Babu Banarasi Das University

BBD City,Faizabad Road,Lucknow UttarPradesh

PROJECT REPORT – Insurance Fraud Prediction
Using IBM SPSS Modeler

SUBMITTED TO:
Mr. AYUSHMAN BHADAURIA

SUBMITTED BY:
VISHWAJIT VISHWAS

Insurance Fraud Detection Using C&R Tree Algorithm

Agenda / Definition

The project aims to detect fraudulent insurance claims using the C&R Tree (Classification (Classification and Regression Tree) method in IBM SPSS Modeler.

By analyzing claim data (such as vehicle details, claim amount, and customer info), the model identifies patterns and predicts whether a claim is fraudulent (Y) or non-fraudulent (N).

Outcomes / Learning

- Import and explore a dataset in IBM SPSS Modeler
- Perform data cleaning (remove irrelevant columns, handle missing values)
- Partition data into training and testing samples
- Build and evaluate a C&R Tree classification model
- Generate and interpret prediction results and graphs

This project demonstrates the full data mining workflow — from preparation to model evaluation.

Required Tools

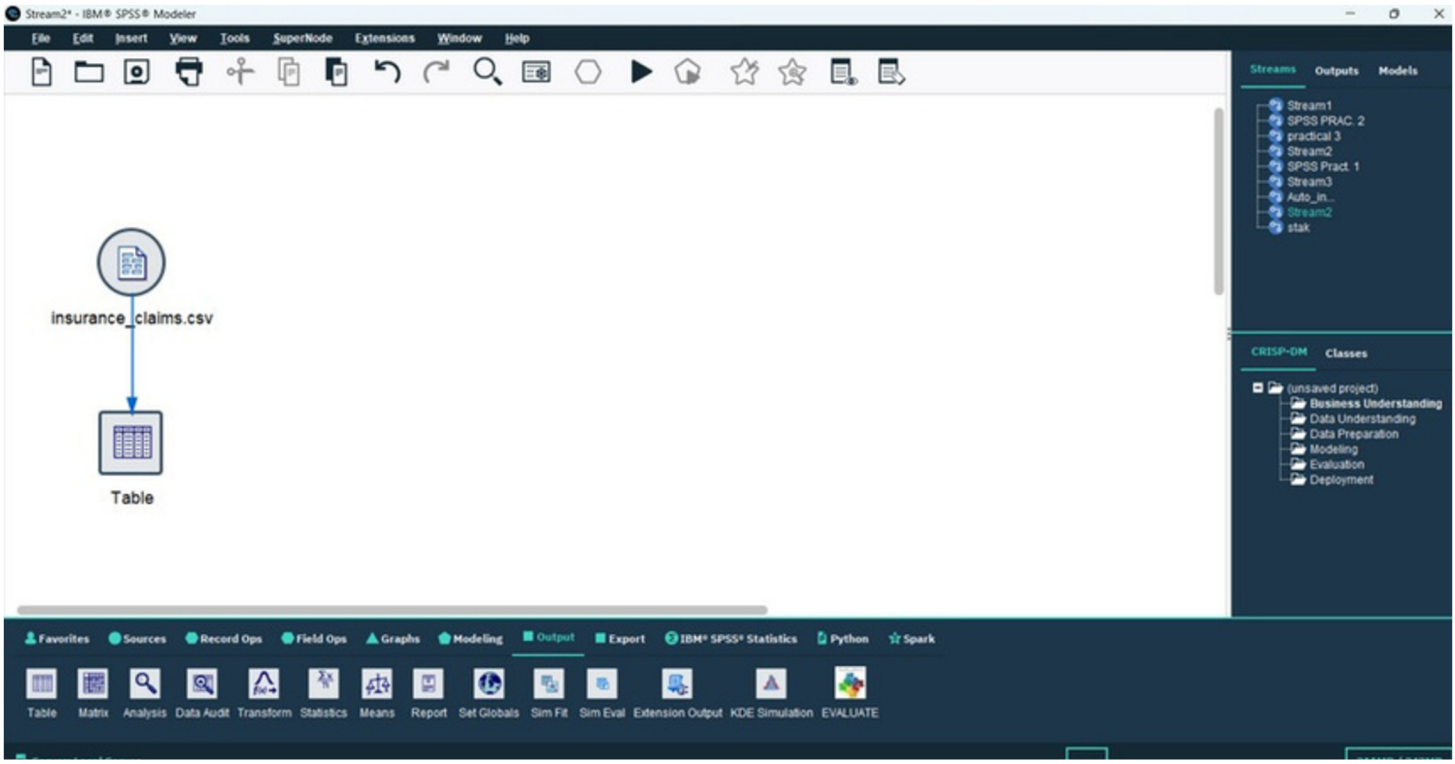
The tool used for this project is IBM SPSS Modeler.

Working

- **The project involves:**
- Importing the insurance claim dataset
  Cleaning and preparing the data
  Setting variable roles and partitioning data
  Configuring and running the C&R Tree model
- Viewing prediction results in a table and histogram
-

Step 1: Import Data

Loaded the dataset (insurance_claims.csv) into SPSS Modeler using the Var.File Node under Sources palette After reading metadata, all fields were correctly recognized.



Step 2: Remove unecessary Data

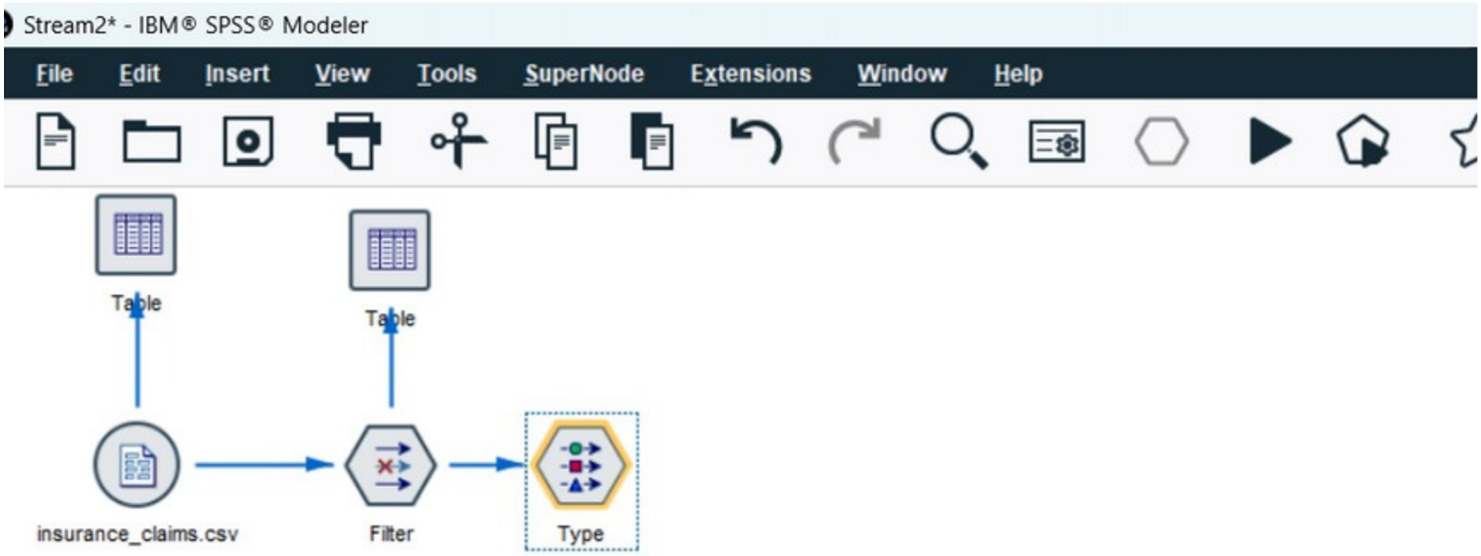
The Filter Node was used to exclude the irrelevant column _c39 from the dataset.
This column contained empty or meaningless values that could interfere with model accuracy.
By filtering it out, we ensured that only useful fields (such as claim amount, vehicle claim, auto make, model, year, and fraud status) were retained for analysis.

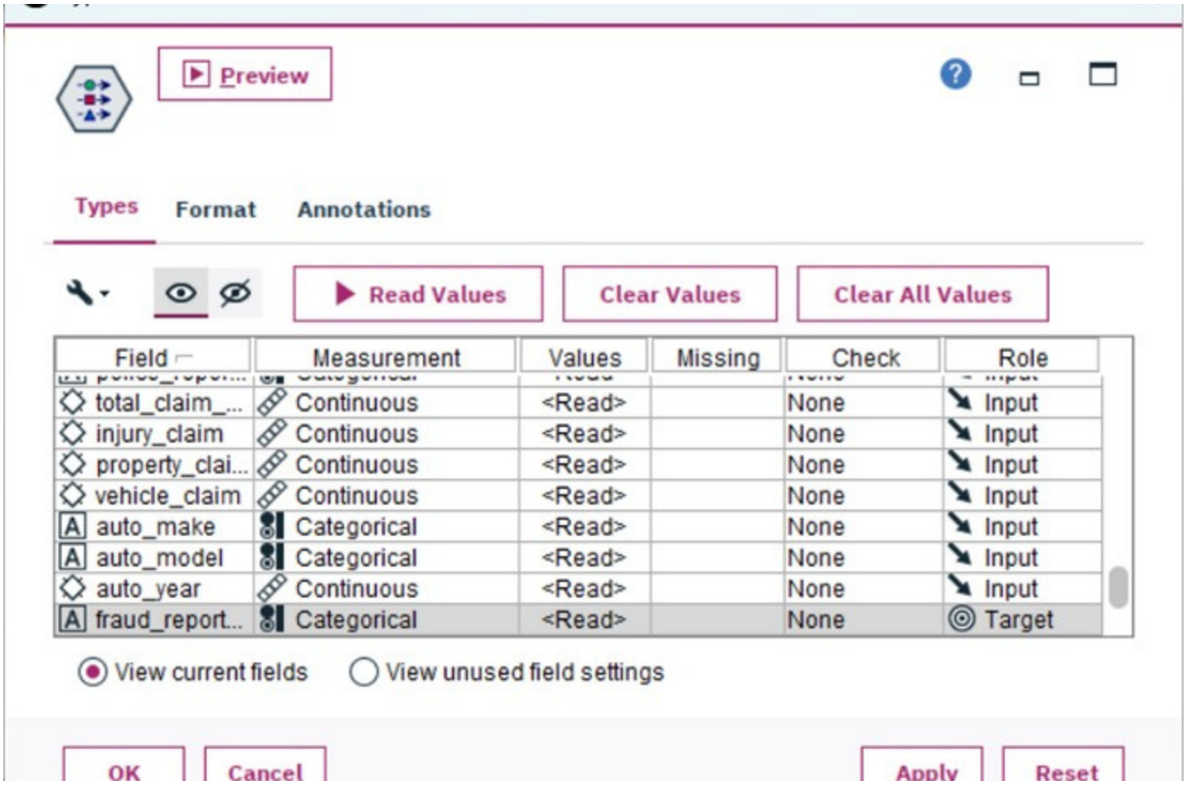
	witnesses	police_report_available	total_claim_amount	injury_claim	property_claim	vehicle_claim	auto_make	auto_model	auto_year	fraud_reported
1	2	YES	71610	6510	13020	52080	Saab	92x	2004	Y
0	0	?	5070	780	780	3510	Mercedes	E400	2007	Y
2	3	NO	34650	7700	3850	23100	Dodge	RAM	2007	N
1	2	NO	63400	6340	6340	50720	Chevrolet	Tahoe	2014	Y
0	1	NO	6500	1300	650	4550	Accura	RSX	2009	N
0	2	NO	64100	6410	6410	51280	Saab	95	2003	Y
0	0	?	78650	21450	7150	50050	Nissan	Pathfinder	2012	N
2	2	YES	51590	9380	9380	32830	Audi	A5	2015	N
1	1	YES	27700	2770	2770	22160	Toyota	Camry	2012	N
2	1	?	42300	4700	4700	32900	Saab	92x	1996	N
1	2	?	67010	7910	15820	43280	Ford	F150	2002	N
1	2	YES	114920	17680	17680	79560	Audi	A3	2006	N
1	0	NO	56520	4710	9420	42390	Saab	95	2000	N
1	1	NO	7280	1120	1120	5040	Toyota	Highlander	2010	N
0	2	YES	46200	4200	8400	33600	Dodge	Neon	2003	Y
0	0	NO	63120	10520	10520	42080	Accura	MDX	1999	Y
1	2	YES	52110	5790	5790	40530	Nissan	Maxima	2012	N
0	2	YES	77880	14160	7080	56640	Suburu	Legacy	2015	N
1	0	NO	72930	6630	13260	53040	Accura	TL	2015	N
2	0	NO	60400	6040	6040	48320	Nissan	Pathfinder	2014	N
1	0	?	47160	0	5240	41920	Suburu	Impreza	2011	N
1	2	?	37840	0	4730	33110	Accura	RSX	1994	N
0	0	YES	71520	17880	5960	47680	Suburu	Forrestor	2000	Y
2	2	?	98160	8180	16360	73620	Dodge	RAM	2011	Y
1	3	NO	77880	7080	14160	56640	Ford	Escape	2005	N
1	3	YES	71500	14500	11000	44000	Ford	Escape	2006	Y
1	3	YES	9020	1640	820	6560	Toyota	Camry	2005	N
2	1	?	5720	1040	520	4160	Suburu	Forrestor	2003	Y
1	0	YES	69040	7760	15520	46560	Dodge	Neon	2009	N
2	2	NO	91650	14100	14100	63450	Accura	TL	2011	N
0	0	?	75600	12600	12600	50400	Toyota	Corolla	2005	N
2	2	?	67140	7460	7460	52220	Ford	F150	2006	Y
2	3	NO	29790	3310	3310	23170	BMW	3 Series	2008	N
1	2	?	77110	14020	14020	49070	Suburu	Impreza	2015	N
0	1	YES	64800	10800	5400	48600	Audi	A3	1999	N
2	0	YES	53100	10620	5310	37170	Mercedes	C300	1995	Y
1	1	YES	60200	6020	6020	48160	Suburu	Forrestor	2004	Y
1	1	YES	5330	1230	820	3280	Suburu	Legacy	2001	N
2	0	?	62300	12460	6230	43610	Jeep	Wrangler	2007	N

Step 4 : Type Node

Defined roles for each field:

- Input Fields: Claim amount, incident severity, age, etc.
- Target Field: fraud_reported





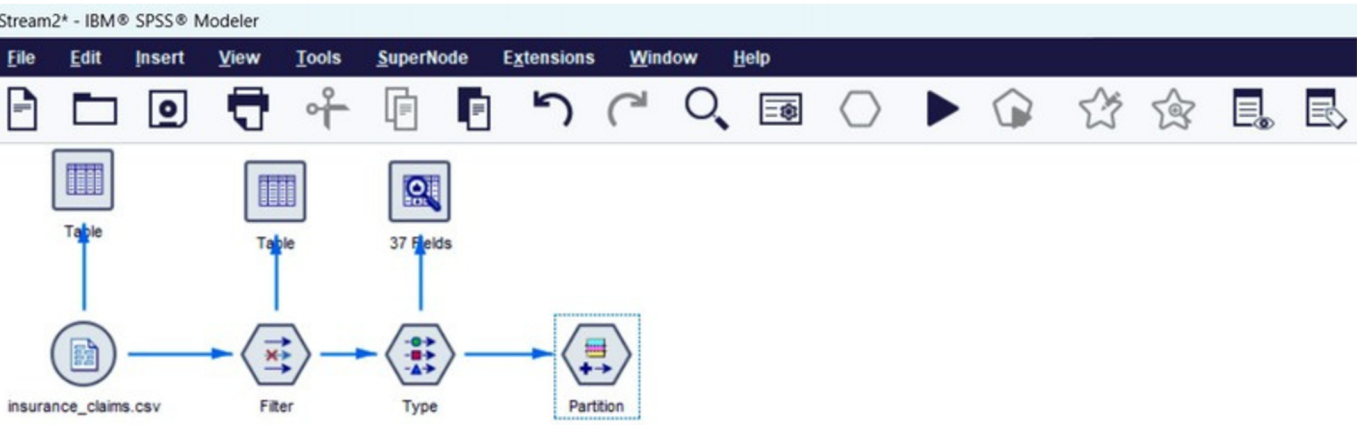
Step 4 : Partition Data

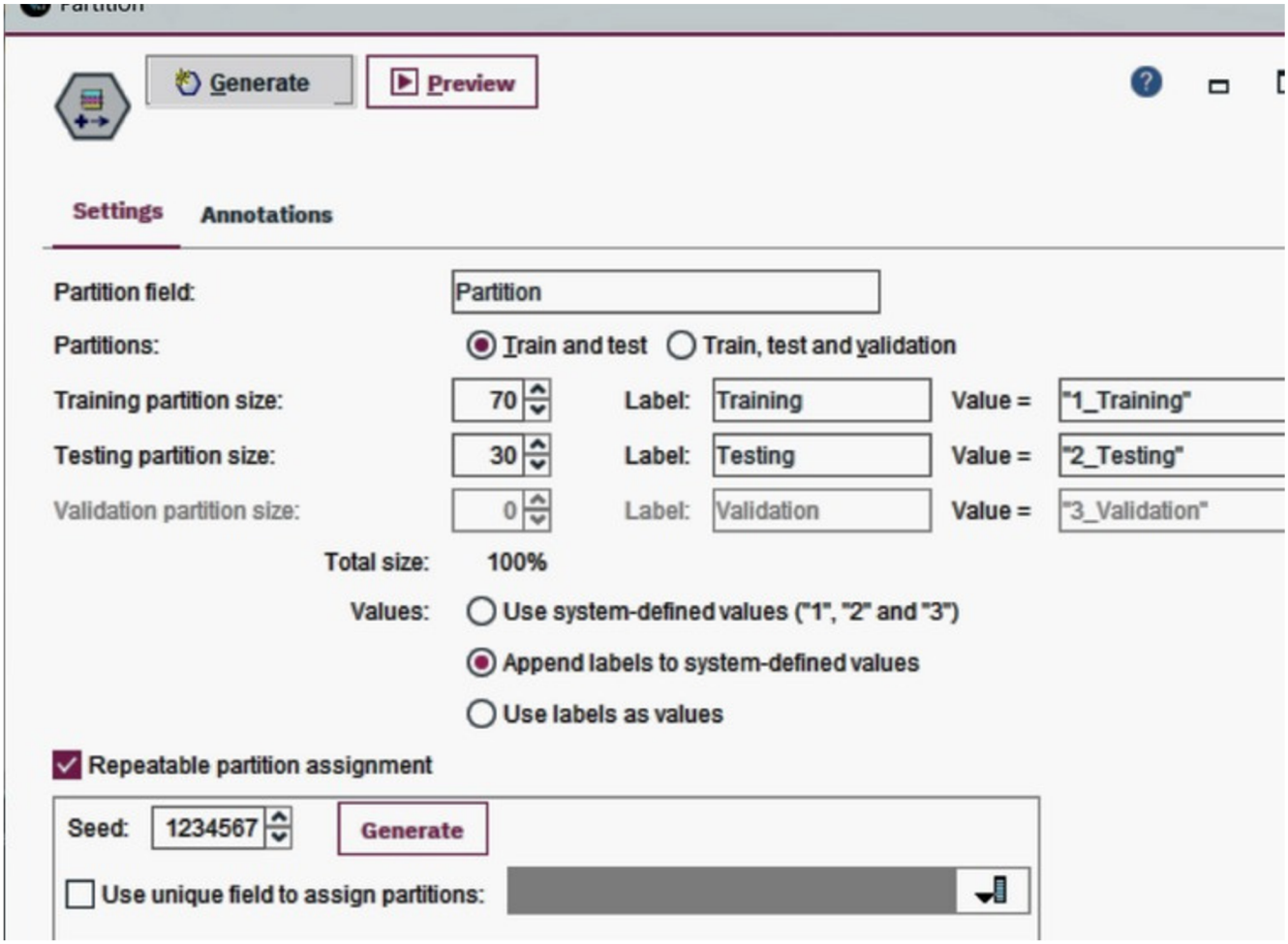
Added Partition Node to split data:

70% for Training

30% for Testing

This allows model evaluation on unseen data

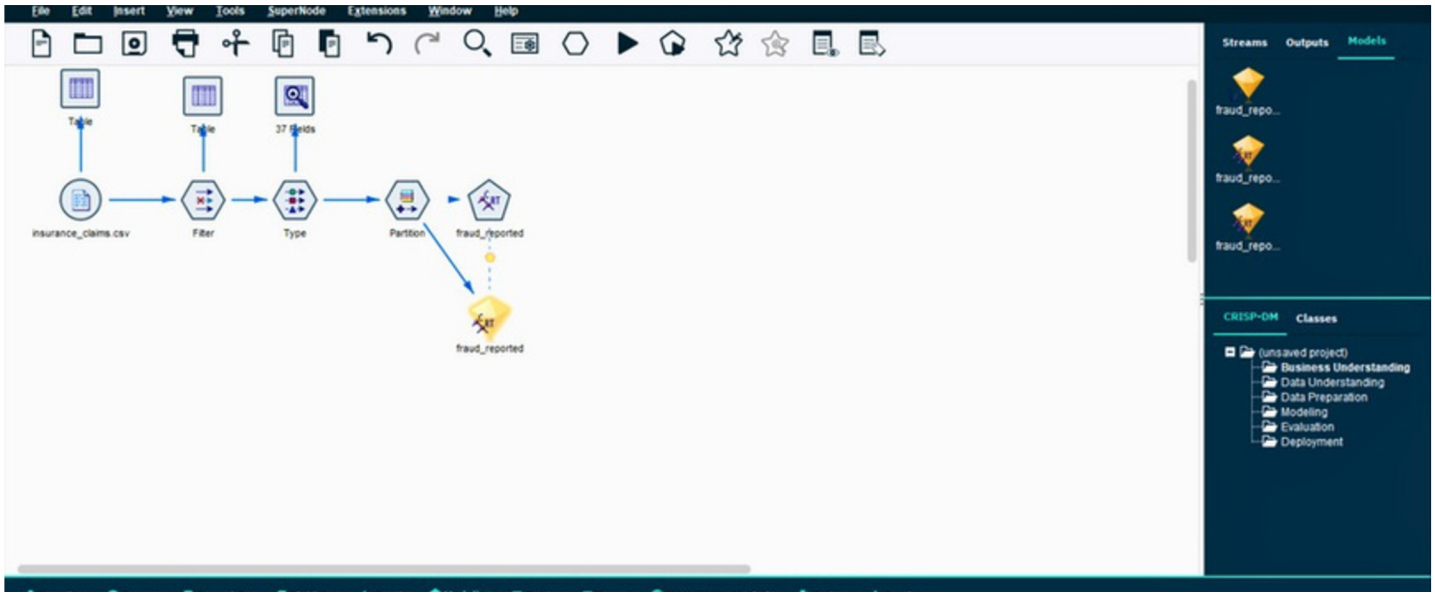




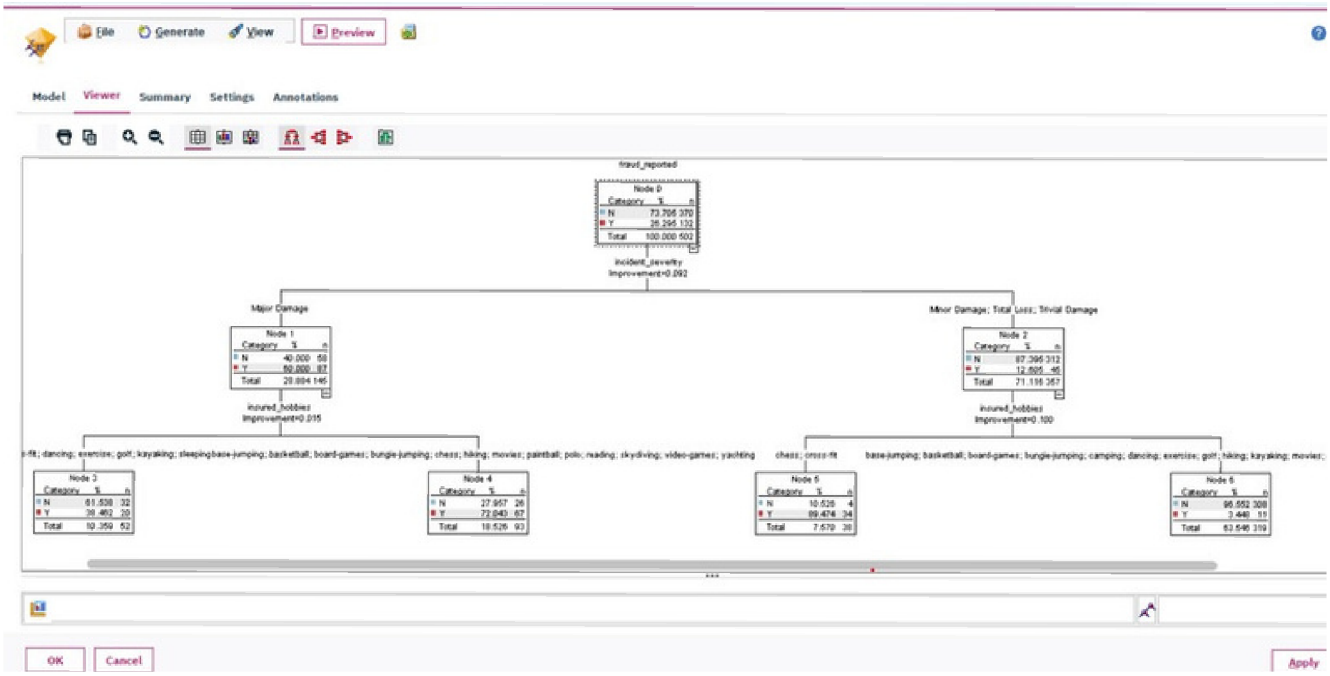
Step 5 : Build Models

Model 1 - C&R Tree

- From the Modeling palette, drag a C&R Tree Node.
- Connect it to the Partition Node.
- Open it → confirm:
- Target: fraud_reported
- Inputs: Auto-selected.
- Click Run → view the decision tree output (splits, accuracy, etc.).

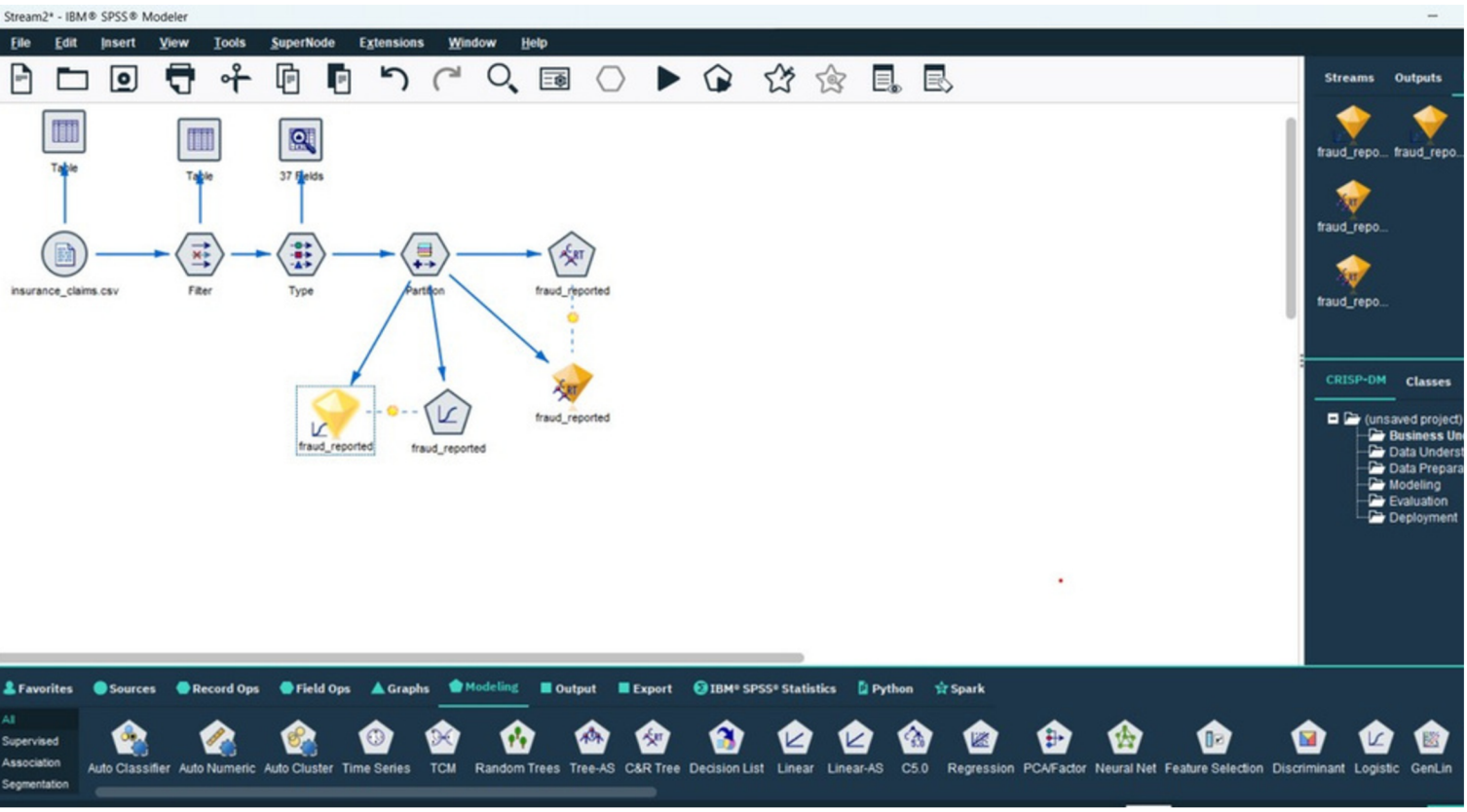


OUTPUT :



Model 2 :Logistic Regression

1.Drag a Logistic Regression Node from the Modeling palette.
2. Connect it to the same Partition Node.
3. Click Run → Check model summary and predictor significance.



Step 7: Generate Predictions

After training, connected the Model Nugget to the dataset and added a Table Node.

Output fields included:

- fraud_reported → Actual
- $L-fraud_reported → Predicted
- $LP-fraud_reported → Prediction probability

ır	fraud_reported	Partition	$L-fraud_reported	$LP-fraud_reported
)04	Y	1_Training	Y	0.800
)07	Y	1_Training	N	0.595
)07	N	1_Training	N	0.840
)14	Y	2_Testing	Y	0.998
)09	N	1_Training	N	1.000
)03	Y	1_Training	Y	0.922
)12	N	1_Training	N	0.998
)15	N	1_Training	N	0.996
)12	N	1_Training	N	1.000
996	N	1_Training	N	0.998
)02	N	1_Training	N	0.981
)06	N	1_Training	Y	0.910
)00	N	2_Testing	N	0.745
)10	N	1_Training	N	0.963
)03	Y	2_Testing	N	0.903
999	Y	2_Testing	N	0.986
)12	N	1_Training	N	0.542
)15	N	1_Training	N	0.905
)15	N	1_Training	N	0.961

TABLE

Table (42 fields, 1,000 records) #4

File Edit Generate

Table Annotations

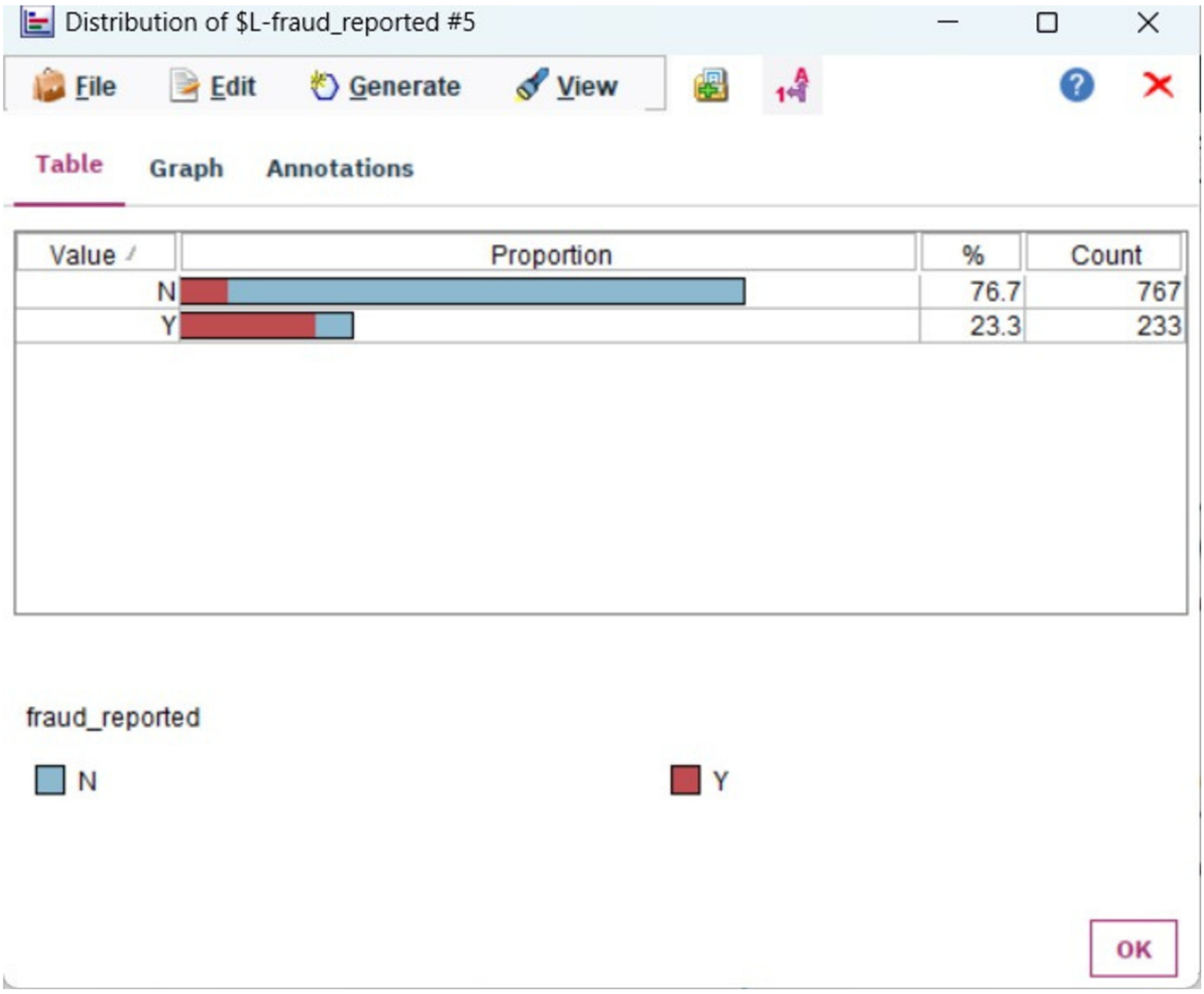
	bodily_injuries	witnesses	police_report_available	total_claim_amount	injury_claim	property_claim	vehicle_claim	auto_make	auto_model	auto_year	fraud_reported	Partition	$L-fraud_reported	$LP-fraud_reported
1	1	2	YES	71610	6510	13020	52080	Saab	92x	2004	Y	1_Training	Y	0.800
2	0	0	?	5070	780	780	3510	Mercedes	E400	2007	Y	1_Training	N	0.595
3	2	3	NO	34650	7700	3850	23100	Dodge	RAM	2007	N	1_Training	N	0.840
4	1	2	NO	63400	6340	6340	50720	Chevrolet	Tahoe	2014	Y	2_Testing	Y	0.998
5	0	1	NO	6500	1300	650	4550	Accura	RSX	2009	N	1_Training	N	1.000
6	0	2	NO	64100	6410	6410	51280	Saab	95	2003	Y	1_Training	Y	0.922
7	0	0	?	78650	21450	7150	50050	Nissan	Pathfinder	2012	N	1_Training	N	0.998
8	2	2	YES	51590	9380	9380	32830	Audi	A5	2015	N	1_Training	N	0.996
9	1	1	YES	27700	2770	2770	22160	Toyota	Camry	2012	N	1_Training	N	1.000
10	2	1	?	42300	4700	4700	32900	Saab	92x	1996	N	1_Training	N	0.999
11	2	2	?	87010	7910	15820	63280	Ford	F150	2002	N	1_Training	N	0.981
12	1	2	YES	114920	17680	17680	79560	Audi	A3	2006	N	1_Training	Y	0.910
13	1	0	NO	56520	4710	9420	42390	Saab	95	2000	N	2_Testing	N	0.745
14	1	1	NO	7280	1120	1120	5040	Toyota	Highlander	2010	N	1_Training	N	0.963
15	0	2	YES	46200	4200	8400	33600	Dodge	Neon	2003	Y	2_Testing	N	0.903
16	0	0	NO	63120	10520	10520	42080	Accura	MDX	1999	Y	2_Testing	N	0.986
17	1	2	YES	52110	5790	5790	40530	Nissan	Maxima	2012	N	1_Training	N	0.542
18	0	2	YES	77880	14160	7080	56640	Suburu	Legacy	2015	N	1_Training	N	0.905
19	1	0	NO	72930	6630	13260	53040	Accura	TL	2015	N	1_Training	N	0.961
20	2	0	NO	60400	6040	6040	48320	Nissan	Pathfinder	2014	N	1_Training	N	0.976
21	1	0	?	47160	0	5240	41920	Suburu	Impreza	2011	N	1_Training	N	0.981
22	1	2	?	37840	0	4730	33110	Accura	RSX	1996	N	2_Testing	N	1.000
23	0	0	YES	71520	17880	5960	47680	Suburu	Forrestor	2000	Y	1_Training	Y	0.785
24	2	2	?	98160	8180	16360	73620	Dodge	RAM	2011	Y	1_Training	Y	0.796
25	1	3	NO	77880	7080	14160	56640	Ford	Escape	2005	N	1_Training	N	0.927
26	1	3	YES	71500	16500	11000	44000	Ford	Escape	2006	Y	1_Training	N	0.592
27	1	3	YES	9020	1640	820	6560	Toyota	Camry	2005	N	2_Testing	N	0.926
28	2	1	?	5720	1040	520	4160	Suburu	Forrestor	2003	Y	2_Testing	Y	0.987
29	1	0	YES	69840	7760	15520	46560	Dodge	Neon	2009	N	1_Training	N	0.992
30	2	2	NO	91650	14100	14100	63450	Accura	TL	2011	N	1_Training	N	0.994
31	0	0	?	75600	12600	12600	50400	Toyota	Corolla	2005	N	1_Training	N	0.985
32	2	2	?	67140	7460	7460	52220	Ford	F150	2006	Y	1_Training	Y	0.708
33	2	3	NO	29790	3310	3310	23170	BMW	3 Series	2008	N	1_Training	N	0.997
34	1	2	?	77110	14020	14020	49070	Suburu	Impreza	2015	N	1_Training	Y	0.792
35	0	1	YES	64800	10800	5400	48600	Audi	A3	1999	N	2_Testing	N	0.994
36	2	0	YES	53100	10620	5310	37170	Mercedes	C300	1995	Y	2_Testing	Y	0.993
37	1	1	YES	60200	6020	6020	48160	Suburu	Forrestor	2004	Y	1_Training	N	0.994
38	1	1	YES	5330	1230	820	3280	Suburu	Legacy	2001	N	1_Training	N	1.000
39	2	0	?	62300	12460	6230	43610	Jeep	Wrangler	2007	N	2_Testing	N	0.536

Step 8: Visualize Results

- After model training (C&R Tree and Logistic Regression), I connected the Distribution Graph Node to visualize the target field fraud_reported.
- The field fraud_reported was selected as the target variable for visualization.
- The output shows a clear bar chart distribution of “Yes” (fraud) and “No” (non-fraud) claims.

- **No (N): 76.7% (767 records)**
- **Yes (Y): 23.3% (233 records)**

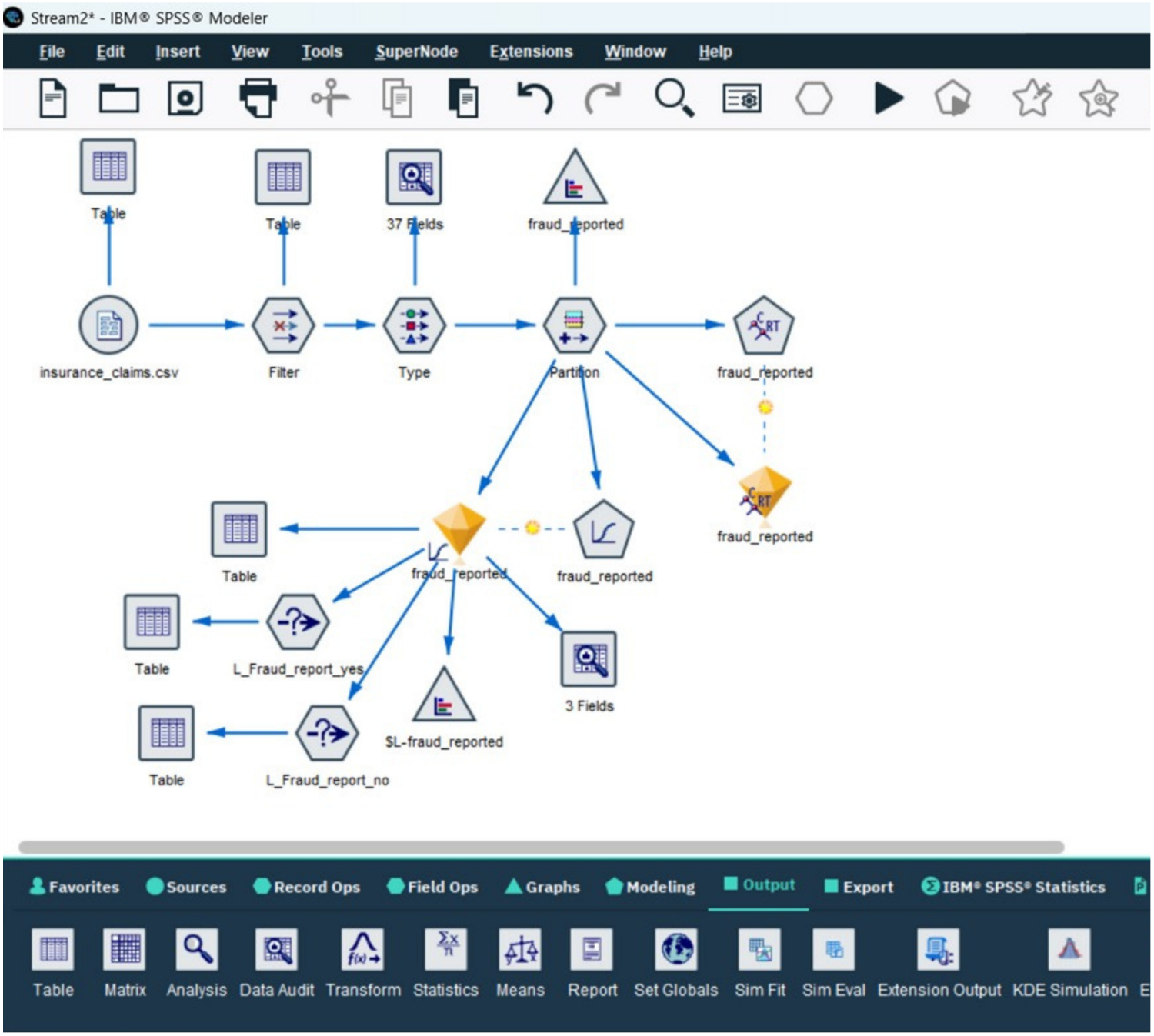
This indicates that most insurance claims are non-fraudulent, but about one-fourth of cases (23.3%) are identified as fraud, which is still a significant number for analysis.



Insights :

- *Major Damage and High Claim Amounts are the most common indicators of fraud.*
- *Customers with risky hobbies (like skydiving or motor racing) are more prone to fraudulent claims.*
- *Fraud detection is better when combining C&R Tree visualization and Logistic Regression probability scoring.*

FINAL VIEW



Model Comparison

Model Type	Description	Performance	Best Use
C&R Tree	Decision tree showing split by features	High accuracy	High accuracy Visual, interpretative classification
Logistic Regression	Statistical model estimating fraud probability	High accuracy	Numerical fraud probability prediction

Conclusion :

The project successfully built and evaluated two models to detect insurance fraud using IBM SPSS Modeler.
The models help the insurance company:

- *Identify potential fraudulent claims early.*
- *Save costs by reducing false claims.*
- *Improve the reliability of claim verification systems.*

Overall, the C&R Tree and Logistic Regression models provided valuable insights into fraudulent behavior patterns.