Babu Banarasi Das University

Bank Customer Churn Prediction Using CHAID Decision Tree Model in IBM SPSS Modeler

SUBMITTED TO:

Mr. Ayushmman

SUBMITTED BY:

Varsha Tiwari

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

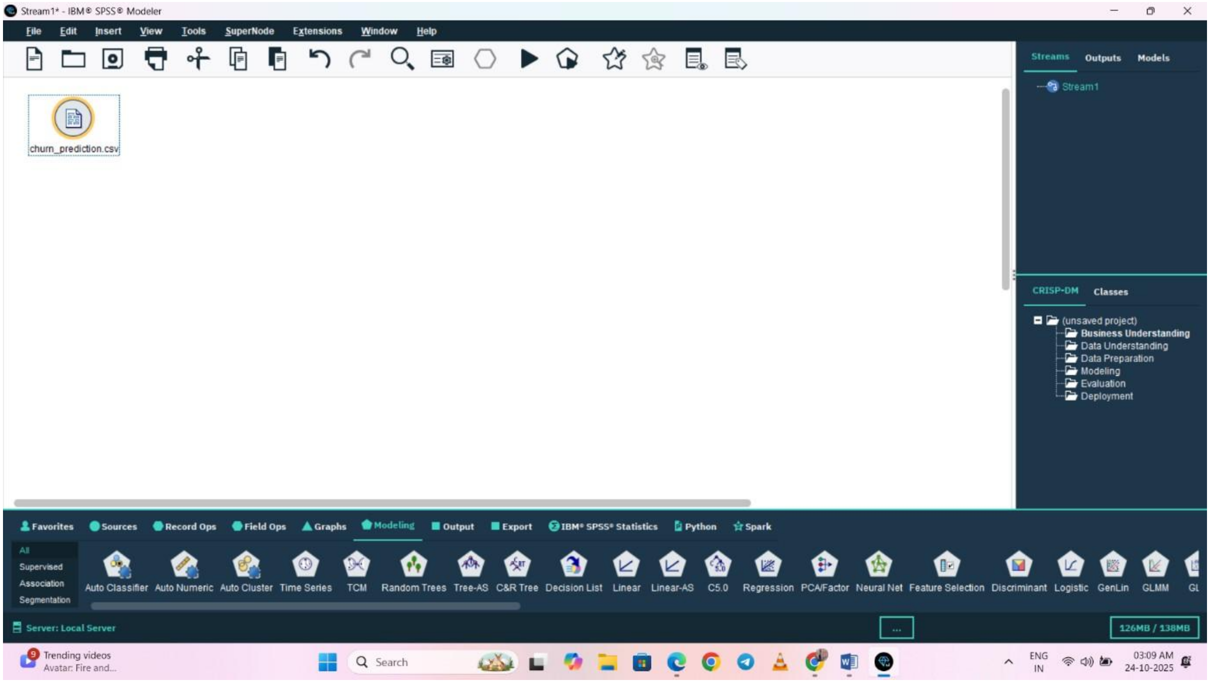
Outcomes/Learning: You will learn how to build a classification model to predict customer churn using CHAID in IBM SPSS Modeler. The project demonstrates the process of data preparation, model configuration, execution, and interpretation of results.

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.

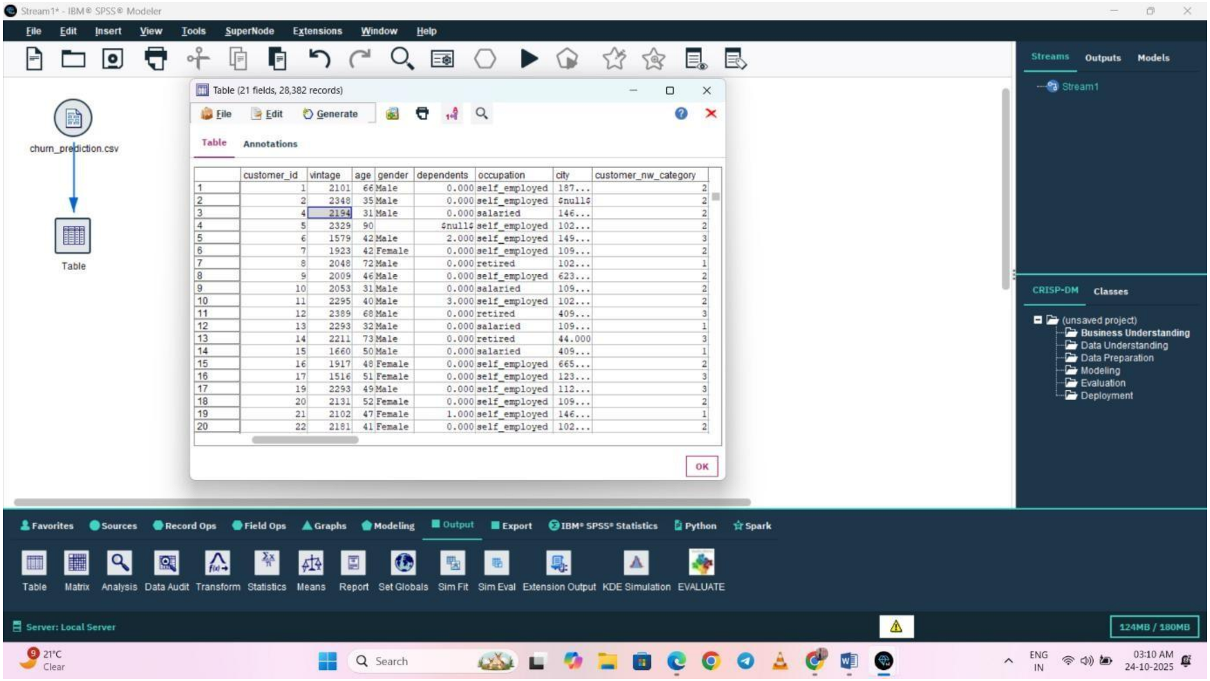
Step 1: Import Data

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



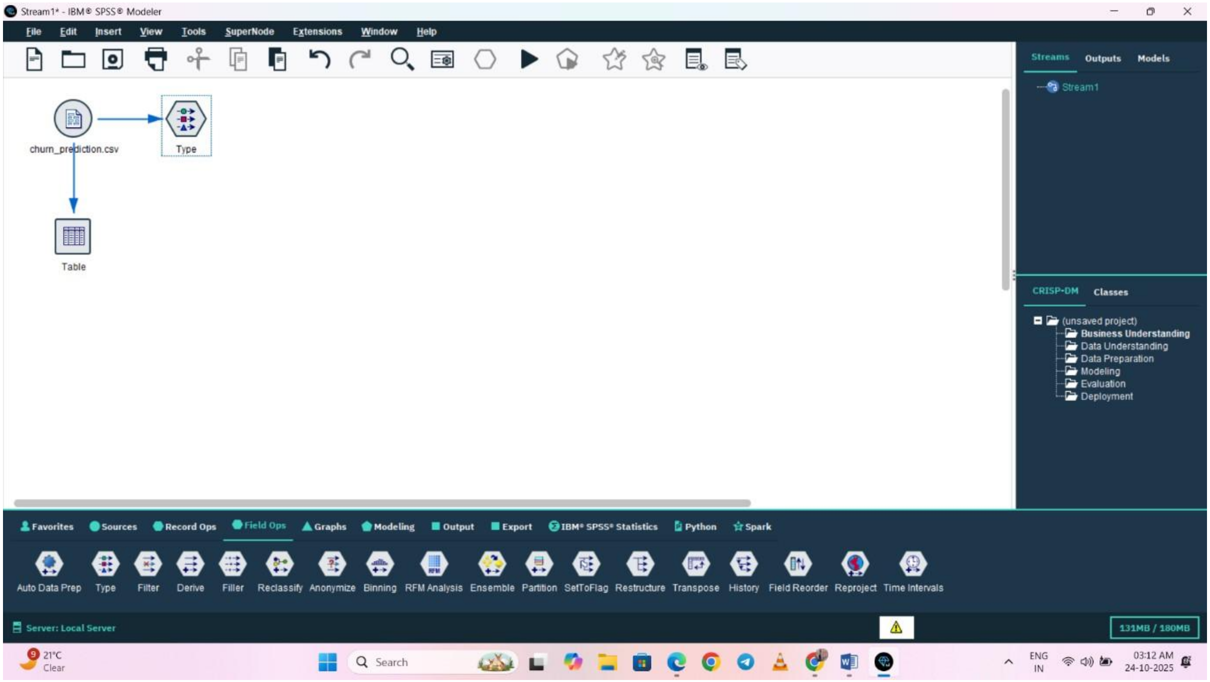
Step 2: Inspect and Prepare Data

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



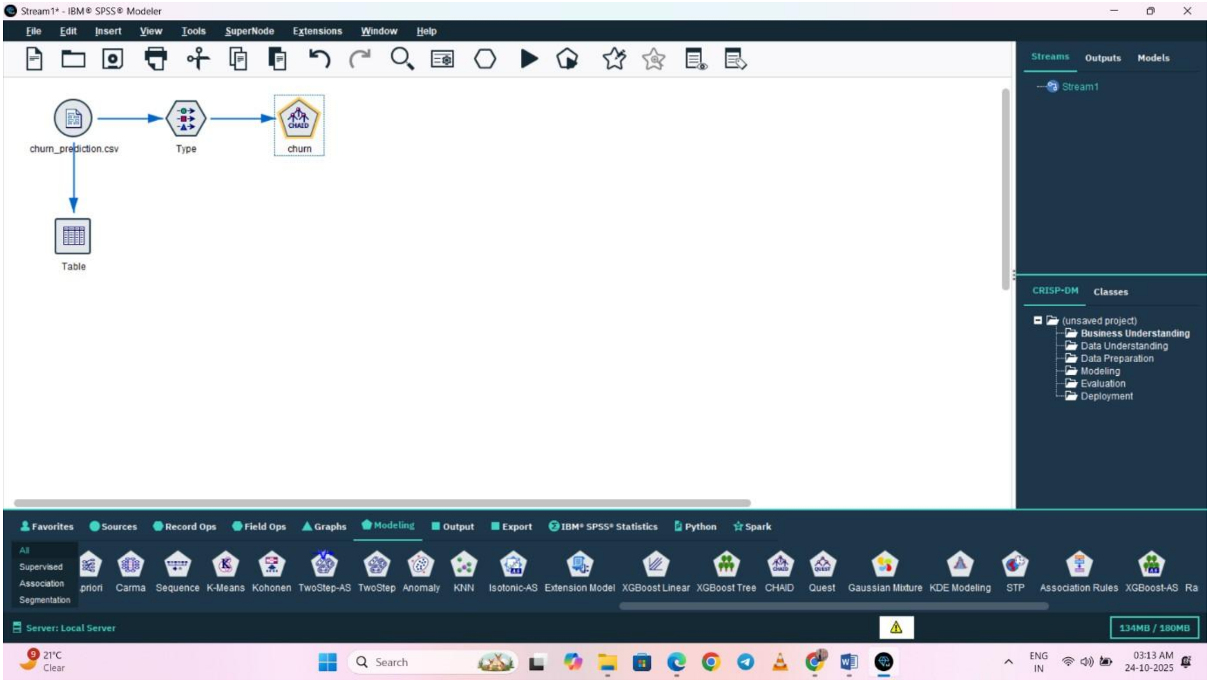
Step 3: Assign Variable Types/Roles

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



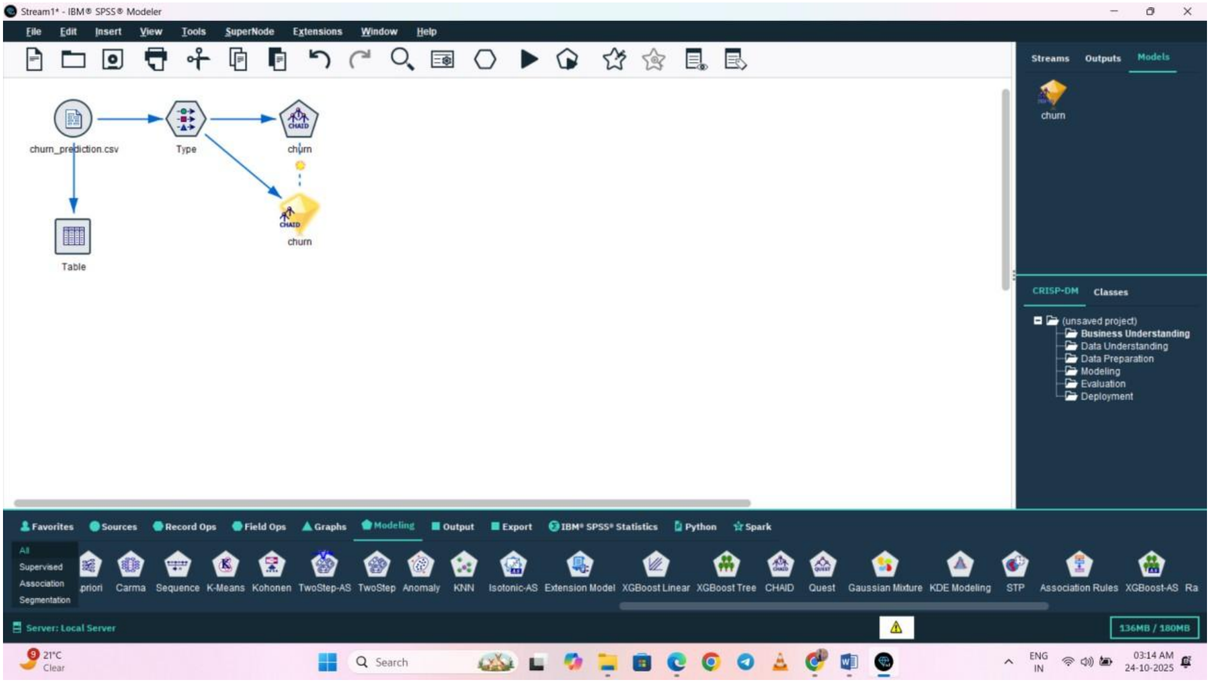
Step 4: Add CHAID Node

Inserted the **CHAID** modeling node and connected it to the Type node output.



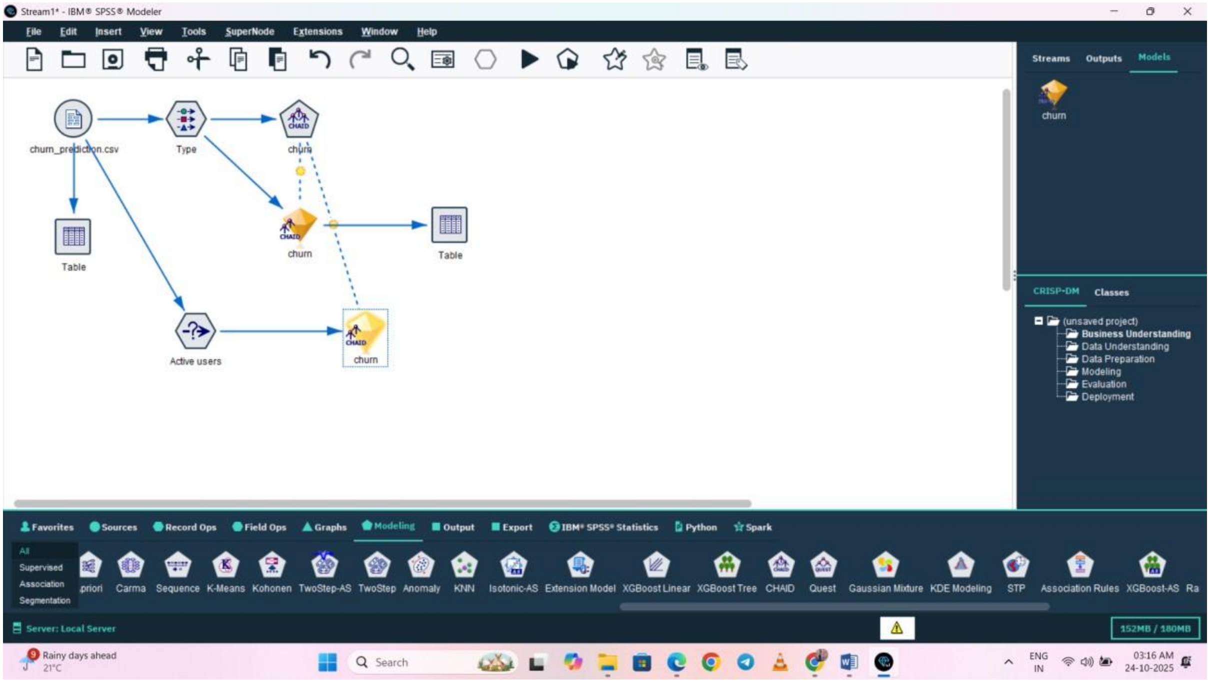
Step 5: Configure CHAID Model

Defined model settings such as significance level, target variable, and maximum depth before saving the configuration.



Step 6: Train the Model (Run CHAID)

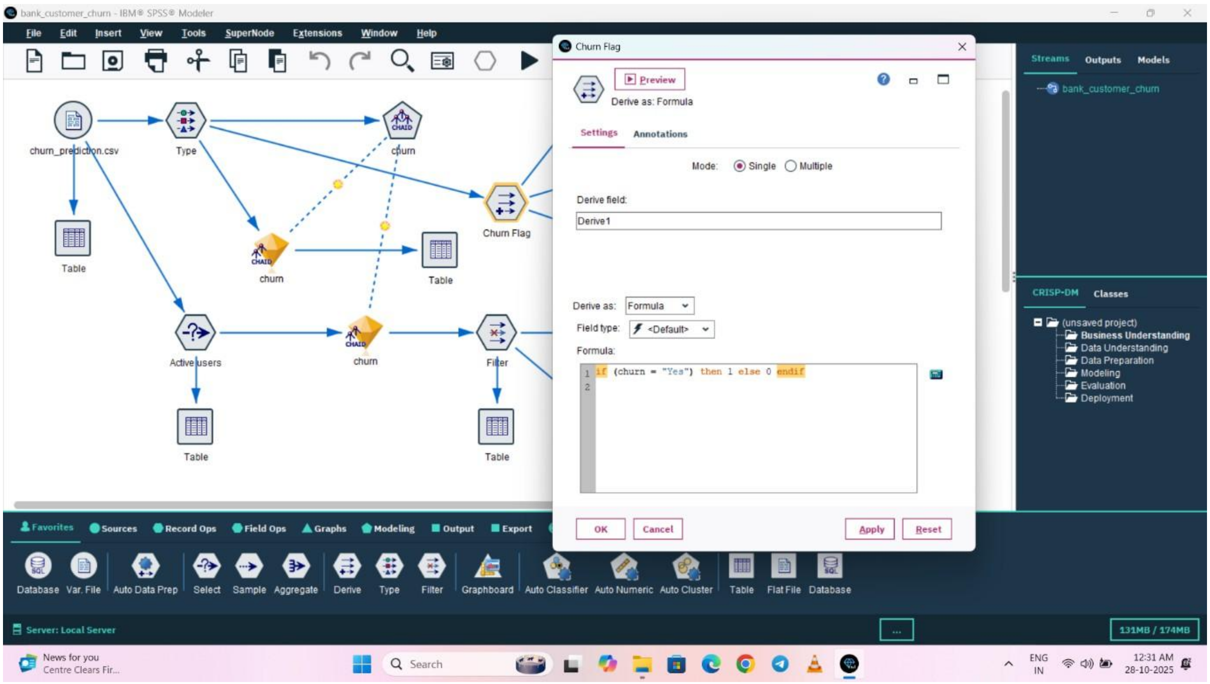
Executed the model stream and generated the CHAID decision tree output.



Step 7: Derive Node Creation

Added a **Derive** node to create a new field `Churn_Flag` using the formula: `if (churn = "Yes") then 1 else 0 endif`

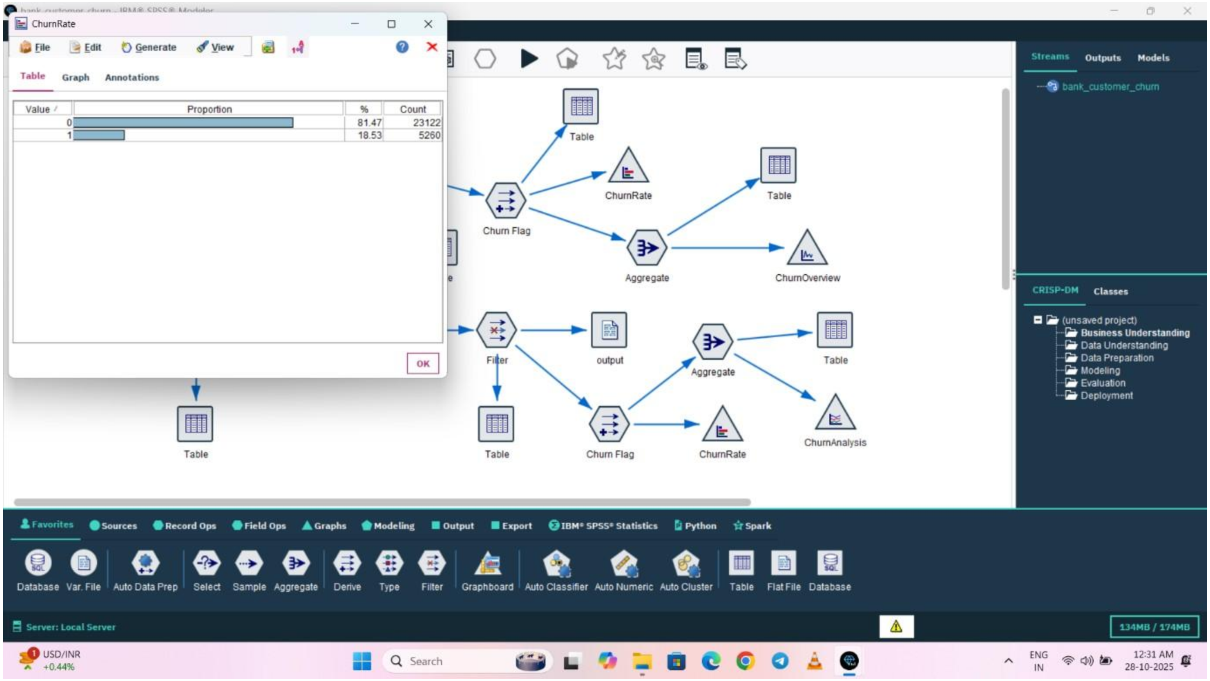
This numeric flag enabled better aggregation and analysis in subsequent nodes.



Step 8: Calculate Churn Rate

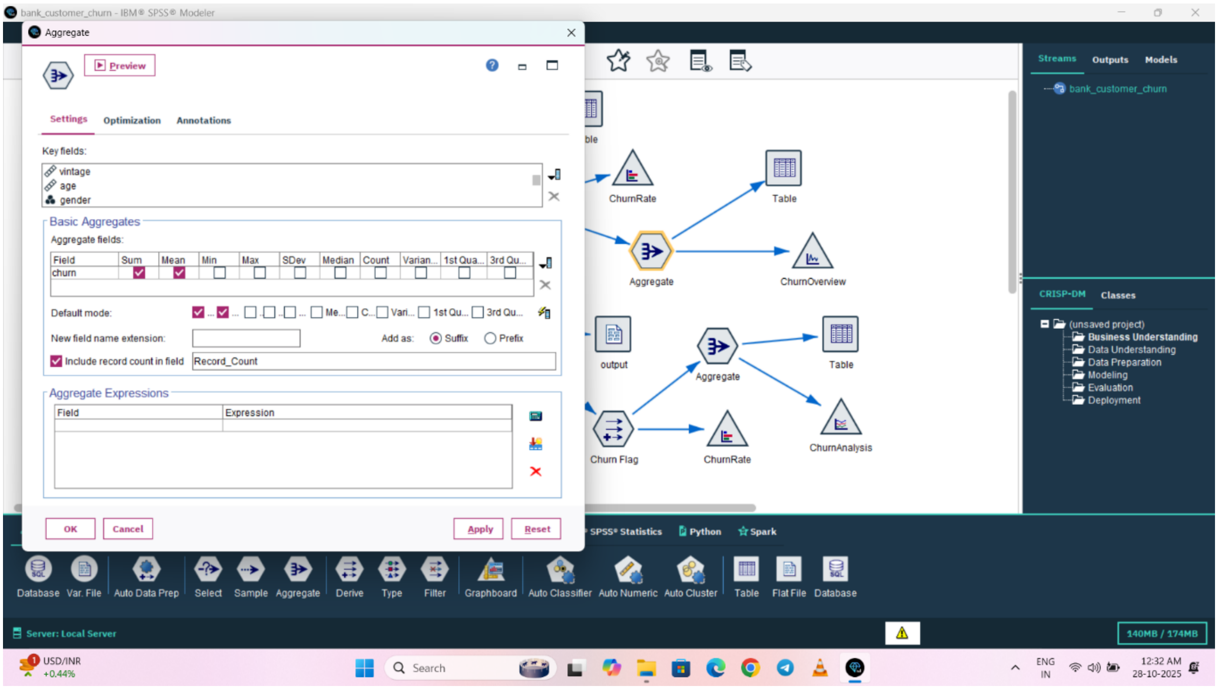
Used **Aggregate** and **Table** nodes to compute churn proportions —

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

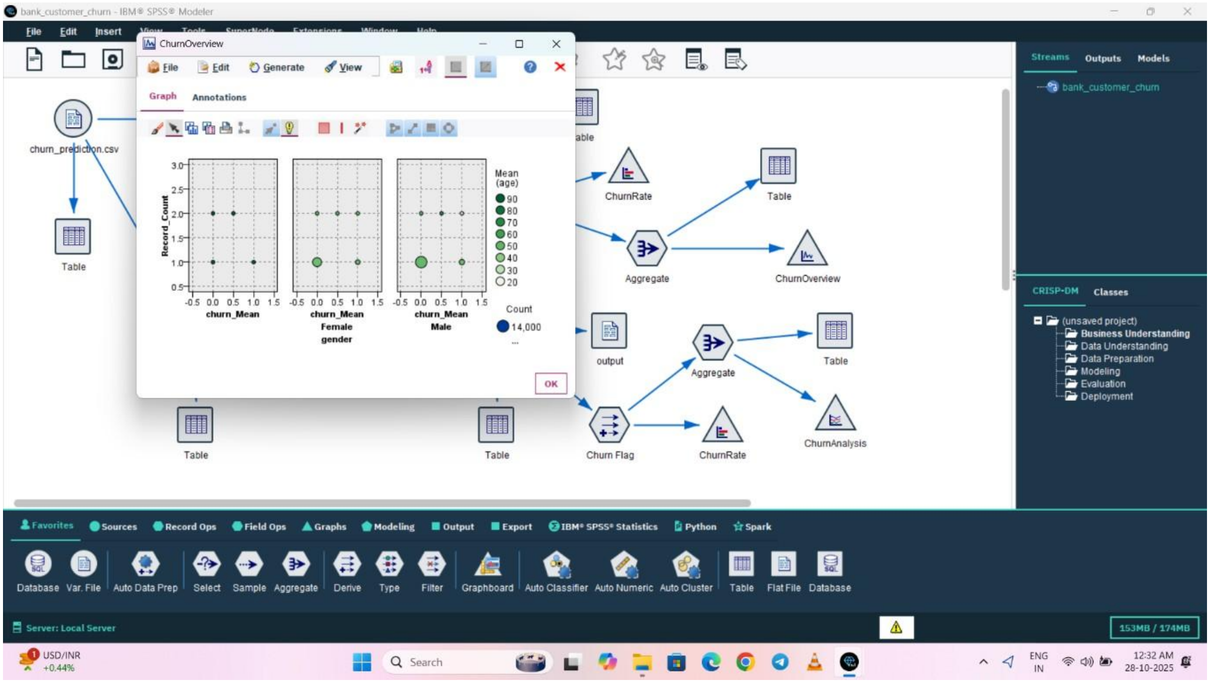


Step 9: Analyze Churn by Demographics

Applied **Aggregate** with `vintage`, `age`, and `gender` as key fields. Calculated mean churn and record count for each group to observe segment-wise trends.

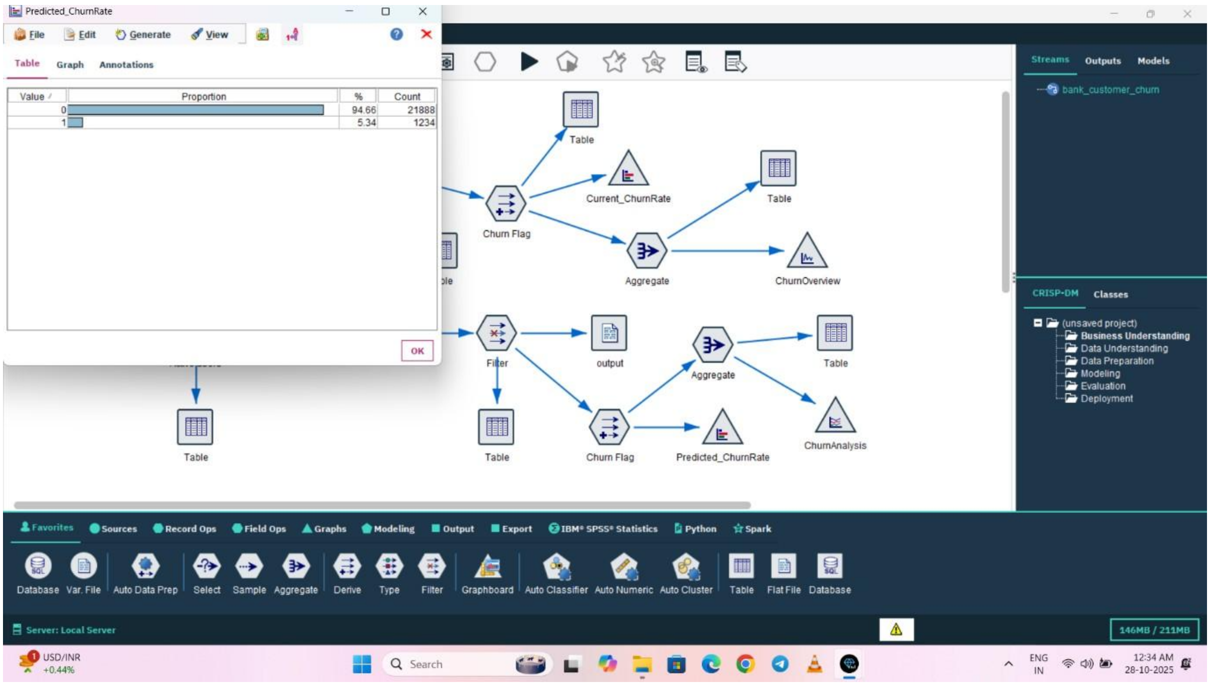


Step 10: Visualize Churn Overview
Displayed a **Bubble Plot** showing churn means by gender and average age, revealing visible patterns in churn tendency.



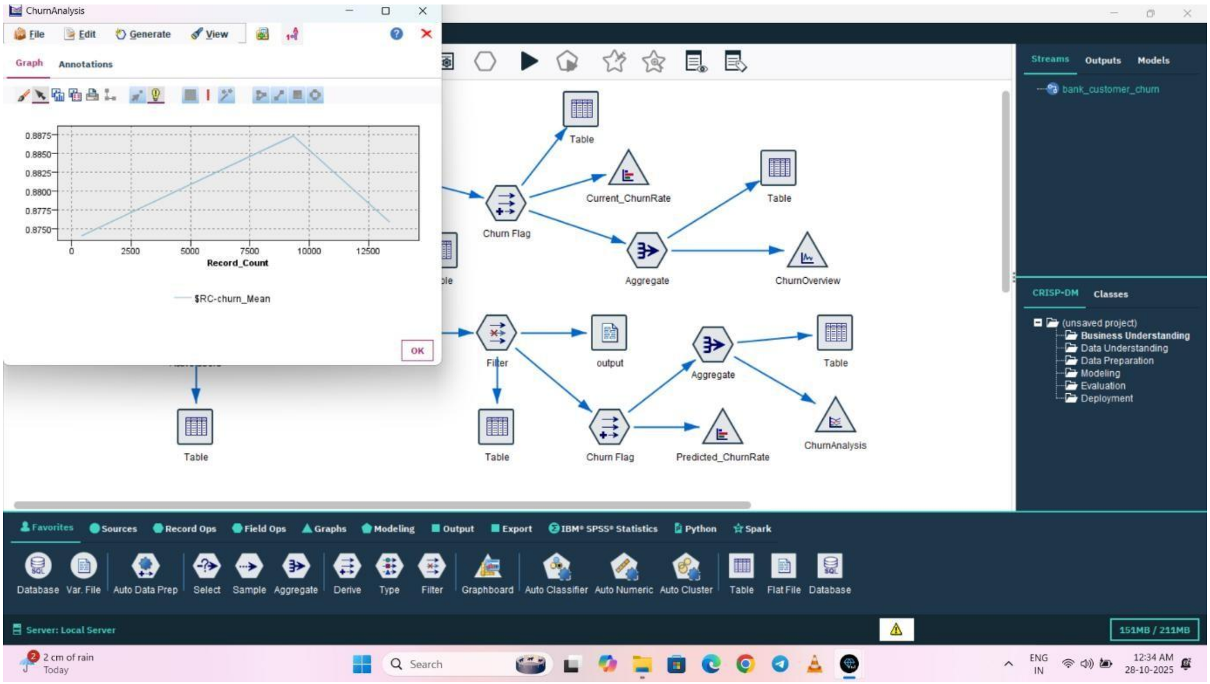
Step 11: Predictive Churn Rate Analysis
Derived **Predicted_ChurnRate** from model output —

- Predicted Non-Churn (0): 94.66%
- Predicted Churn (1): 5.34%



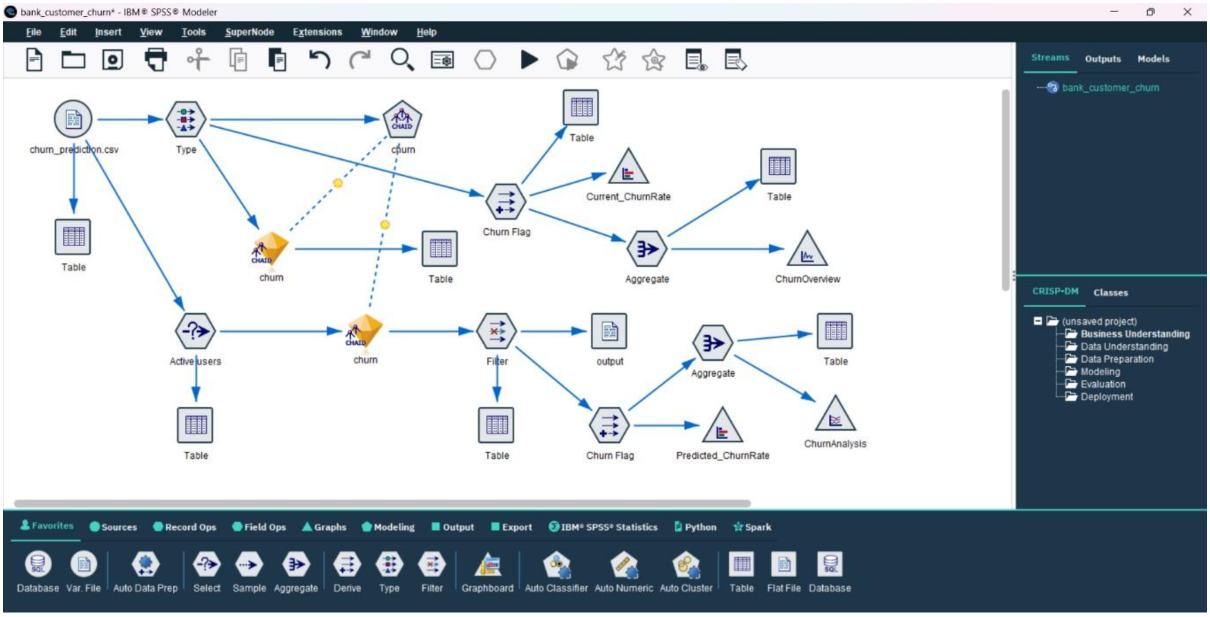
Step 12: Trend Visualization (ChurnAnalysis)

Generated a **Line Graph** displaying churn mean across record counts, reflecting gradual churn variation across customer segments.



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 project **“Bank Customer Churn Prediction Using CHAID Decision Tree Model in IBM SPSS Modeler”** effectively demonstrates how predictive analytics can help identify customers likely to leave a bank. The objective was to analyze customer data, discover key churn factors, and use the CHAID decision tree algorithm to predict churn behavior accurately.

The workflow began with importing and inspecting the dataset, ensuring proper formatting and resolving missing values. Variable roles were assigned through the **Type node**, with the *churn* field defined as the target. A **Derive node** was created to generate a numeric `Churn_Flag`, enabling more precise aggregation and visualization of churn rates. Using **Aggregate** and **Table** nodes, the overall churn rate was calculated — about **18.53% churned** and **81.47% retained** — offering a clear understanding of the data distribution.

The **CHAID model** was then configured with parameters such as maximum tree depth and significance level, producing a decision tree that divided customers into segments based on churn likelihood. The model identified important predictors such as customer age, vintage, and account balance. These insights revealed that newer and younger customers were more prone to churn, while long-tenured clients showed stronger retention.

Further analysis and visualizations — including bubble and line plots — provided a deeper view of churn patterns across demographics and segments. The **final model evaluation** compared actual and predicted churn values, confirming the CHAID model’s reliability. The **stream diagram** summarizes this entire workflow, from data preparation to model interpretation, visually representing how each analytical stage connects within IBM SPSS Modeler.

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.