

Name: Snigdha Singh

Data Analytics

VIT Vellore, Tamil Nadu

Assignment-3 Dataset: diabetes.csv

New project

Define details

Name

Description

Storage

Cloud Object Storage-hv

Choose project options

☐ Restrict who can be a collaborator ⓘ

Project includes integration with [Cloud Object Storage](#) for storing project assets.

Cancel

Create

Fig: Create a new Project

Projects / Classification of Diabetes

Overview


Assets

Jobs

Manage

Assets

No assets created with tools yet.



[View all](#)

Resource usage

For this month in this project

0 CUH

Readme

Type project notes, reminders, or instructions

Project history

No notifications

You will see your most recent notifications here.

Data in this project

Drop data files here or browse for files to upload

diabetes.csv

Uploaded!

[Dismiss](#)

Fig: upload the dataset to the project

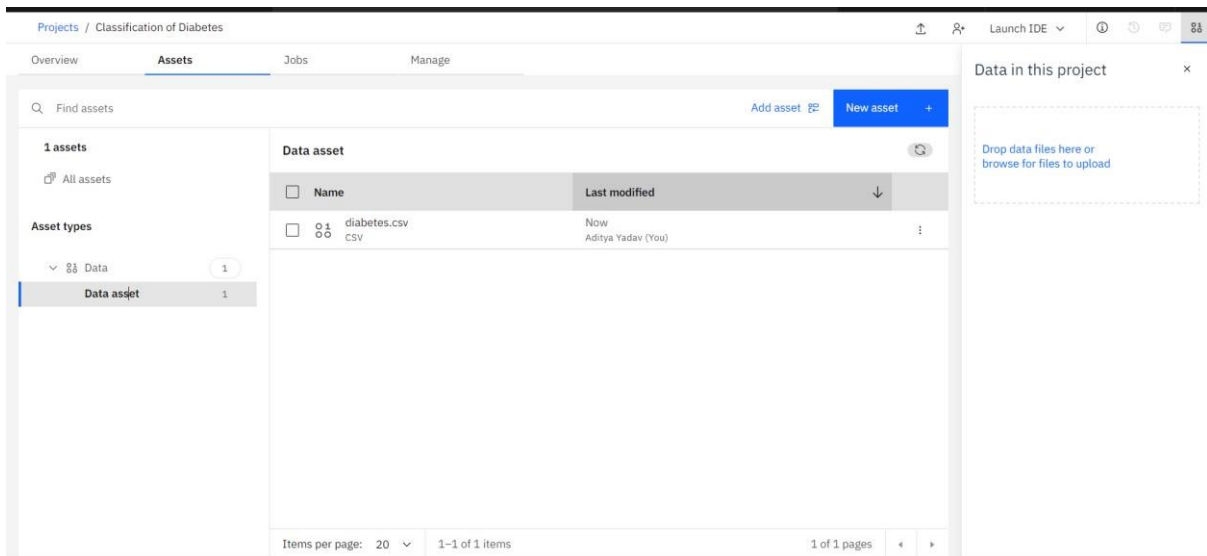


Fig: data asset available

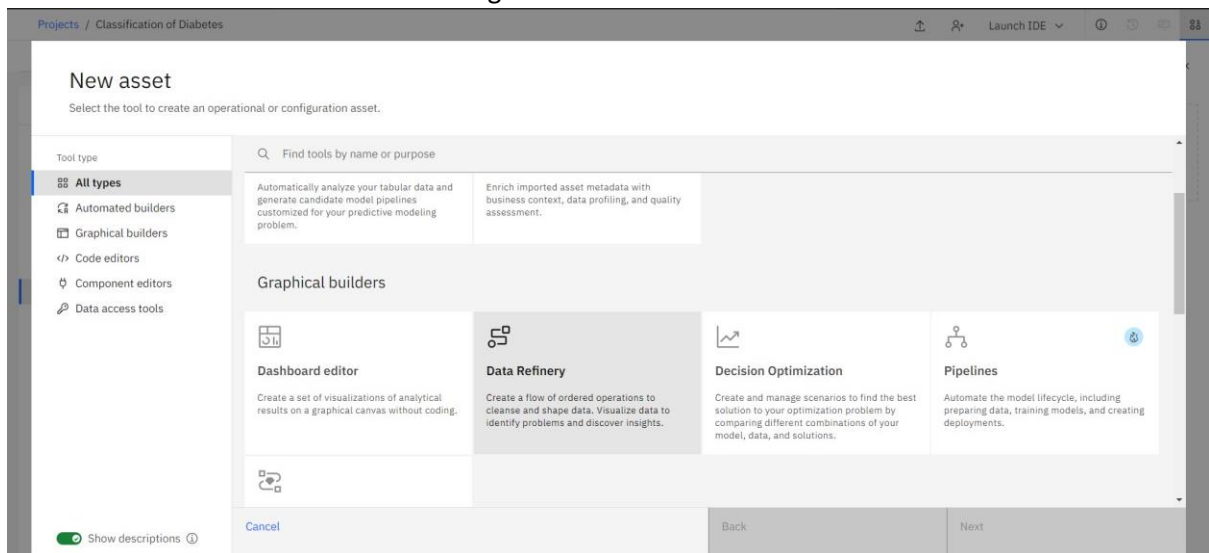


Fig: create data refinery

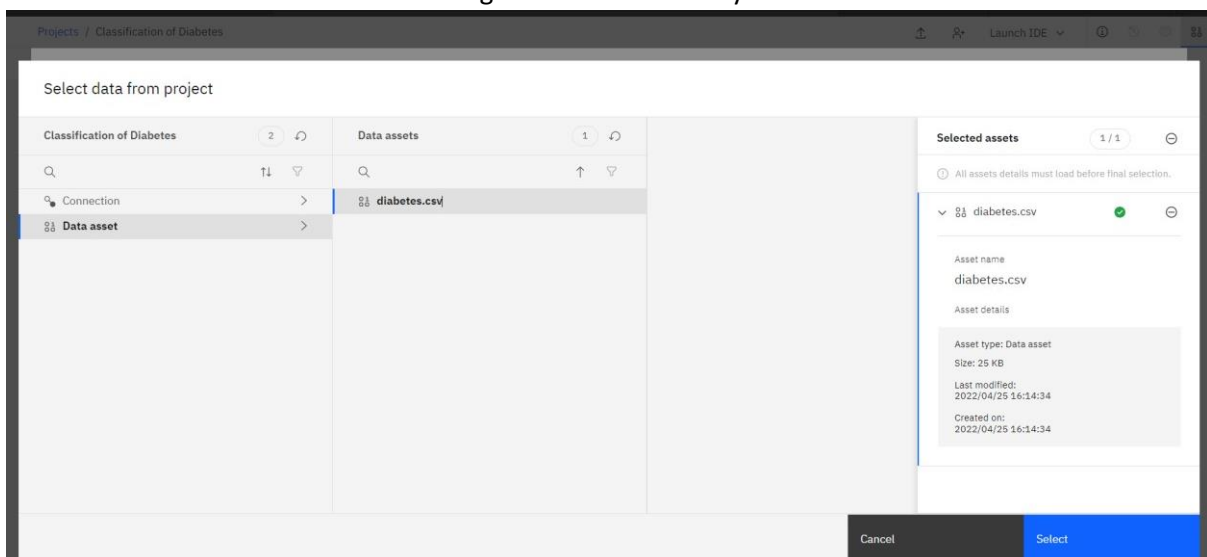


Fig: add the dataset for data refining

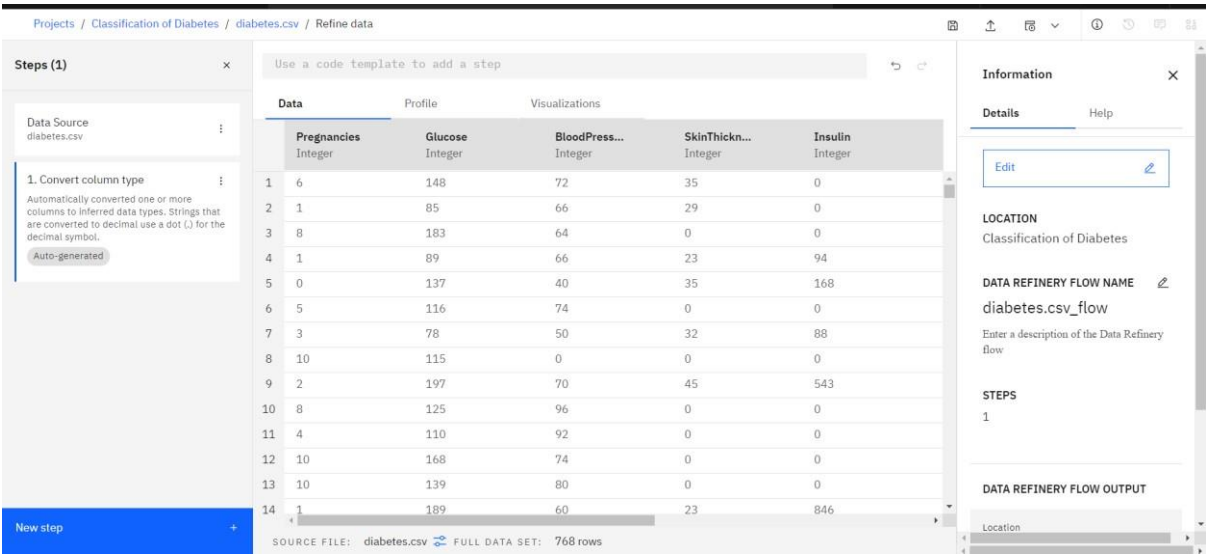


Fig: descriptive details of the dataset

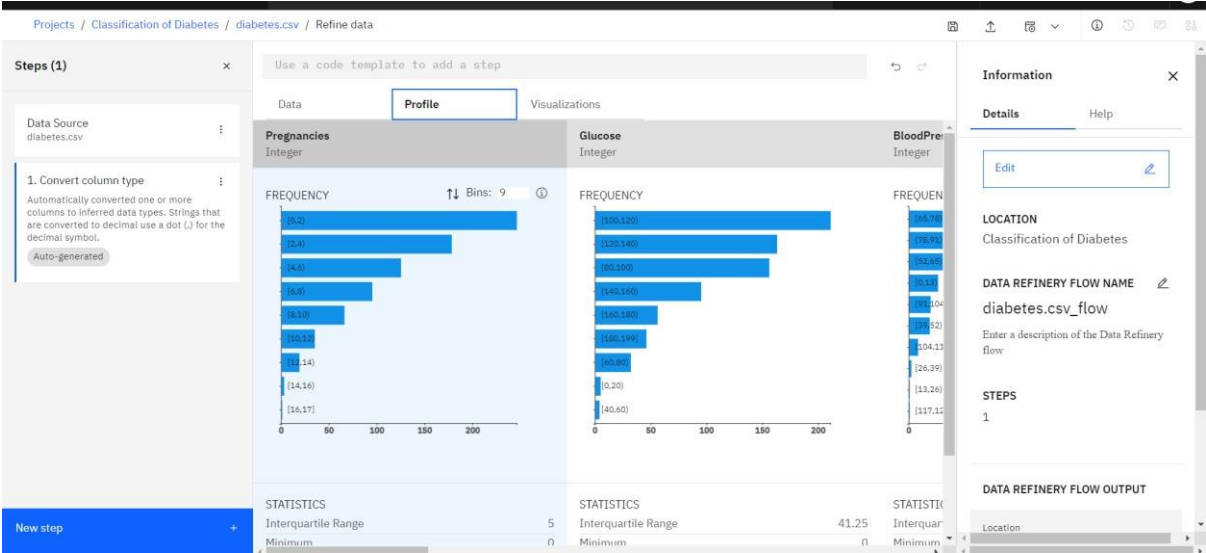
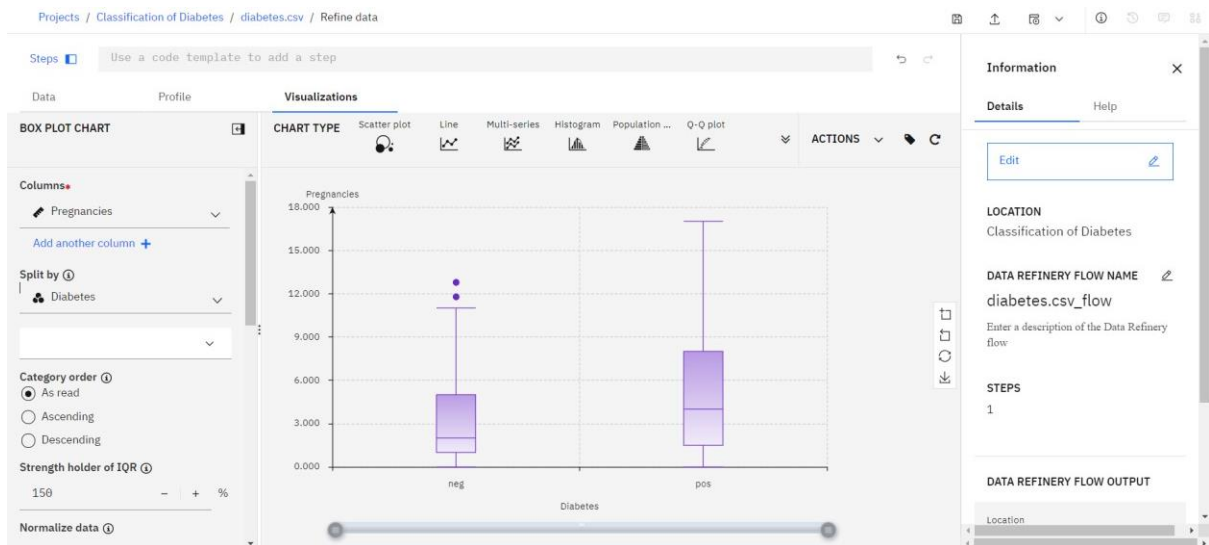


Fig: profile of our data



Visualize on the basis of box plot chart on pregnancies and diabetes.

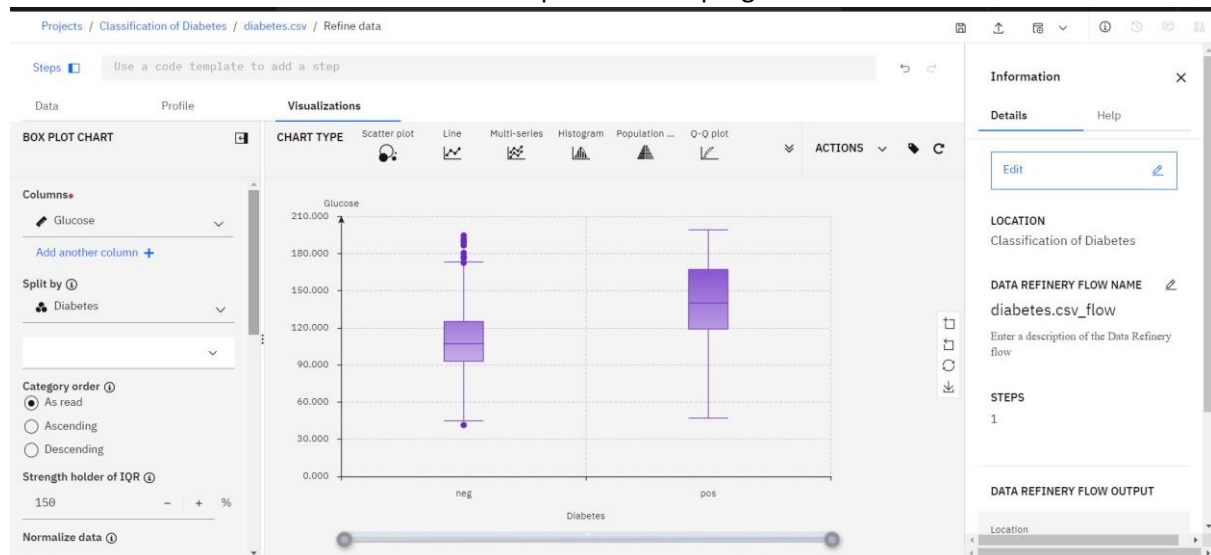


Fig: Visualize on the basis of box plot chart on glucose and diabetes.

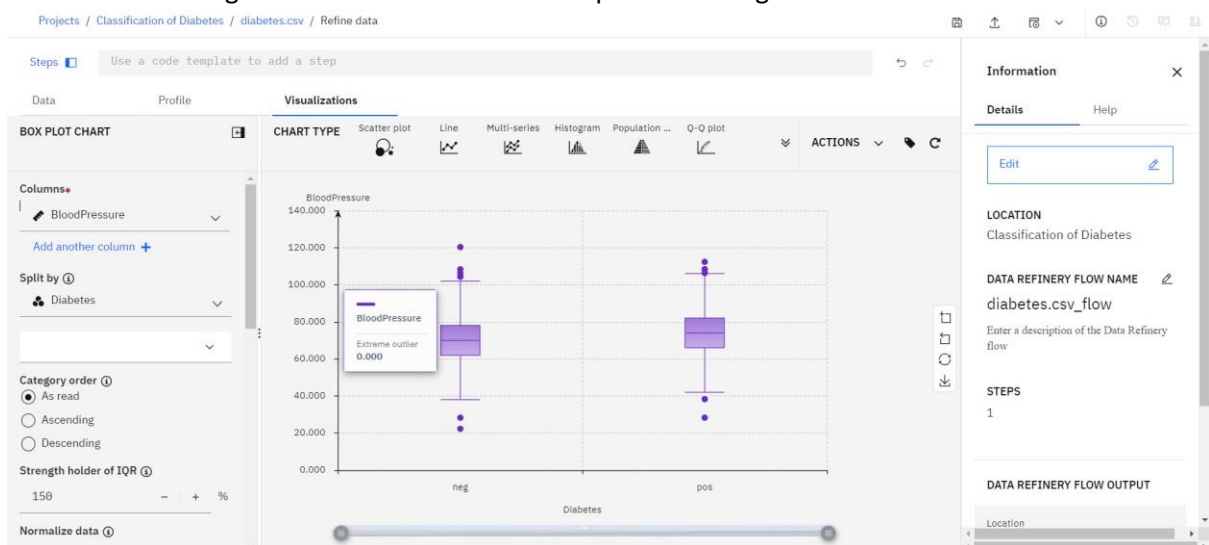


Fig: graph for BloodPressure and diabetes

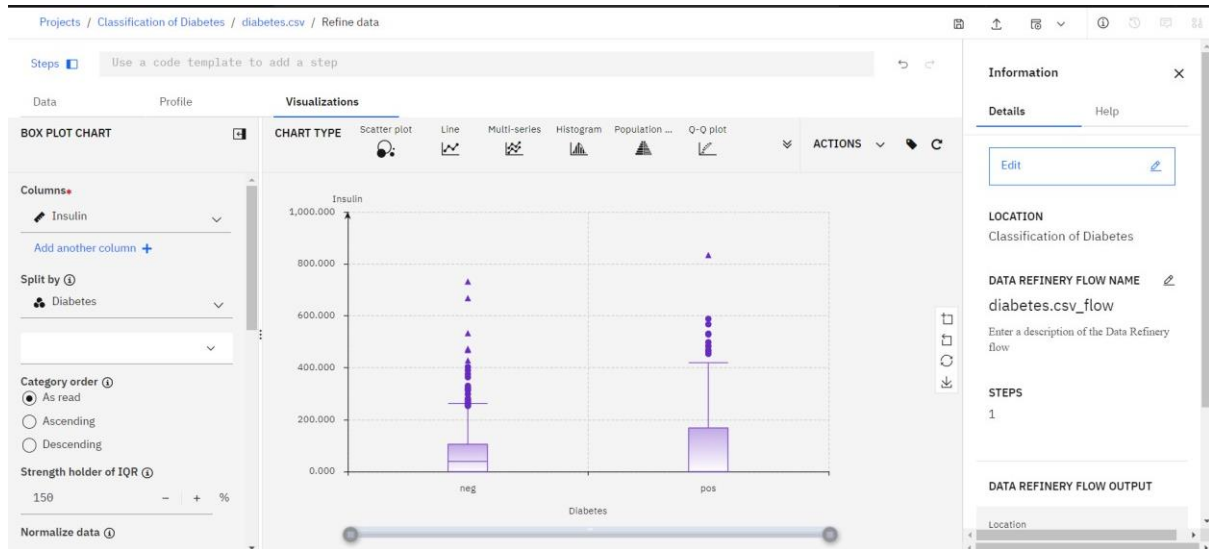


Fig: graph for Insulin and diabetes

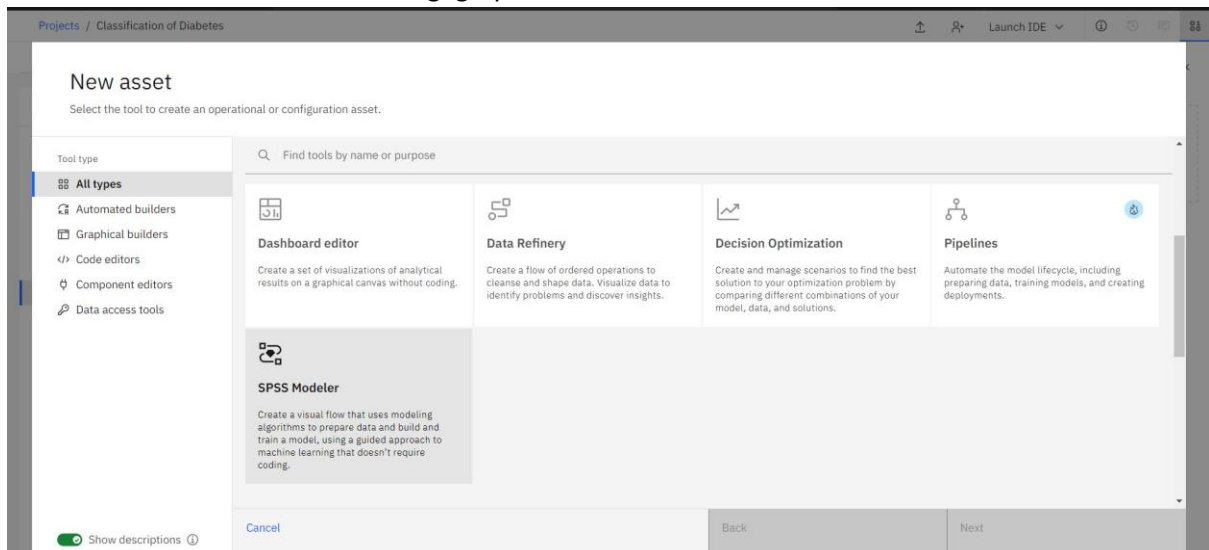


Fig: create a SPSS modeler

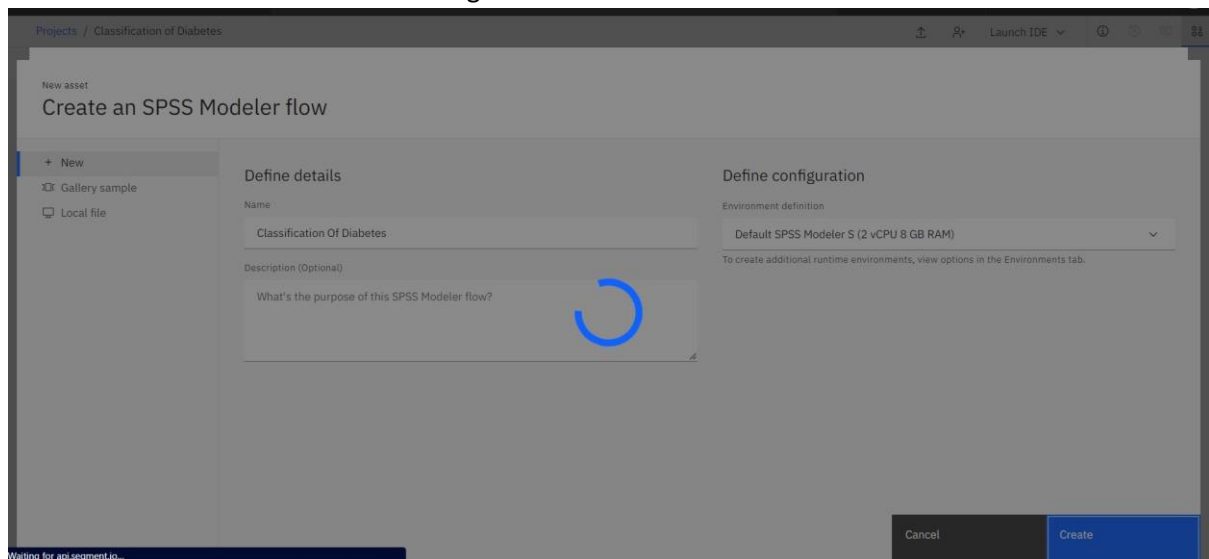


Fig: creating workbench

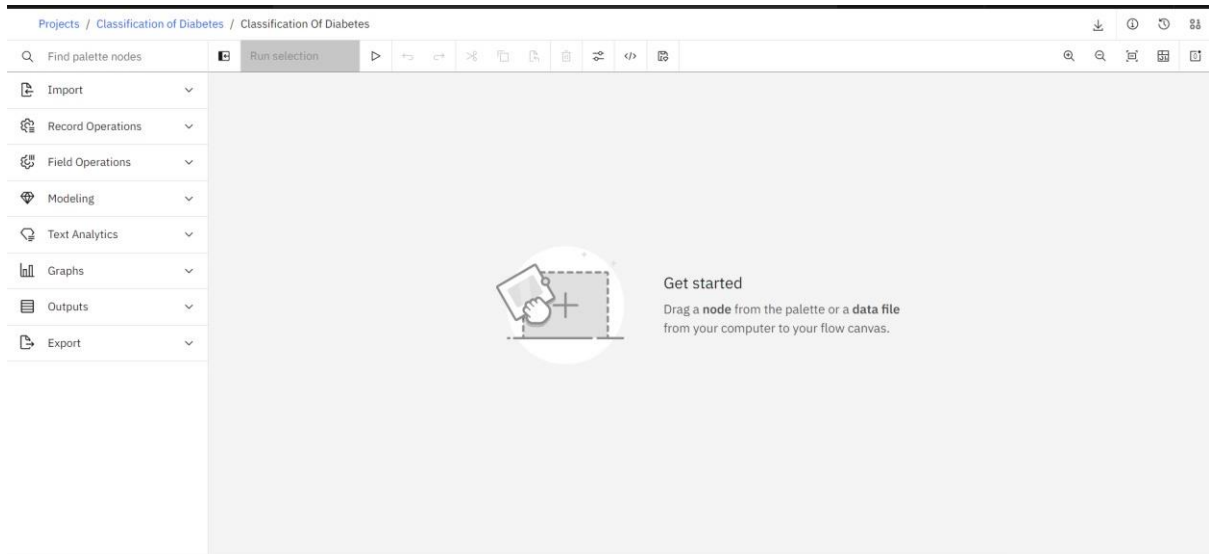
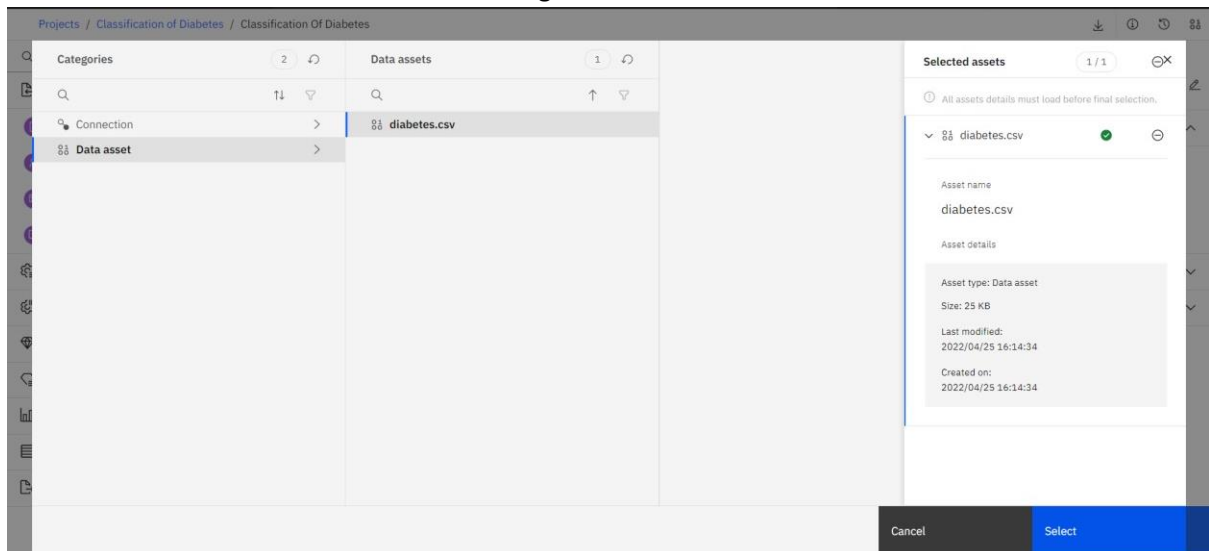
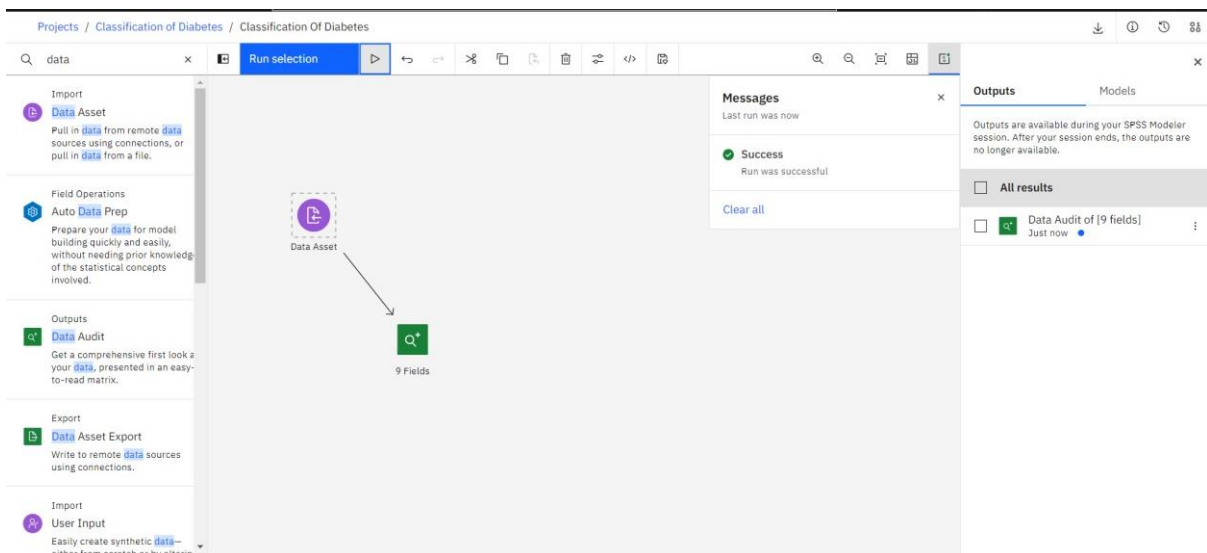


Fig: workbench



add dataset node to the workplace



Add data audit node and run the model.

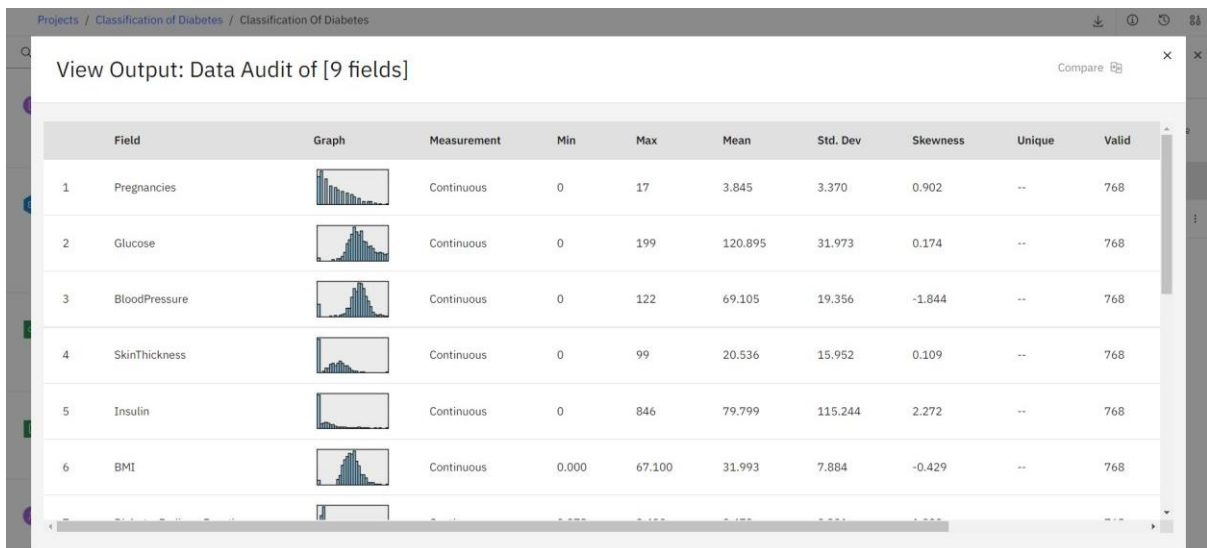


Fig: output of audit

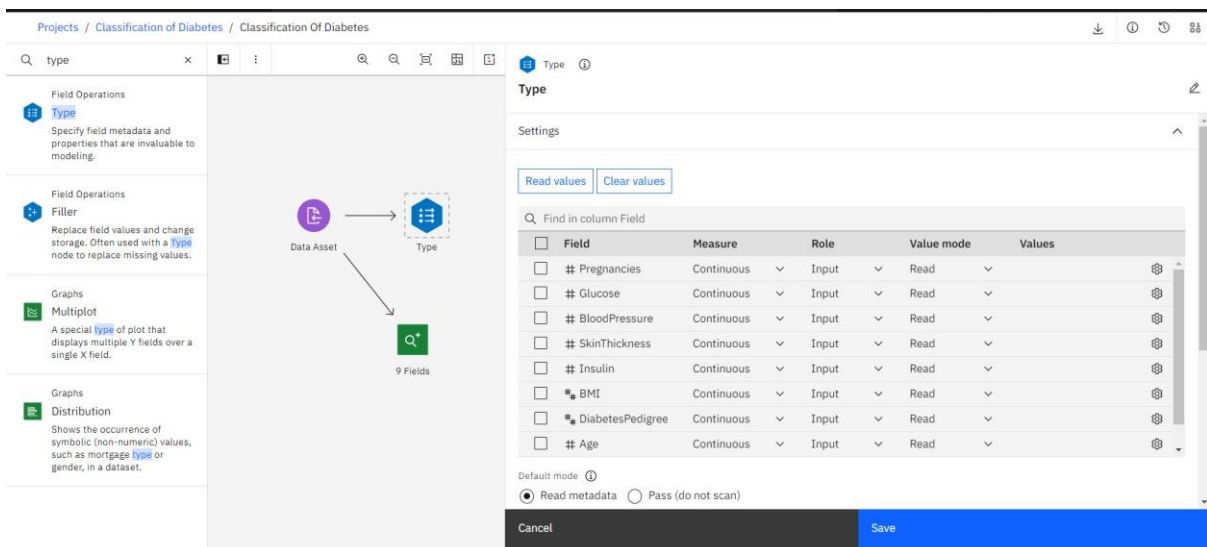
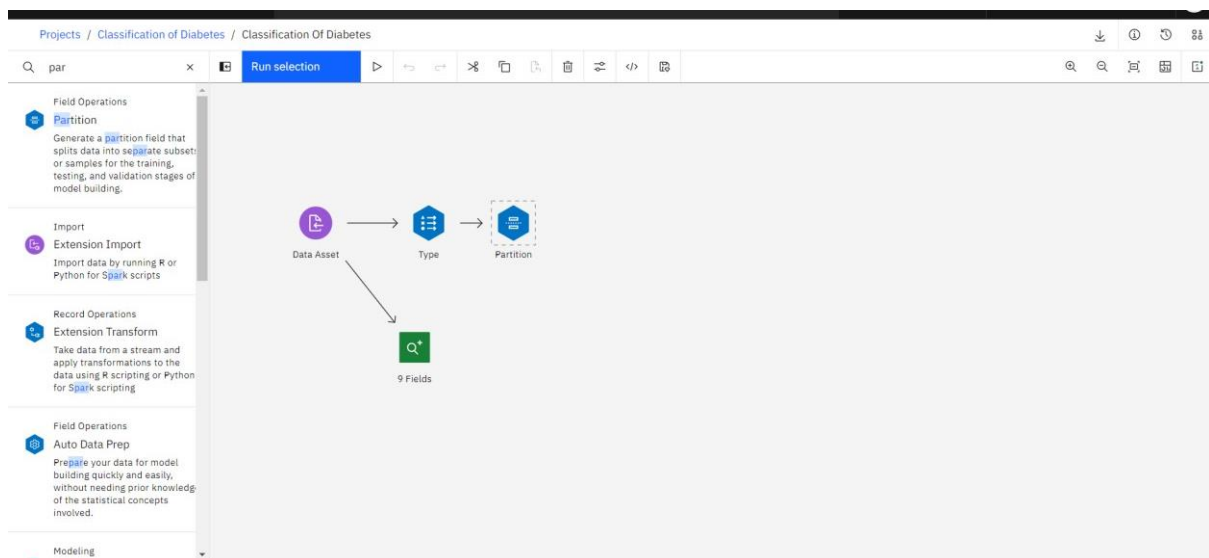


Fig: drag type node and update with the target and input data.

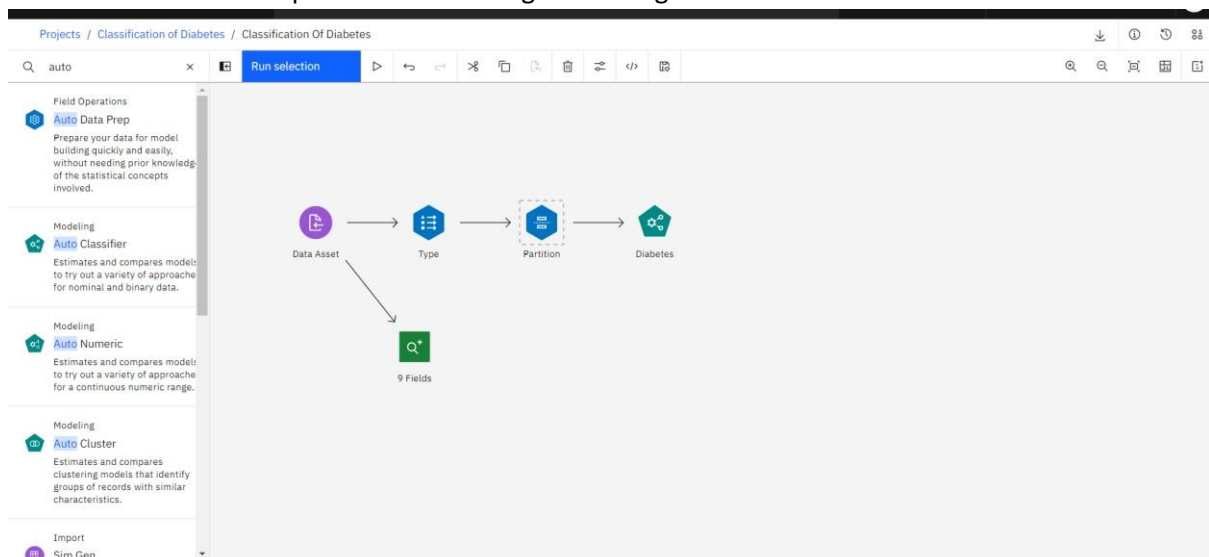
View Output: 9 Fields

	Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty S
1	Pregnancies	Continuous	0	768	None	Never	Fixed	100.000	768	0	0
2	Glucose	Continuous	0	768	None	Never	Fixed	100.000	768	0	0
3	BloodPressure	Continuous	0	768	None	Never	Fixed	100.000	768	0	0
4	SkinThickness	Continuous	0	768	None	Never	Fixed	100.000	768	0	0
5	Insulin	Continuous	0	768	None	Never	Fixed	100.000	768	0	0
6	BMI	Continuous	0	768	None	Never	Fixed	100.000	768	0	0
7	DiabetesPedigreeFunction	Continuous	0	768	None	Never	Fixed	100.000	768	0	0
8	Age	Continuous	0	768	None	Never	Fixed	100.000	768	0	0
9	Diabetes	Categorical	--	--	--	Never	Fixed	100.000	768	0	0

Updated outliers



Add partition node and give training and test as 90% and 10%



Select auto classifier and run the model

The screenshot shows the SPSS Modeler interface with the project 'Classification Of Diabetes'. The workflow is as follows:

- Data Asset** (purple icon) connects to **Type** (blue icon) and **9 Fields** (green icon).
- Type** connects to **Partition** (blue icon).
- Partition** connects to **Diabetes** (green icon).
- Diabetes** connects to **Diabetes** (yellow icon).

The left sidebar shows the following options:

- Field Operations: **Auto Data Prep**
- Modeling: **Auto Classifier**
- Modeling: **Auto Numeric**
- Modeling: **Auto Cluster**
- Import: **Sim Gen**

The right sidebar shows the **Outputs** tab with the following results:

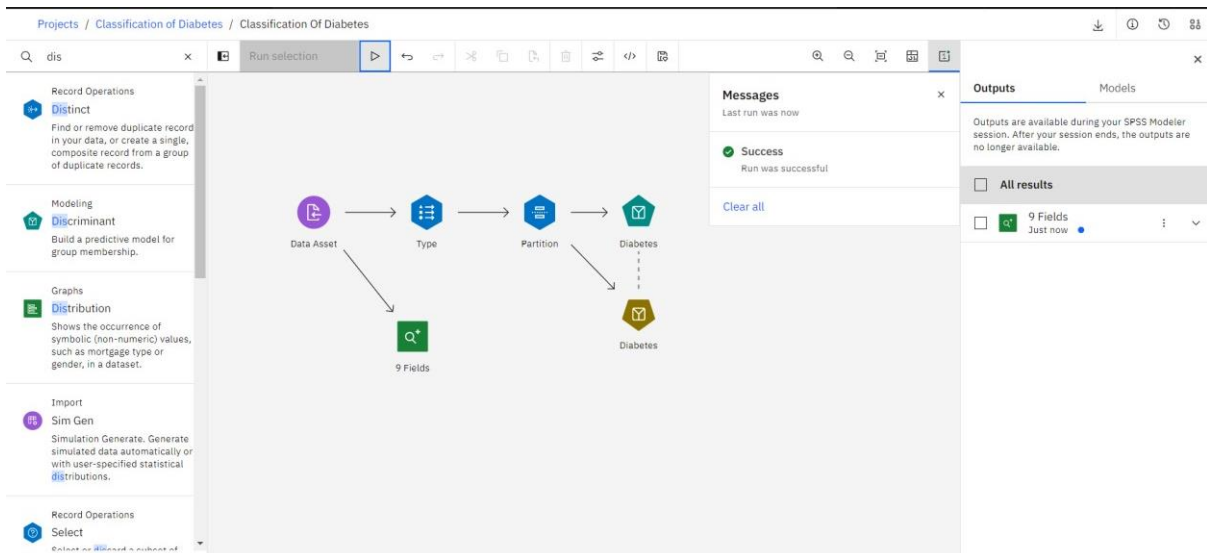
- All results**
- 9 Fields** (6 minutes ago)

Run the auto classifier model

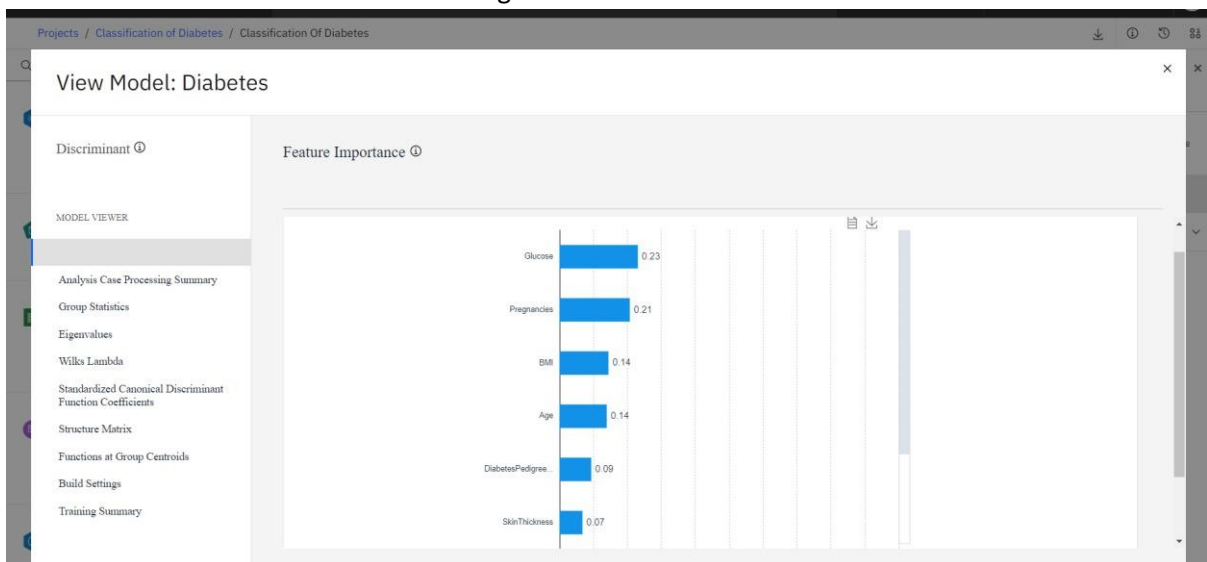
The 'View Model: Diabetes' dialog box shows the 'Auto Classifier - Models' table. The target is 'DIABETES'.

USE	MODEL NAME	ESTIMATOR	BUILD TIME (MINS)	NO. FIELDS USED	ACCURACY	ACCUMULATED ACCURACY	AREA UNDER CURVE	ACCUMULATED AUC	RECALL	PRECISION
<input checked="" type="checkbox"/>	Logistic regression 1	Nominal Regression	< 1	8	68.293	68.293	0.779	0.779	0.412	0.700
<input checked="" type="checkbox"/>	Discriminant 1	Discriminant	< 1	8	73.171	73.171	0.776	0.776	0.588	0.714
<input checked="" type="checkbox"/>	Tree-AS 1	CHAID	< 1	4	73.171	73.171	0.766	0.766	0.471	0.800
<input checked="" type="checkbox"/>	CHAID 1	CHAID	< 1	4	73.171	73.171	0.749	0.749	0.529	0.750
<input checked="" type="checkbox"/>	C5.1	C5.0	< 1	7	69.512	69.512	0.712	0.712	0.529	0.667

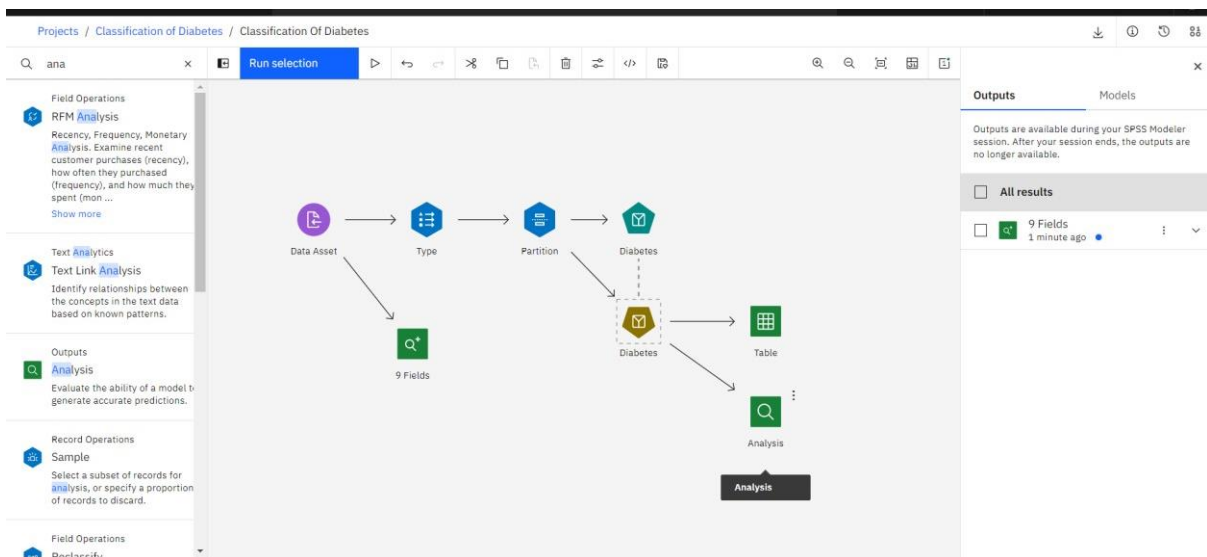
examine the model by view model



Perform again for discriminant model



View model



Create output in form of table and analysis form

Projects / Classification of Diabetes / Classification Of Diabetes

View Output: Table (12 fields, 768 records) [Compare](#)

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Diabetes	Partition	\$D-Diabetes	\$DP-Diabetes
1	6	148	72	35	0	33.600	0.627	50	pos	1_Training	pos	0.842
2	1	85	66	29	0	26.600	0.351	31	neg	1_Training	neg	0.932
3	8	183	64	0	0	23.300	0.672	32	pos	1_Training	pos	0.908
4	1	89	66	23	94	28.100	0.167	21	neg	2_Testing	neg	0.943
5	0	137	40	35	168	43.100	2.288	33	pos	1_Training	pos	0.938
6	5	116	74	0	0	25.600	0.201	30	neg	1_Training	neg	0.777
7	3	78	50	32	88	31.000	0.248	26	pos	1_Training	neg	0.908
8	10	115	0	0	0	35.300	0.134	29	neg	1_Training	pos	0.775
9	2	197	70	45	543	30.500	0.158	53	pos	1_Training	pos	0.873
10	8	125	96	0	0	0.000	0.232	54	pos	1_Training	neg	0.911
11	4	110	92	0	0	37.600	0.191	30	neg	1_Training	neg	0.726
12	10	168	74	0	0	38.000	0.537	34	pos	1_Training	pos	0.947
13	10	139	80	0	0	27.100	1.441	57	neg	2_Testing	pos	0.889
14	1	189	60	23	846	30.100	0.398	59	pos	1_Training	pos	0.830
15	5	166	72	19	175	25.800	0.587	51	pos	1_Training	pos	0.803

Table

Projects / Classification of Diabetes / Classification Of Diabetes

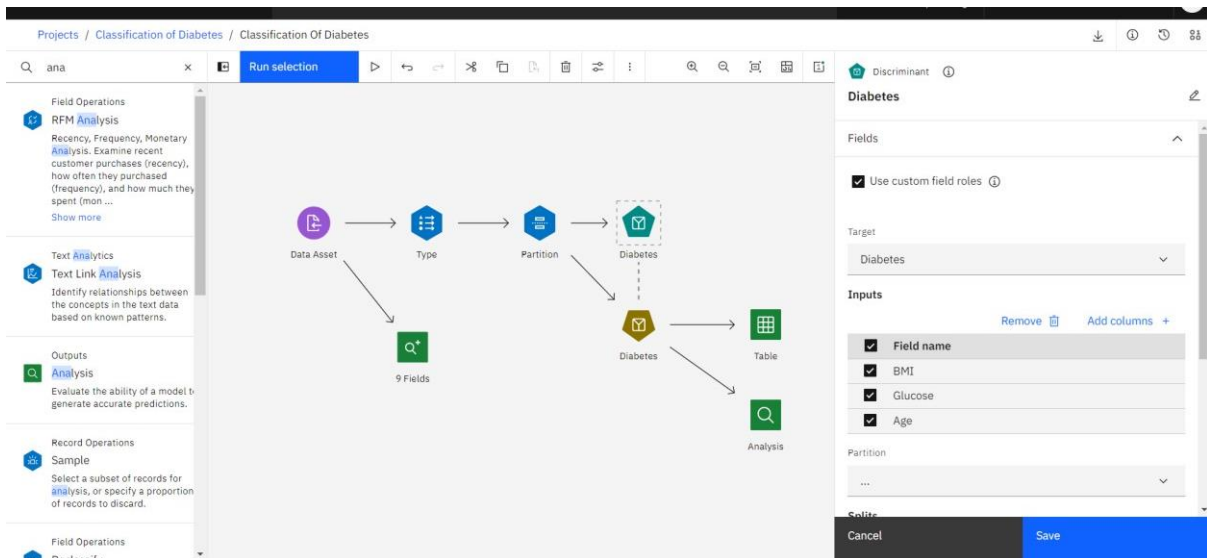
View Output: Analysis of [Diabetes] [Compare](#) [Collapse All](#)

Results for output field Diabetes

Comparing \$D-Diabetes with Diabetes

'Partition'	1_Training		2_Testing	
Correct	528	76.97%	60	73.17%
Wrong	158	23.03%	22	26.83%
Total	686		82	

Analysis



Use custom fields

Projects / Classification of Diabetes / Classification Of Diabetes

View Output: Table (12 fields, 768 records) #1

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Diabetes	Partition	\$D-Diabetes	\$DP-Diabetes
1	6	148	72	35	0	33.600	0.627	50	pos	1_Training	pos	0.793
2	1	85	66	29	0	26.600	0.351	31	neg	1_Training	neg	0.877
3	8	183	64	0	0	23.300	0.672	32	pos	1_Training	pos	0.775
4	1	89	66	23	94	28.100	0.167	21	neg	2_Testing	neg	0.884
5	0	137	40	35	168	43.100	2.288	33	pos	1_Training	pos	0.755
6	5	116	74	0	0	25.600	0.201	30	neg	1_Training	neg	0.729
7	3	78	50	32	88	31.000	0.248	26	pos	1_Training	neg	0.885
8	10	115	0	0	0	35.300	0.134	29	neg	1_Training	neg	0.585
9	2	197	70	45	543	30.500	0.158	53	pos	1_Training	pos	0.949
10	8	125	96	0	0	0.000	0.232	54	pos	1_Training	neg	0.859
11	4	110	92	0	0	37.600	0.191	30	neg	1_Training	neg	0.578
12	10	168	74	0	0	38.000	0.537	34	pos	1_Training	pos	0.865
13	10	139	80	0	0	27.100	1.441	57	neg	2_Testing	pos	0.684
14	1	189	60	23	846	30.100	0.398	59	pos	1_Training	pos	0.943
15	5	166	72	19	175	25.800	0.587	51	pos	1_Training	pos	0.807

After custom input

Projects / Classification of Diabetes / Classification Of Diabetes

View Output: Analysis of [Diabetes] #1

Results for output field Diabetes

Comparing \$D-Diabetes with Diabetes

'Partition'	1_Training	2_Testing
Correct	515 75.07%	59 71.95%
Wrong	171 24.93%	23 28.05%
Total	686	82

After custom input

Dataset: Insurance.csv

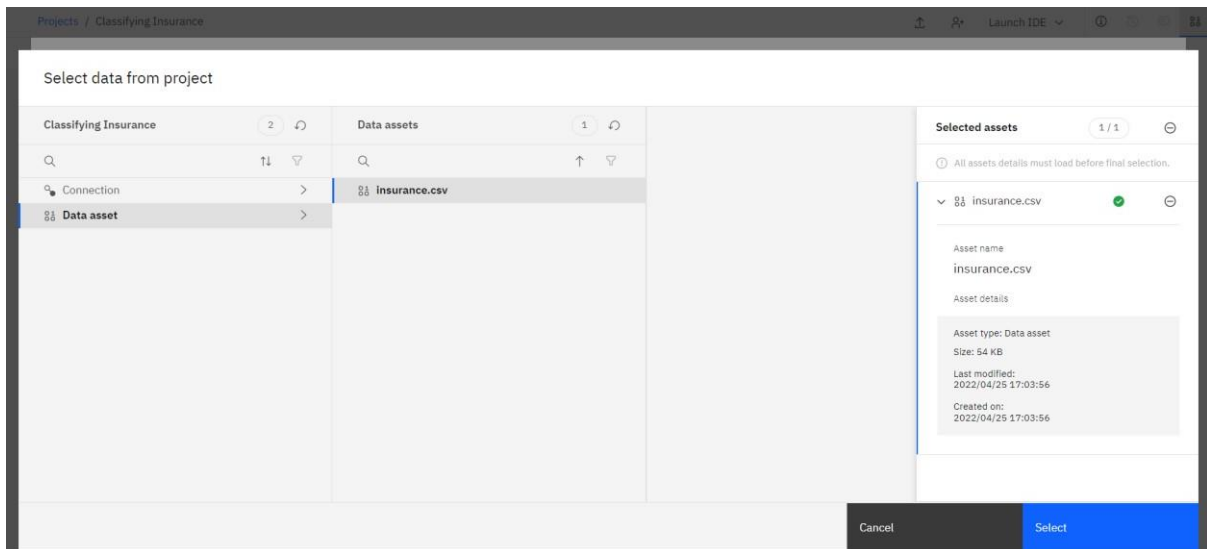
The screenshot shows the 'Overview' tab of a project named 'Classifying Insurance'. The interface includes a top navigation bar with 'Projects / Classifying Insurance', a 'Launch IDE' button, and several utility icons. Below the navigation bar, there are four main sections: 'Assets' (with a 'View all' link), 'Resource usage' (showing '0 CUH' for the month), 'Readme' (with a text area for project notes), and 'Project history' (showing 'No notifications').

Create a new project

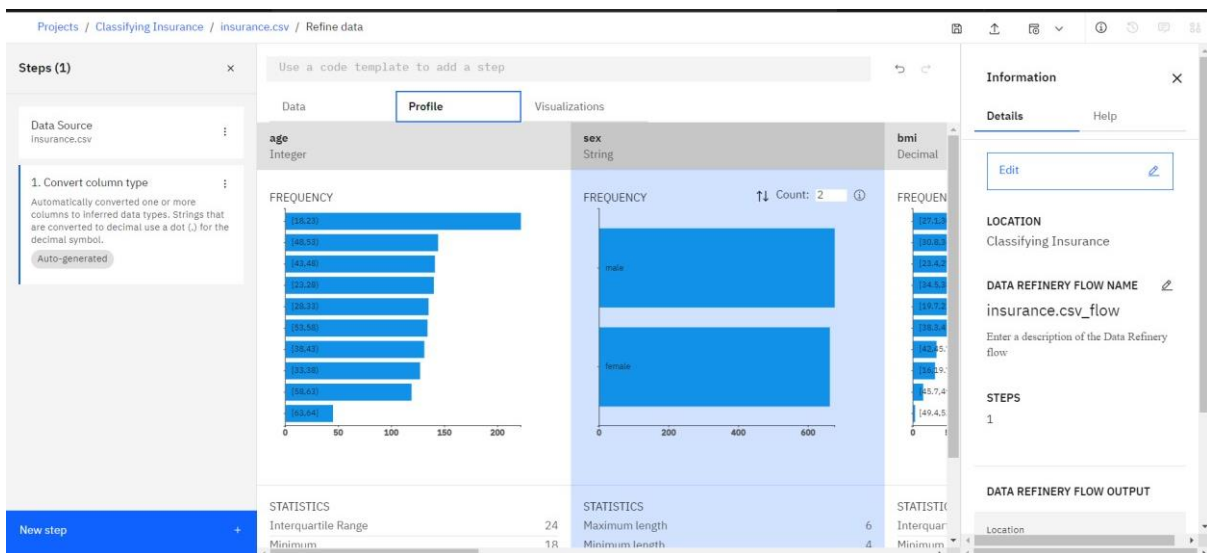
The screenshot shows the 'Assets' tab of the 'Classifying Insurance' project. It features a search bar, a list of assets (currently showing 1 asset), and a table of data assets. The table has columns for 'Name' and 'Last modified'. A single asset, 'insurance.csv', is listed with a CSV icon and a modification time of 'Now' by 'Aditya Yadav (You)'. On the right, a 'Data in this project' panel shows a dashed box for dropping files. At the bottom, there are pagination controls showing '1 of 1 items'.

Name	Last modified
insurance.csv CSV	Now Aditya Yadav (You)

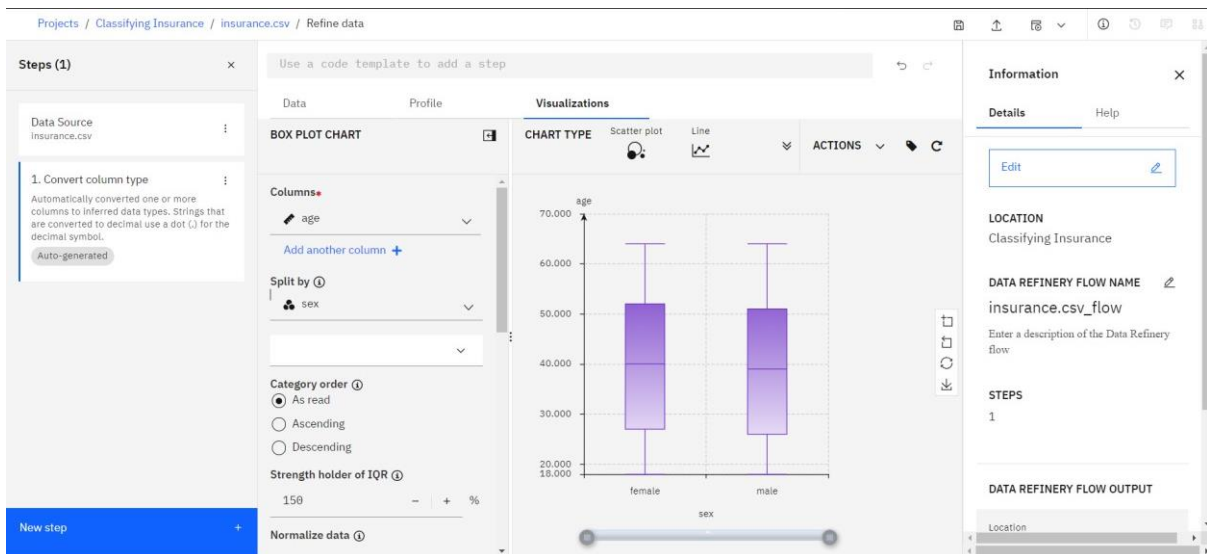
Upload dataset



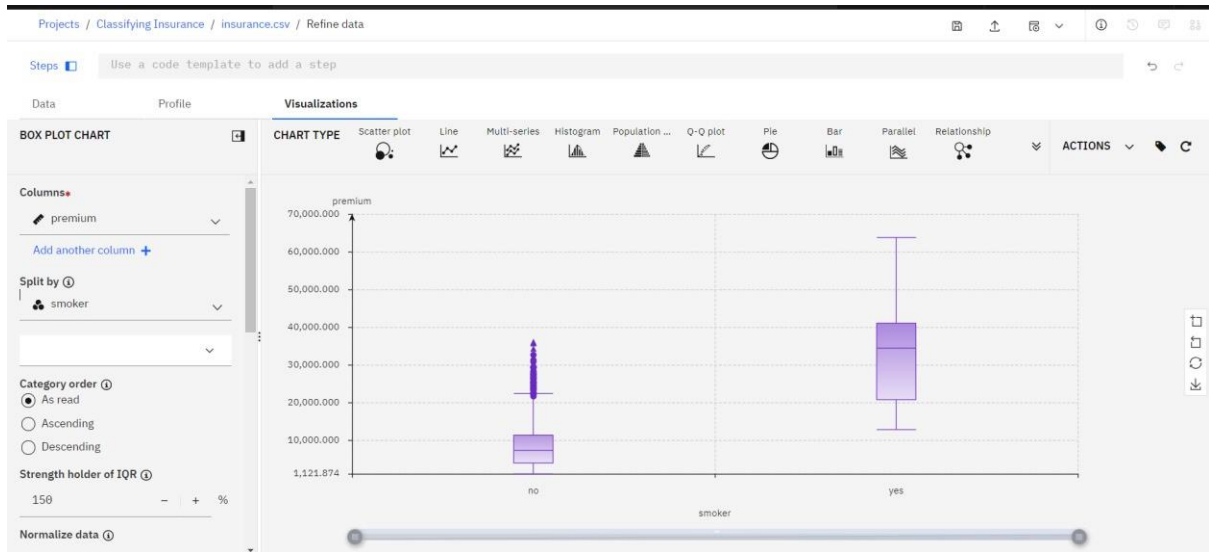
Go for data refining for the given dataset



Profile for dataset



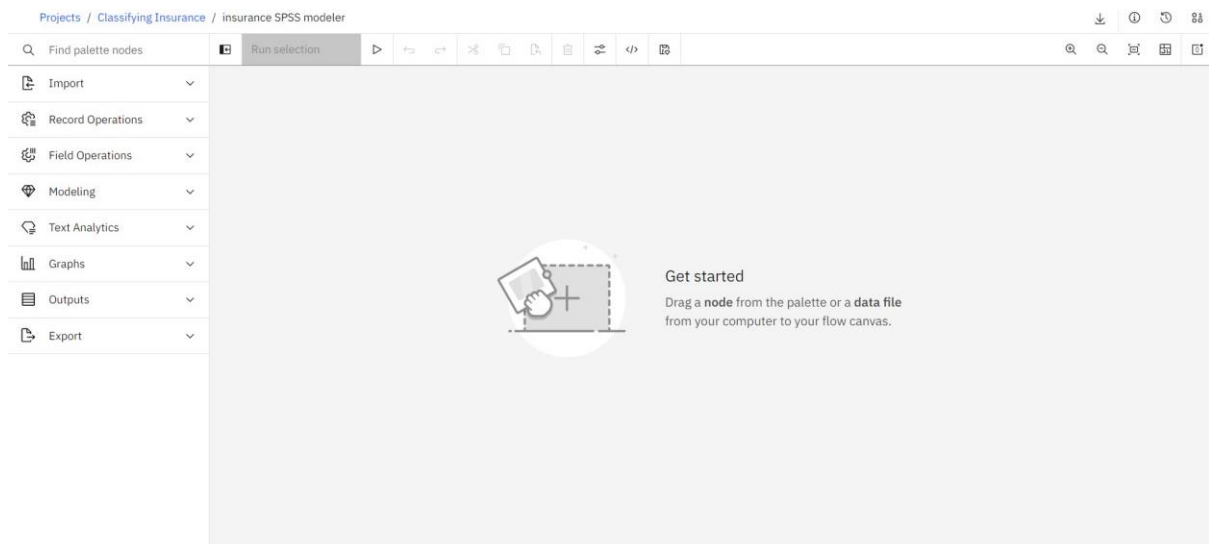
Visualize the data



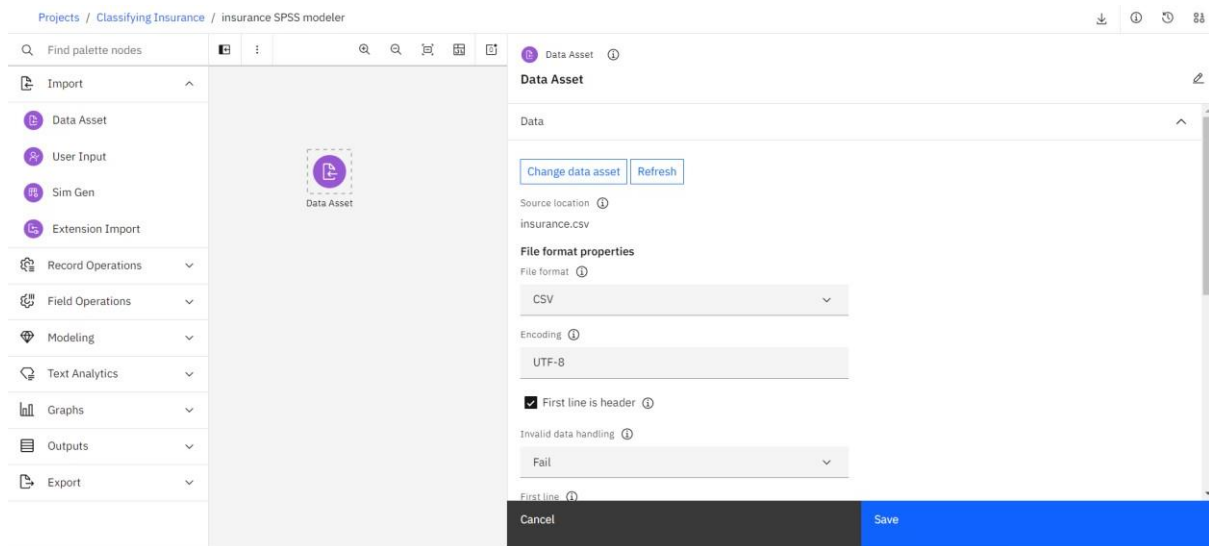
Visualize with other parameters

The screenshot shows a 'Create an SPSS Modeler flow' dialog box. It has a sidebar on the left with options: '+ New', 'Gallery sample', and 'Local file'. The main area is divided into two sections: 'Define details' and 'Define configuration'. In 'Define details', there's a 'Name' field with the value 'insurance SPSS modeler' and a 'Description (Optional)' text area with the placeholder text 'What's the purpose of this SPSS Modeler flow?'. In 'Define configuration', there's an 'Environment definition' dropdown menu set to 'Default SPSS Modeler S (2 vCPU 8 GB RAM)'. At the bottom right, there are 'Cancel' and 'Create' buttons.

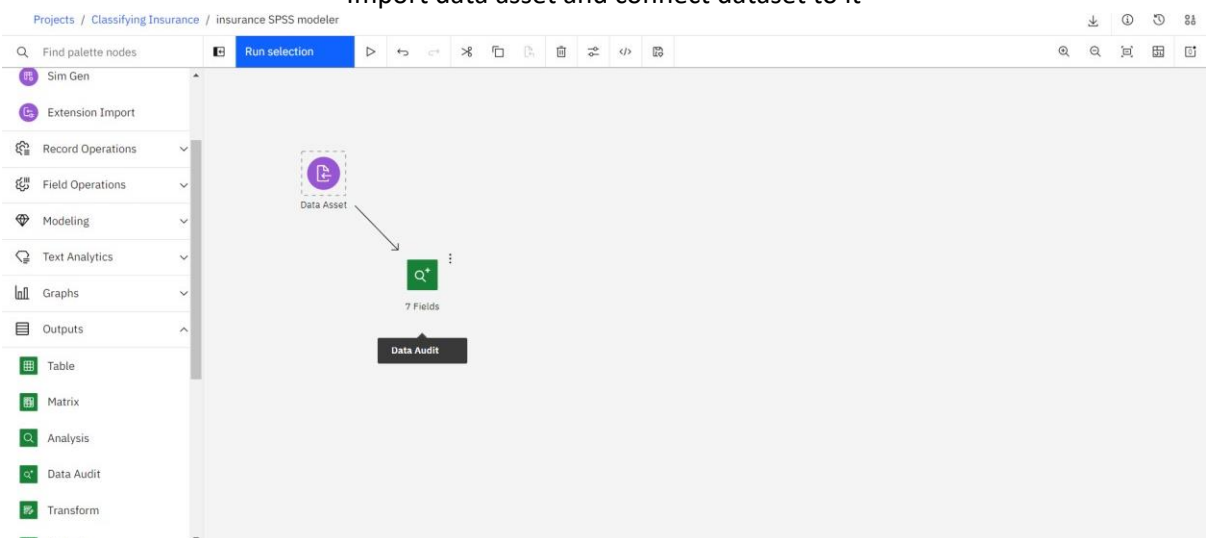
Create a SPSS modeler



Workspace



Import data asset and connect dataset to it



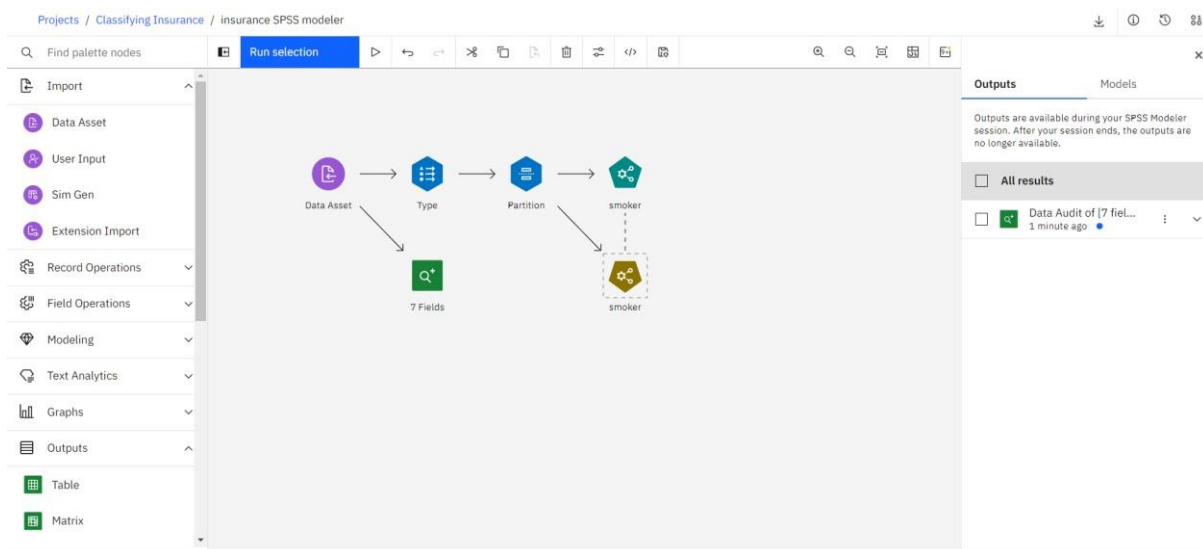
Connect data audit to data asset and run model

Projects / Classifying Insurance / insurance SPSS modeler

View Output: Data Audit of [7 fields]

	Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
1	age		Continuous	18	64	39.207	14.050	0.056	--	1338
2	sex		Categorical	--	--	--	--	--	2	1338
3	bmi		Continuous	15.960	53.130	30.663	6.098	0.284	--	1338
4	children		Continuous	0	5	1.095	1.205	0.938	--	1338
5	smoker		Categorical	--	--	--	--	--	2	1338
6	region		Categorical	--	--	--	--	--	4	1338

Data Audit output



Auto Classifier model

Projects / Classifying Insurance / insurance SPSS modeler

View Model: smoker

Auto Classifier

Models

Auto Classifier - Models

TARGET : SMOKER

USE	MODEL NAME	ESTIMATOR	BUILD TIME (MINS)	NO. FIELDS USED	ACCURACY	ACCUMULATED ACCURACY	AREA UNDER CURVE	ACCUMULATED AUC	RECALL	PRECISION
<input checked="" type="checkbox"/>	XGBoost Tree 1	XGBoost Binary Classification Model	< 1	3	78.870	78.870	0.501	0.501	0.025	17.000
<input checked="" type="checkbox"/>	Neural Net 1	MLP Neural Network	< 1	3	80.098	80.098	0.540	0.540	0.000	0.000
<input checked="" type="checkbox"/>	C5.1	C5.0	< 1	3	80.098	80.098	0.500	0.500	0.000	15.000
<input checked="" type="checkbox"/>	C&R Tree 1	C&RT	< 1	3	80.098	80.098	0.500	0.500	0.000	16.000

View model

Projects / Classifying Insurance / insurance SPSS modeler

View Output: Table (10 fields, 1,338 records) [Compare](#)

	age	sex	bmi	children	smoker	region	premium	Partition	\$XF-smoker	\$XFC-smoker
1	19	female	27.900	0	yes	southwest	16884.924	1_Training	no	0.751
2	18	male	33.770	1	no	southeast	1725.552	1_Training	no	0.803
3	28	male	33.000	3	no	southeast	4449.462	1_Training	no	0.790
4	33	male	22.705	0	no	northwest	21984.471	2_Testing	no	0.787
5	32	male	28.880	0	no	northwest	3866.855	1_Training	no	0.803
6	31	female	25.740	0	no	southeast	3756.622	1_Training	no	0.799
7	46	female	33.440	1	no	southeast	8240.590	1_Training	no	0.792
8	37	female	27.740	3	no	northwest	7281.506	1_Training	no	0.758
9	37	male	29.830	2	no	northeast	6406.411	1_Training	no	0.818
10	60	female	25.840	0	no	northwest	28923.137	1_Training	no	0.829
11	25	male	26.220	0	no	northeast	2721.321	1_Training	no	0.821
12	62	female	26.290	0	yes	southeast	27808.725	1_Training	no	0.638
13	23	male	34.400	0	no	southwest	1826.843	2_Testing	no	0.803
14	56	female	39.820	0	no	southeast	11090.718	1_Training	no	0.826
15	27	male	42.130	0	yes	southeast	39611.758	2_Testing	no	0.783

Table output

Projects / Classifying Insurance / insurance SPSS modeler

View Output: Analysis of [smoker] [Compare](#)

[Collapse All](#)

Results for output field smoker

Comparing \$XF-smoker with smoker

'Partition'	1_Training		2_Testing	
Correct	738	79.27%	326	80.1%
Wrong	193	20.73%	81	19.9%
Total	931		407	

Analysis output