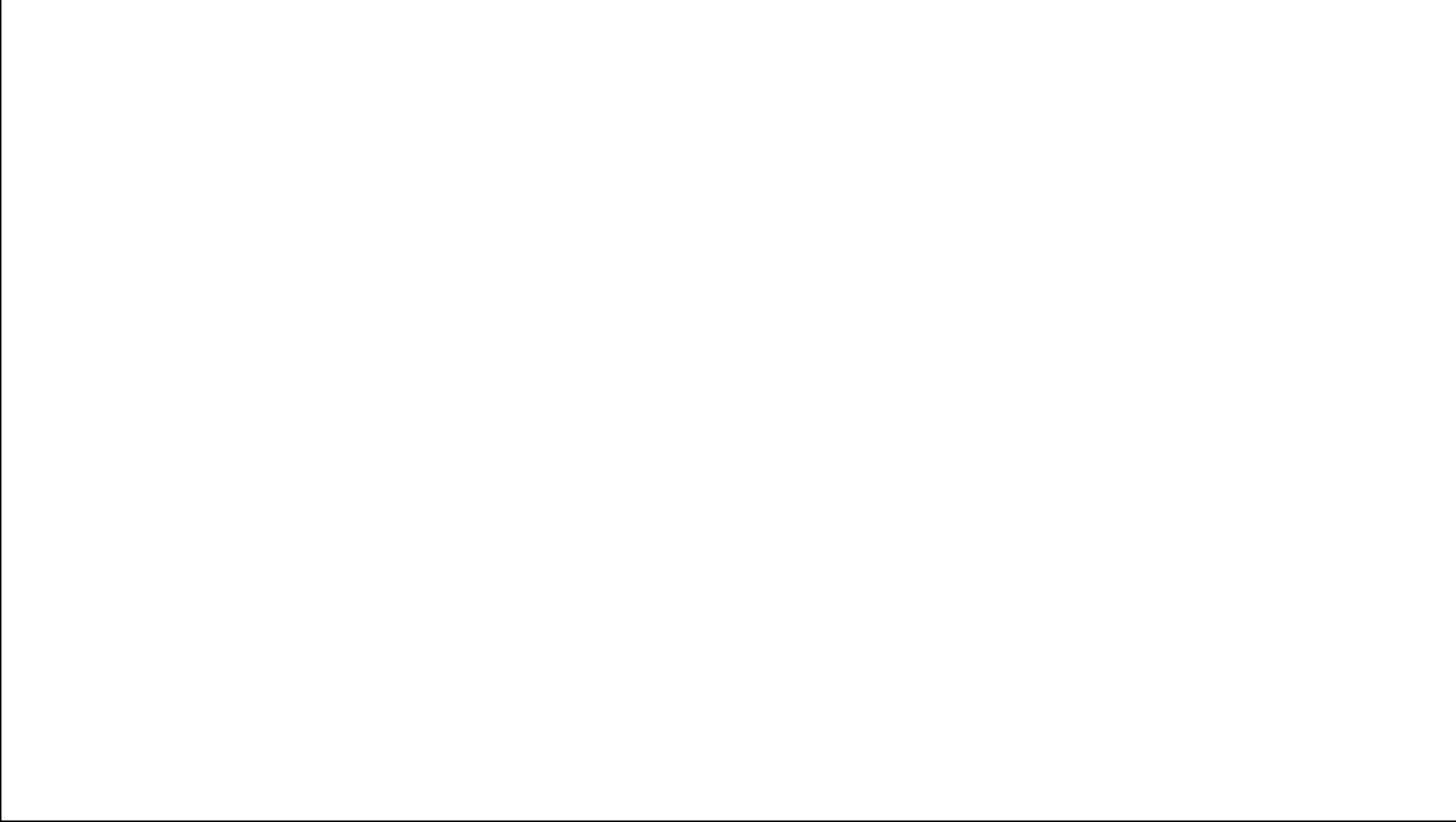


Conclusion

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.