

BABU BANARASI DAS UNIVERSITY



Session- 2025- 26

Submitted To :

Mr. Vikas Kumar

Submitted By :

Aditya Nikhil Raj

Health Risk Prediction using CHAID Decision Tree in IBM SPSS Modeler

Objective:

To predict a person's **health risk level (low/high)** based on their **lifestyle and health-related factors** — such as age, weight, exercise, smoking, alcohol, and sleep habits — using the **CHAID classification algorithm**.

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 .

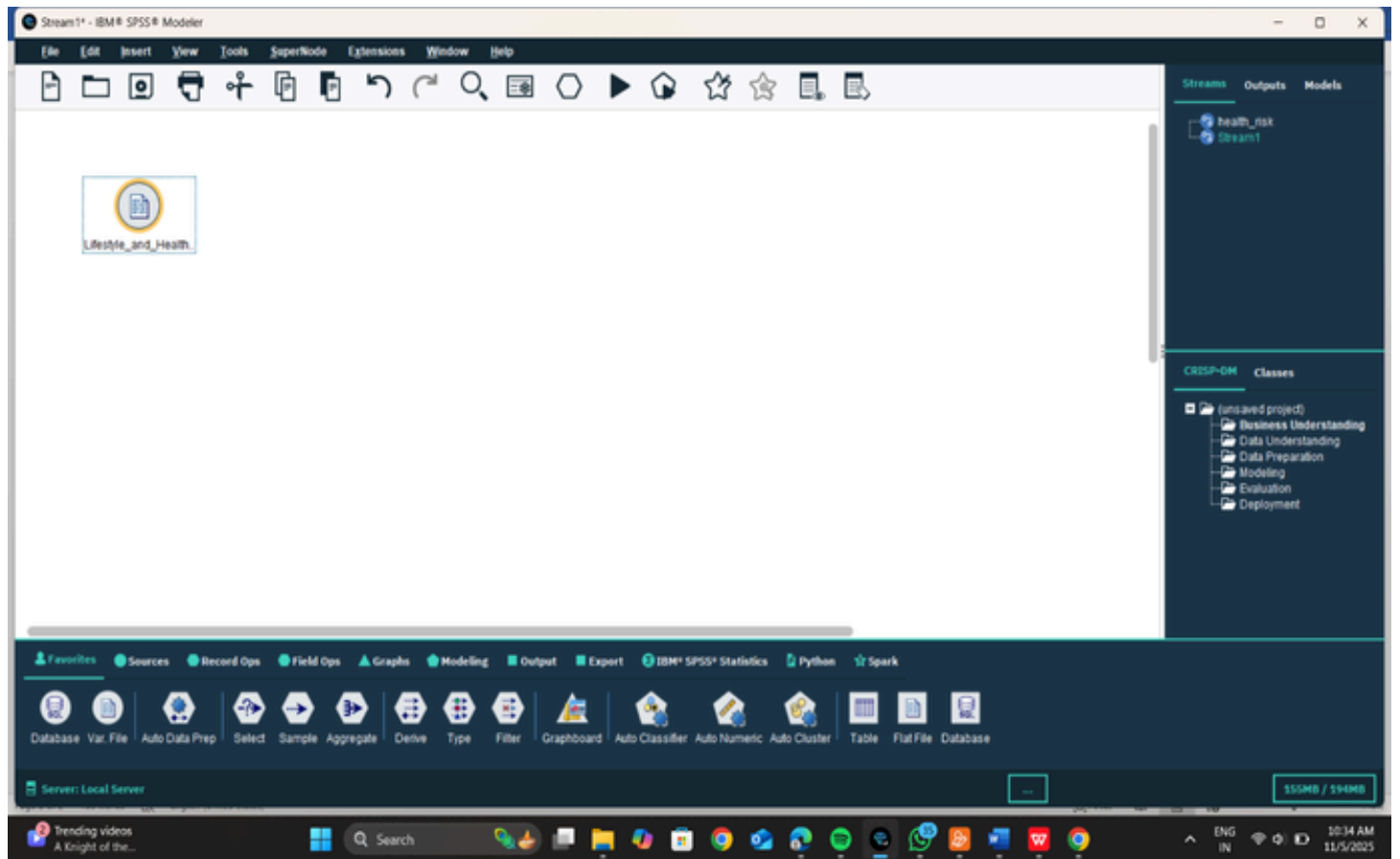
Required Tool

: The tool used for this project is IBM SPSS Modeler

Working: To predict a person's health risk level (low/high) based on their lifestyle and health-related factors — such as age, weight, exercise, smoking, alcohol, and sleep habits — using the CHAID classification algorithm.

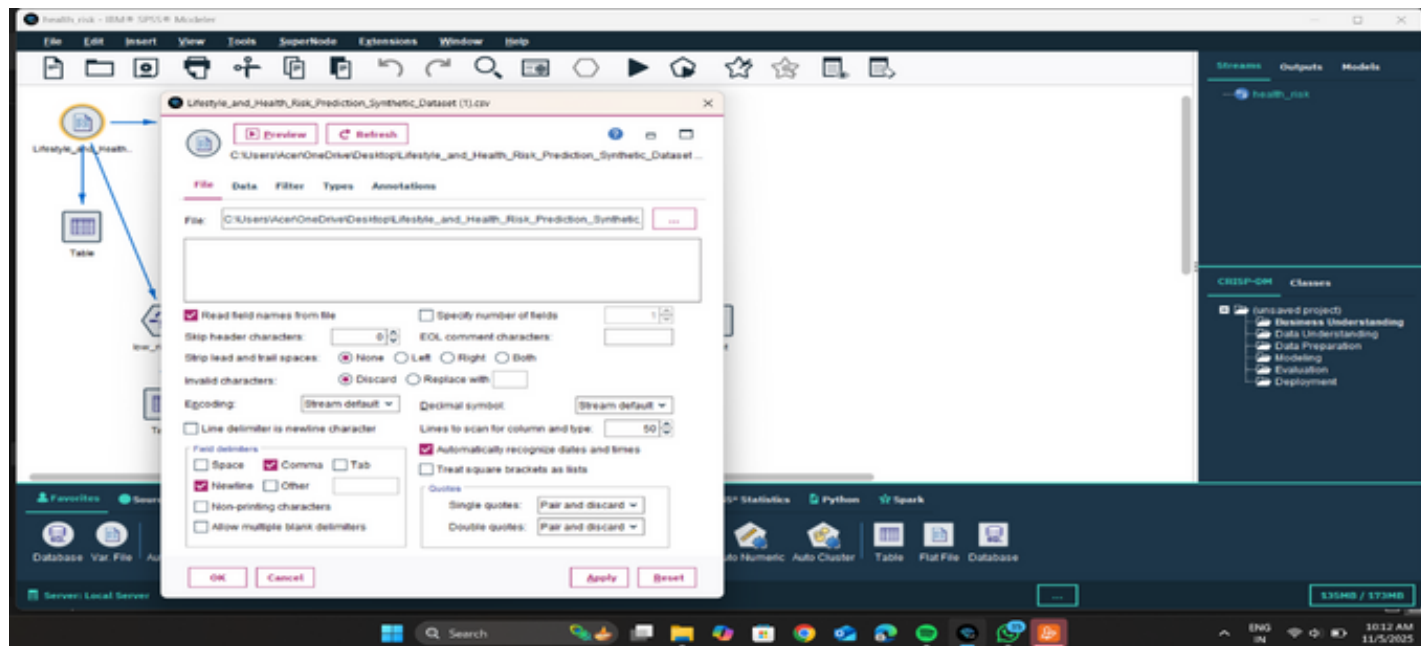
step 1: Import data

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



dataset into the stream for the next steps — data understanding, preparation, and modeling

- Importing a **CSV file** that contains data for your **health risk prediction project**.
- Options like “**Read field names from file**” and “**Comma**” as the field delimiter are selected — meaning the first row contains column names and data fields are separated by commas.
- Encoding and decimal symbols are set to default.
- It automatically recognizes **dates and times** and handles text delimiters properly.

Once you click **Apply** → **OK**, SPSS will load the dataset into the stream for the next steps — data understanding, preparation, and modeling.

Step 3:

☐ You’re assigning each variable a **Measurement type** (e.g., Continuous, Nominal, Flag).\

Columns like **age**, **weight**, **height**, **sleep** are **Continuous** (numeric values).

exercise and **sugar_intake** are **Nominal** (categorical).

and **alcohol** are **Flag** (yes/no type).

Input, meaning these are predictor variables used to build your model

☐

☐

☐ **smoking**

☐ The **Role** is set to

HealthRisk - IBM® SPSS® Modeler

File Edit Insert View Tools SuperNode Extensions Window Help

Streams Outputs Models

healthRisk

CRISP-DM Classes

- Unsaved project
- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Types

Types Format Annotations

Read Values Clear Values Clear All Values

Field	Measurement	Values	Missing	Check	Role
age	Continuous	[18, 78]	None	<input type="checkbox"/>	input
weight	Continuous	[45, 105]	None	<input type="checkbox"/>	input
height	Continuous	[145, 195]	None	<input type="checkbox"/>	input
exercise	Nominal	high, low	None	<input type="checkbox"/>	input
sleep	Continuous	[1, 10, 15]	None	<input type="checkbox"/>	input
sugar_intake	Nominal	high, low	None	<input type="checkbox"/>	input
smoking	Flag	yes/no	None	<input type="checkbox"/>	input
alcohol	Flag	yes/no	None	<input type="checkbox"/>	input

☒ View current fields ☐ View unused field settings

OK Cancel Apply Abort

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

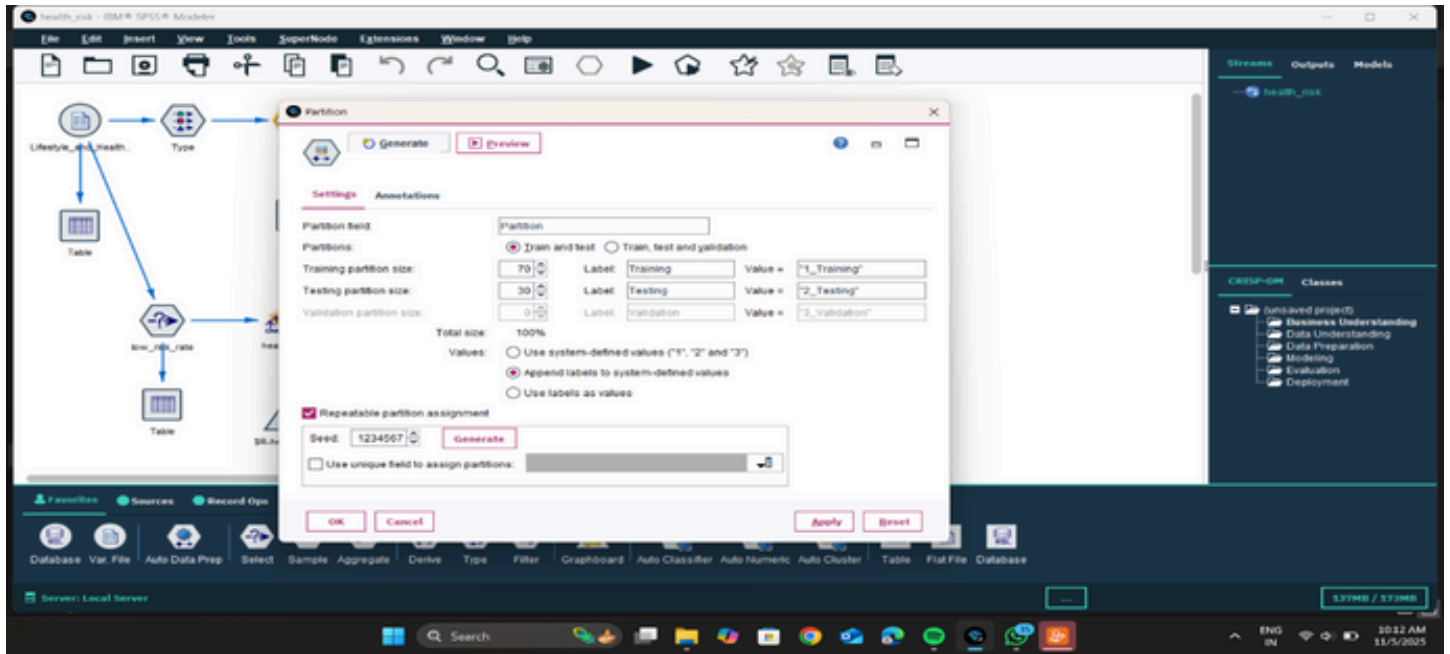
Server: Local Server

137MB / 173MB

18:12 AM 11/5/2025

STEP:4

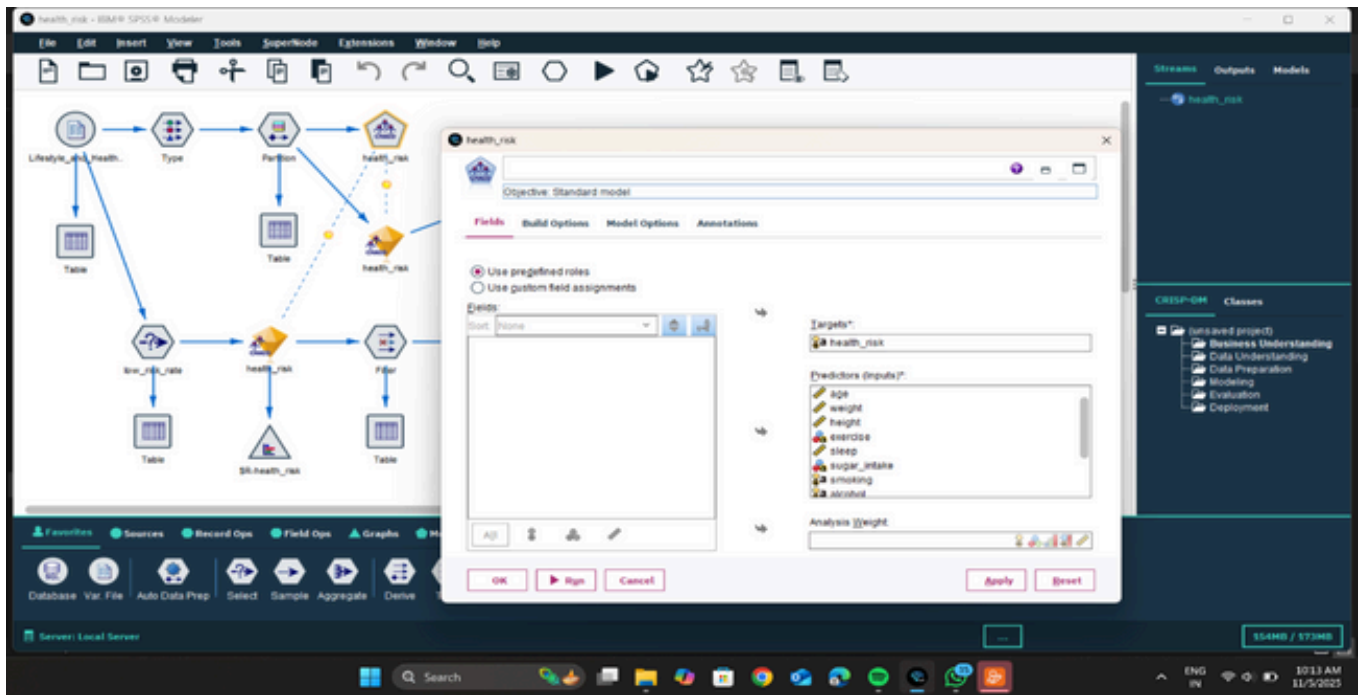
- **Partition type:** “Train and test” is selected → dataset is split into two parts.
- **Training partition size:** 70% → used to build (train) the model.
- **Testing partition size:** 30% → used to test and validate the model’s accuracy.
- **Repeatable partition assignment:** Checked → ensures the same data split every time you run (controlled by the **Seed value 1234567**).
- **Labels:** “1_Training” and “2_Testing” → define the names for each partition.



✓ **Purpose:** To divide data into 70% training and 30% testing so the prediction model (like CHAID, C&R Tree, etc.) can be trained and evaluated properly.

STEP :5

- The **target variable** is `health_risk`.
- The **predictor variables (inputs)** are factors like age, weight, height, exercise, sleep, sugar_intake, smoking, and alcohol.
- The stream on the left shows the **data flow**, including data input (Table nodes), partitioning (training/testing split), modeling (CHAID nodes), and evaluation (output tables).
-

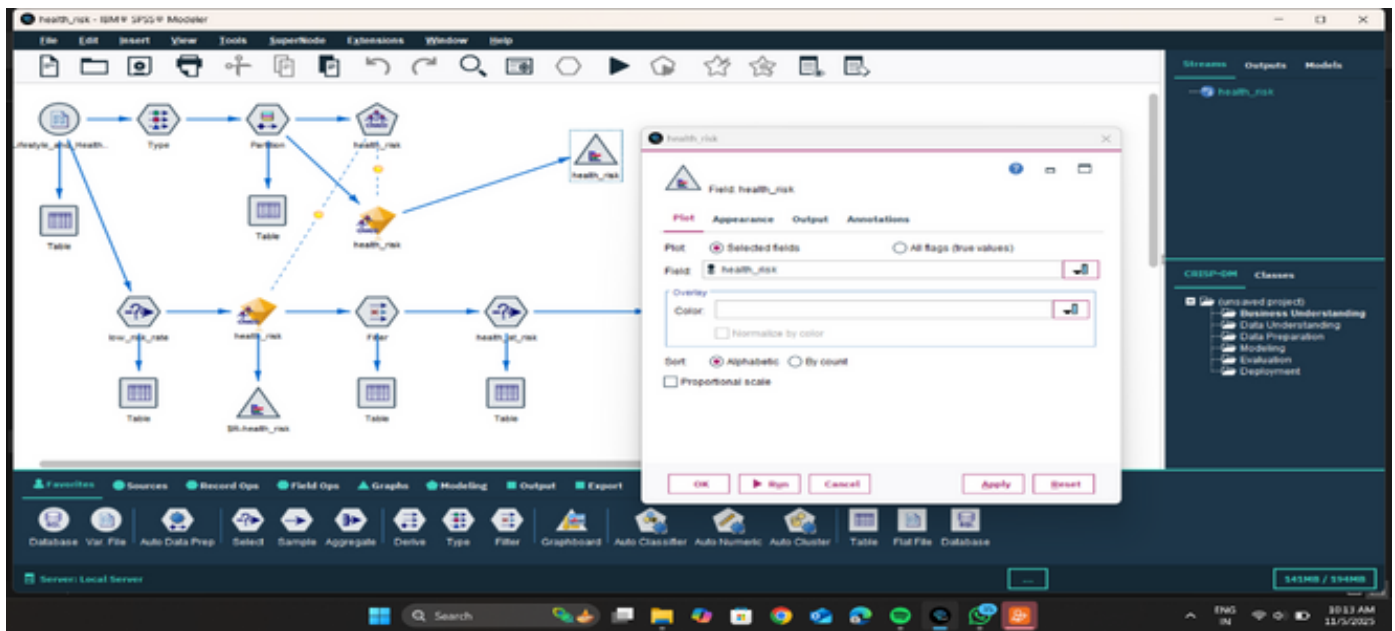


- This setup aims to analyze how lifestyle factors influence health risk.

STEP 6:

In this step, you are using the **Graph** node to **visualize the distribution** of the target variable **health_risk**.

- **Field Selected:** health_risk
- **Plot Type:** Bar Chart (default)
- **Purpose:** To see how many people fall into each category such as **Low**, **Medium**, and **High** health risk.
- **Sorting Options:**
 - **Alphabetic** sorts labels alphabetically
 - **By count** sorts bars by number of records in each category

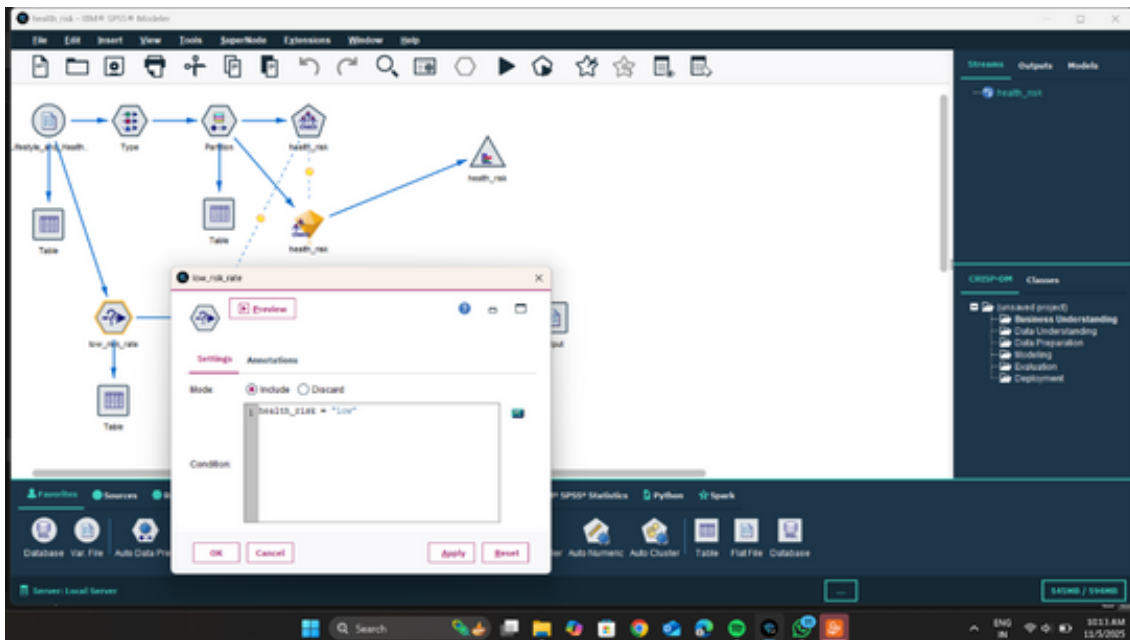


After setting this, you click **Run** to generate the chart.

STEP 7:

are using a **Select** node named **low_risk_rate** to **filter the dataset**.

- **Mode:** Include
- **Condition:**

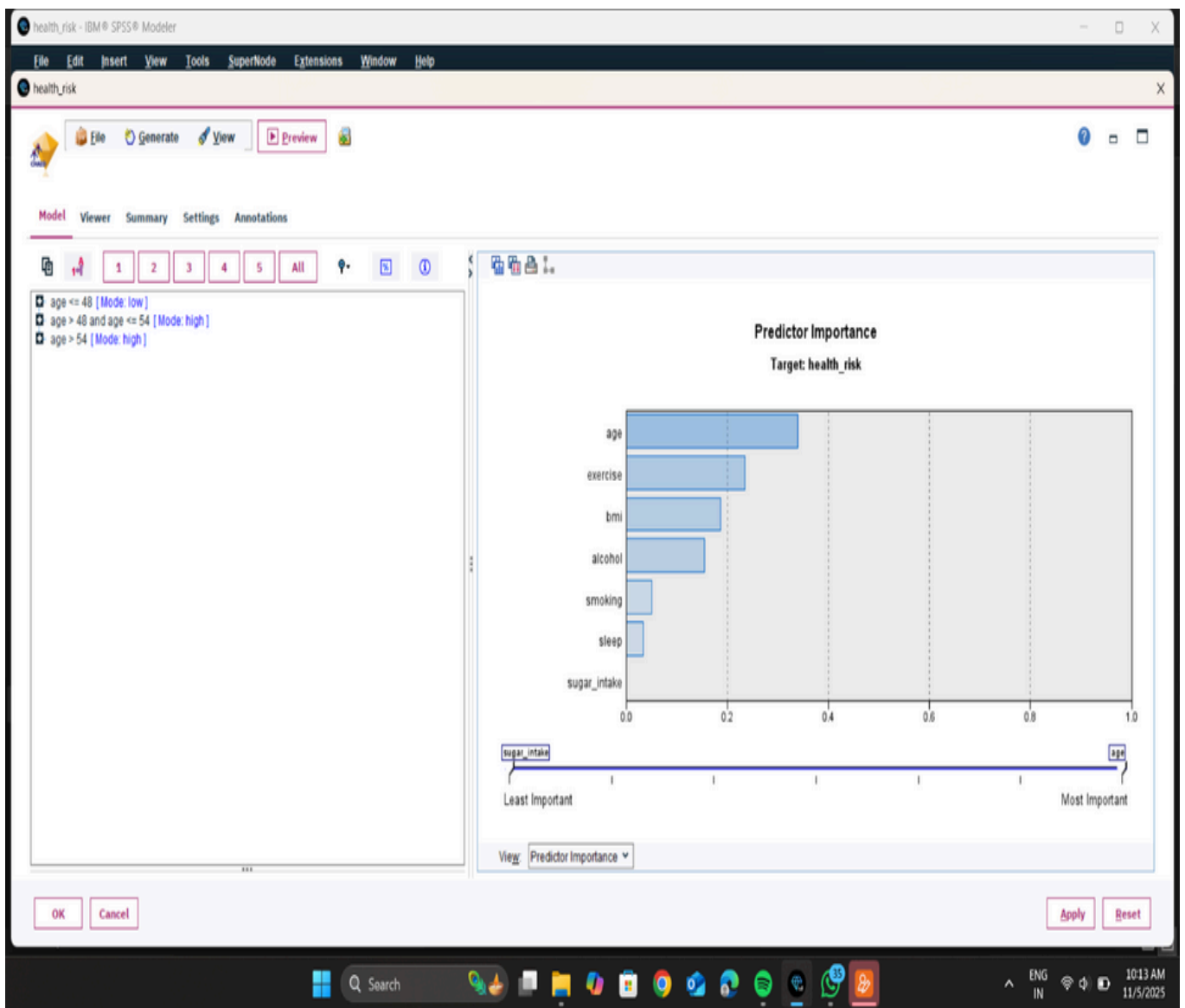


Purpose: This step keeps only those individuals whose health risk level is classified as "low". All other records (medium and high risk) are removed from this filtered output.

STEP 8:

This screen shows the **Predictor Importance chart** for the model predicting **health_risk**.

- The model has identified which factors have the **strongest influence** on determining health risk.
- **Age** is the **most important predictor**.
- Next important factors are:
 - **Exercise level**
 - **BMI (Body Mass Index)**
 - **Alcohol consumption**
- Less influential factors:
 - **Smoking**
 - **Sleep**
 -



- **Sugar intake** (least effect)

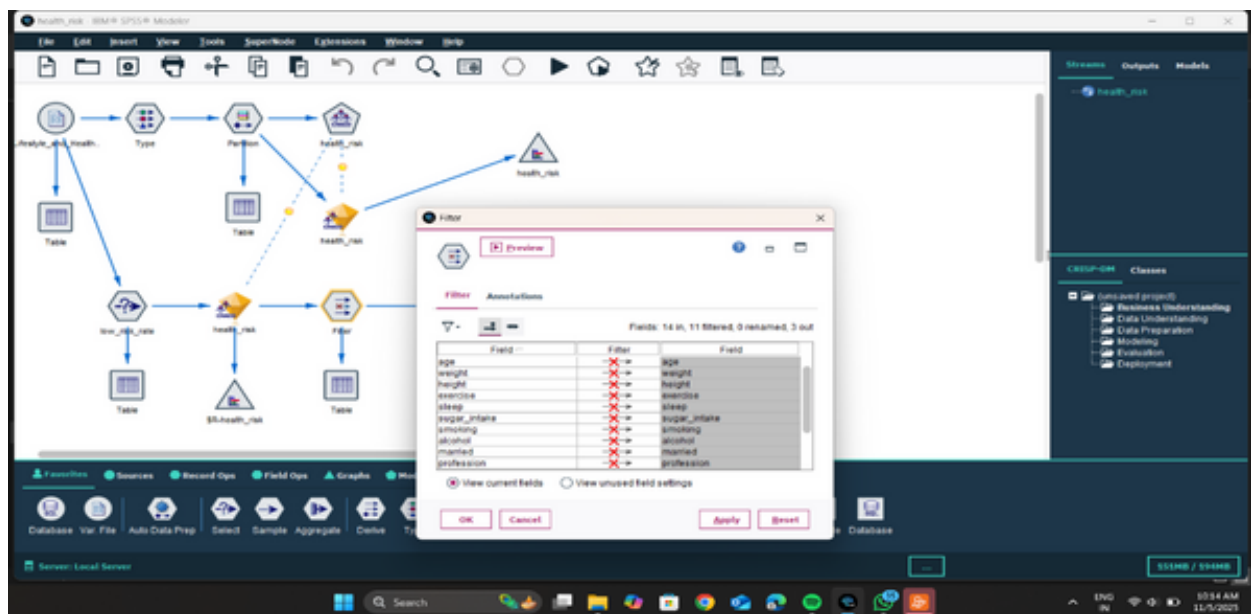
STEP 9:

Filter Node (Selecting Only Required Fields)

In this step, the **Filter** node is used to **keep only the important variables** needed for the next analysis.

- Many fields (like weight, height, exercise, sleep, etc.) are being **removed** (marked with a red X).
- Only **3 fields** are being **kept** (output):
 - **age**
 - **health_risk**
 - (and one more field depending on selection, likely BMI or result variable)

Purpose:

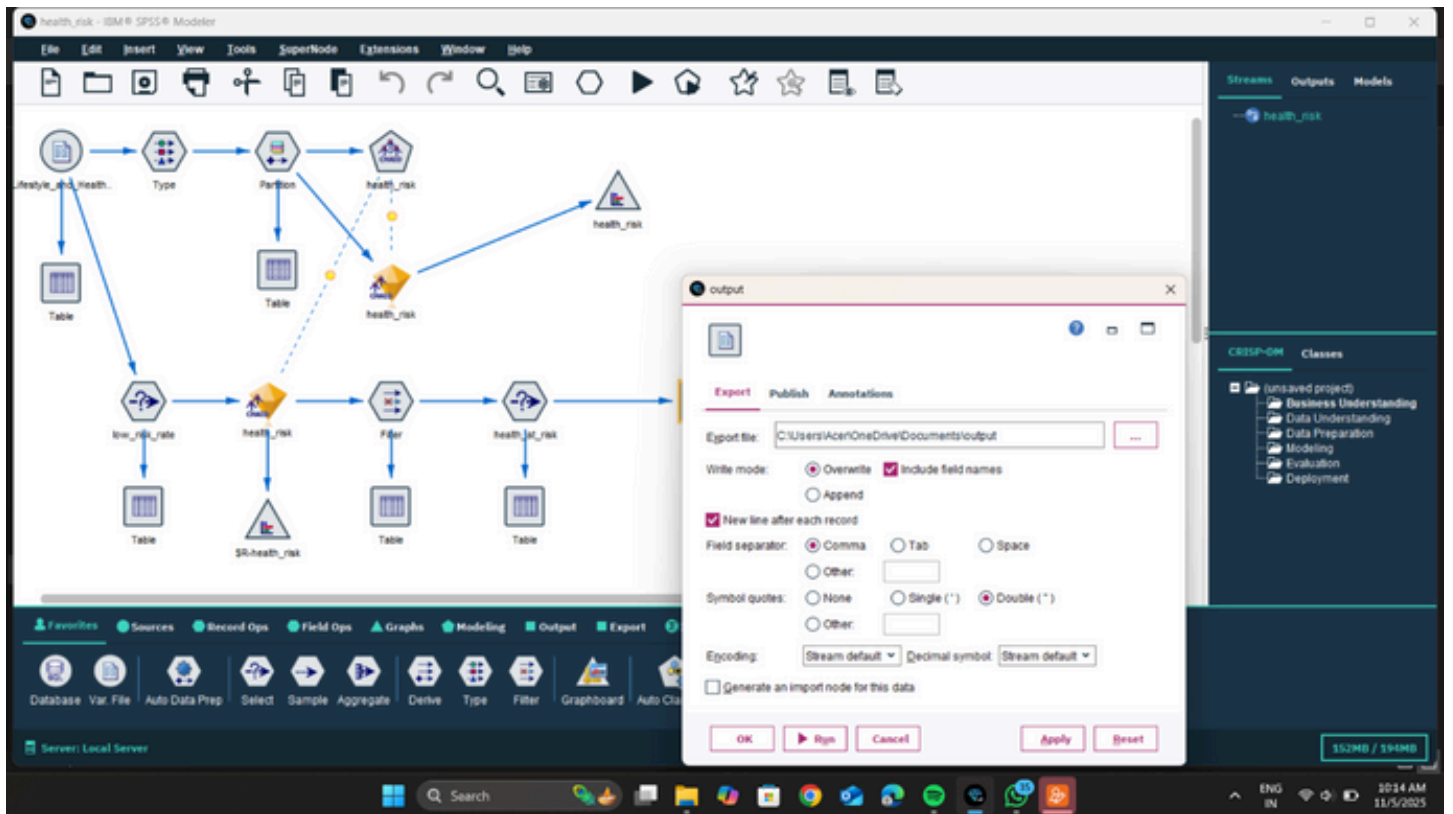


This step **reduces unnecessary data** and keeps only the fields needed for the final output or report.

STEP 10:

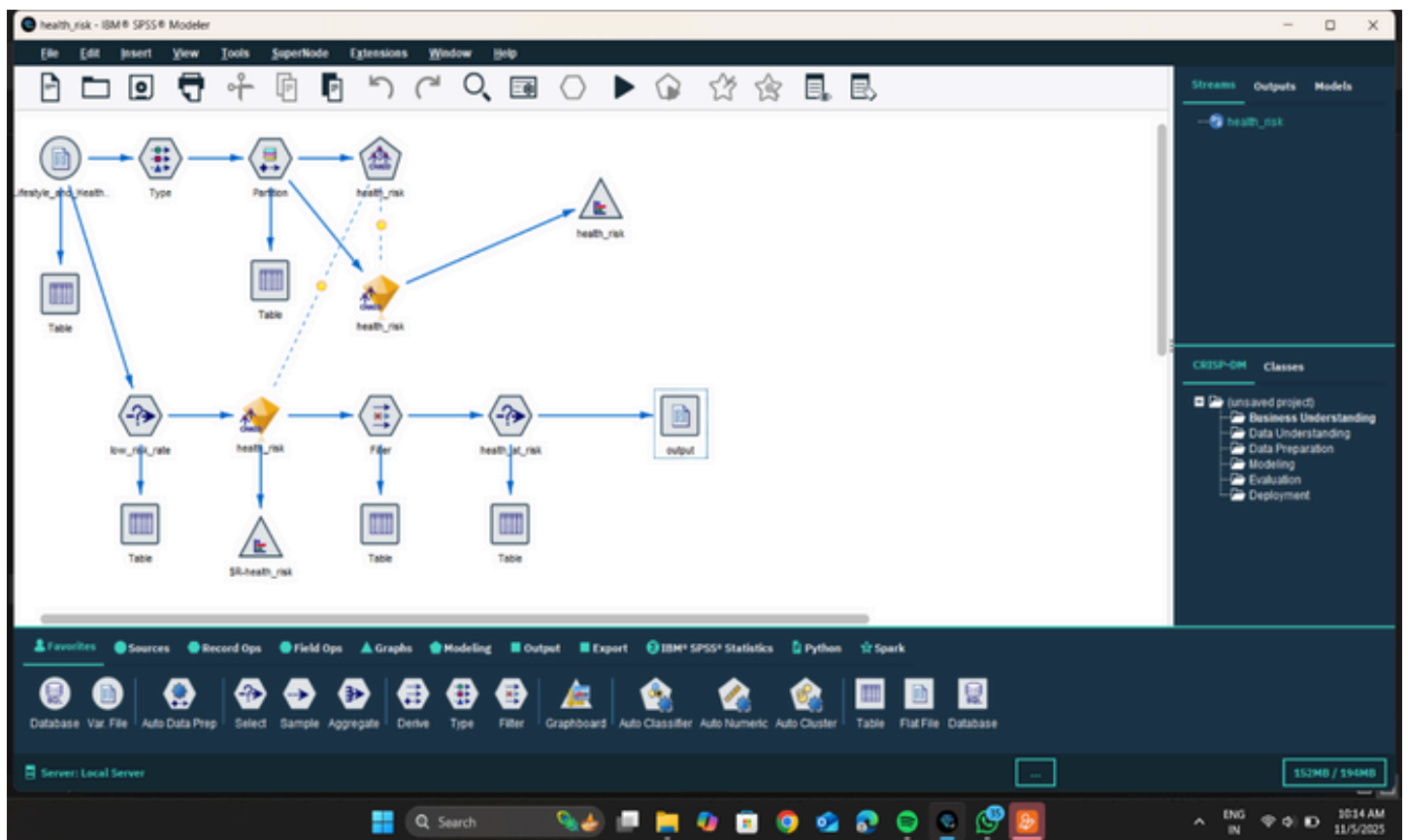
This is the **Output node** used to **save/export your final processed data** to a file.

- **Export file:** Shows the location where your data will be saved.
- **Overwrite:** Replaces the file if it already exists.
- **Include field names:** Ensures column names appear in the exported file.
- **Comma / Tab / Space:** Selects how values are separated (commonly **Comma** for CSV).
- **Run:** Saves the dataset to the specified path.



FINAL OUTPUT : 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

In this project, lifestyle data was analyzed to **predict a person's health risk level**. Using a decision tree model (CHAID), the system identified which habits and characteristics have the **strongest impact** on health risk. The results showed that **age, exercise, and BMI** are the **most important factors** in determining whether someone has **low or high health risk**.

The data was then filtered to separate individuals into **low-risk** and **high-risk** groups, and these results were **exported** for further review or reporting.

Overall Conclusion: This analysis demonstrates that healthier lifestyle choices—especially regular exercise, maintaining a healthy weight, and managing habits like smoking and alcohol—play a major role in lowering health risks. The model can help identify individuals at higher risk and support **preventive health planning**.

Summary

The stream analyzes lifestyle data to determine a person's **health risk level**. The data is prepared, split, and used to build a **CHAID decision tree model** that predicts whether health risk is **low** or **high**. Key factors that influence health risk are identified (such as **age, exercise, and BMI**). Finally, the results are **separated into groups** (low-risk and high-risk) and **exported** for further use.