

Sayak Dey 19BOE10072

VIT Bhopal

SmartInternz

Assignment-3

Dataset : Diabetes.csv

Overview:

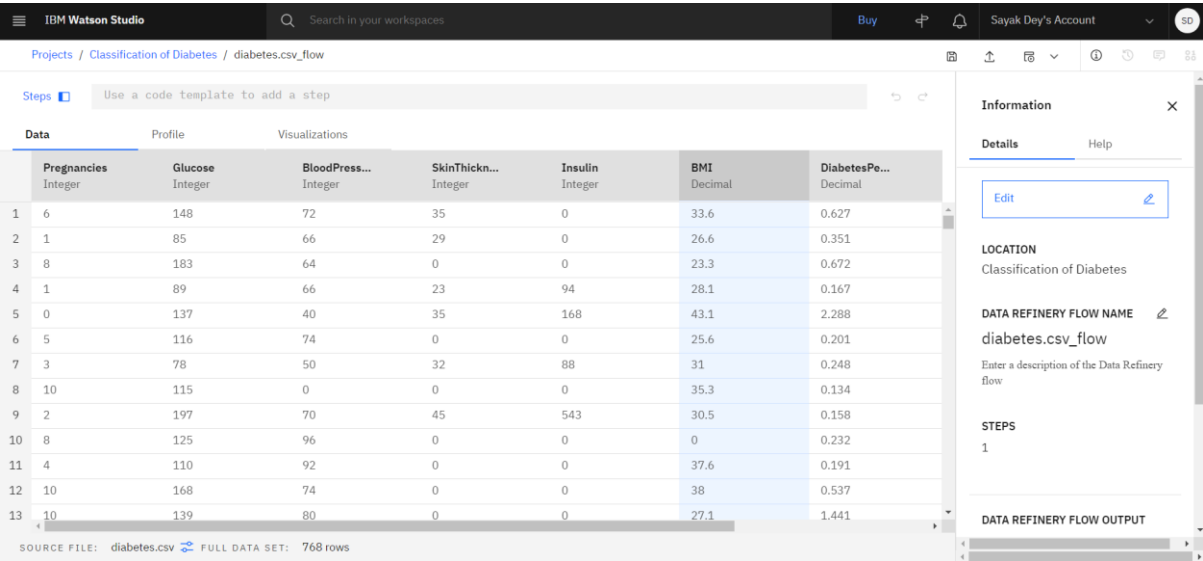
The screenshot shows the IBM Watson Studio interface. The top navigation bar includes the IBM Watson Studio logo, a search bar, and user account information. The main content area is titled 'Classification of Diabetes' and has tabs for Overview, Assets, Jobs, and Manage. The 'Assets' tab is active, displaying a list of assets. On the left, there's a sidebar with '3 assets' and 'Asset types' including Data (1) and Flows (2). The main table lists two flows: 'Classification of Diabetes SPSS Modeler flow' (modified 6 days ago) and 'diabetes.csv_flow Data Refinery flow' (modified 1 week ago). At the bottom, it shows 'Items per page: 20' and '1-2 of 2 items'.

Uploaded Dataset:

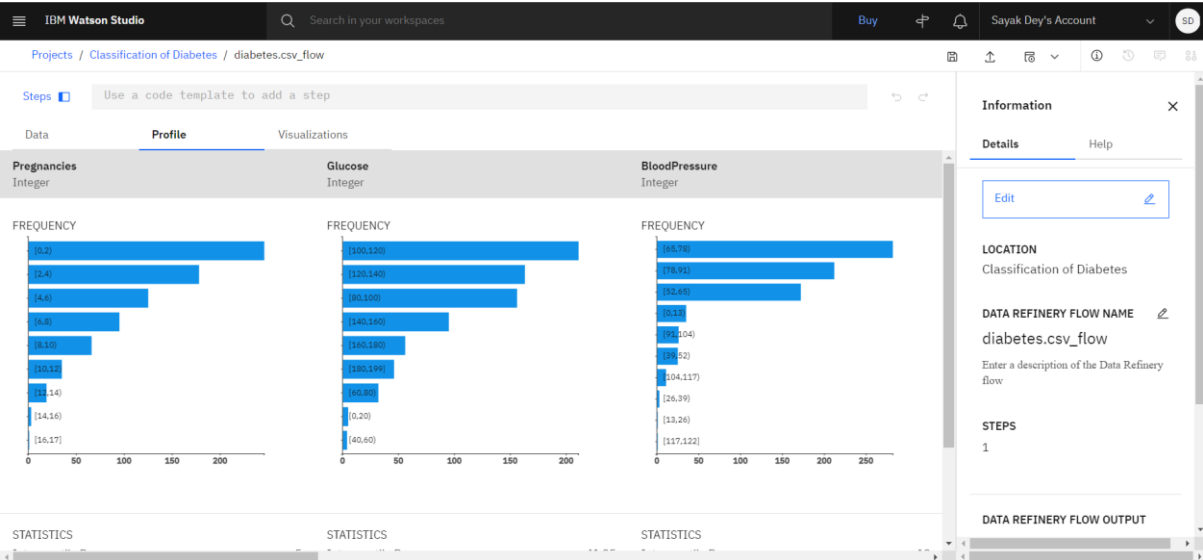
The screenshot shows the IBM Watson Studio interface for the 'diabetes.csv' dataset. The top navigation bar is the same as the previous screenshot. The main content area is titled 'My projects / Classification of Diabetes / diabetes.csv' and has tabs for Preview and Visualization. The 'Preview' tab is active, displaying a table with 9 columns and 10 rows. The columns are: Pregnanc... String, Glucose String, BloodPres... String, SkinThick... String, Insulin String, BMI String, DiabetesPedigreeFun... String, Age String, and Diabetes String. The table shows various numerical and categorical values. On the right, there's an 'Information' panel with details about the dataset, including its name, description, tags, creator (Sayak Dey), usage, creation date (Apr 23, 2022, 12:27 PM), and size (25,412 KB).

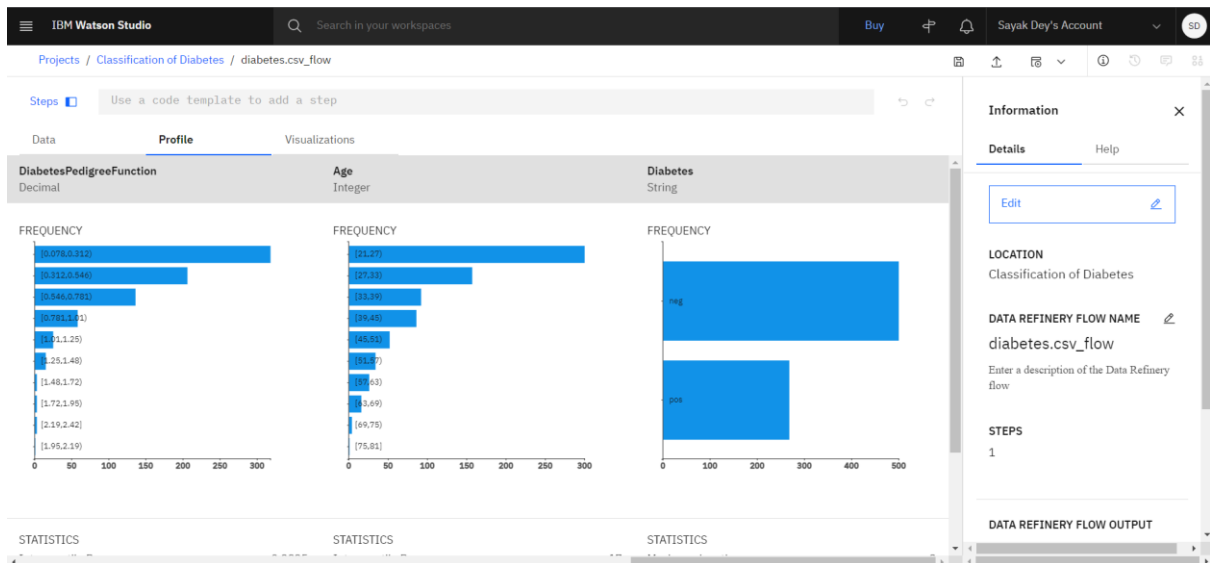
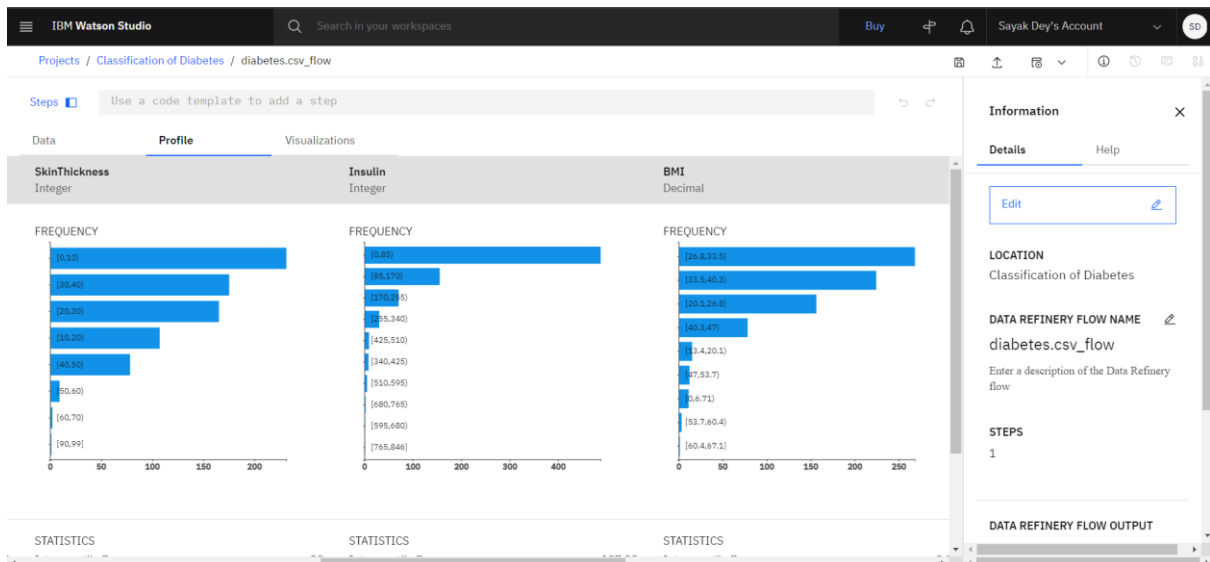
Pregnanc... String	Glucose String	BloodPres... String	SkinThick... String	Insulin String	BMI String	DiabetesPedigreeFun... String	Age String	Diabetes String
6	148	72	35	0	33.6	0.627	50	pos
1	85	66	29	0	26.6	0.351	31	neg
8	183	64	0	0	23.3	0.672	32	pos
1	89	66	23	94	28.1	0.167	21	neg
0	137	40	35	168	43.1	2.288	33	pos
5	116	74	0	0	25.6	0.201	30	neg
3	78	50	32	88	31	0.248	26	pos
10	115	0	0	0	35.3	0.134	29	neg
2	197	70	45	543	30.5	0.158	53	pos

Data refinery Flow:

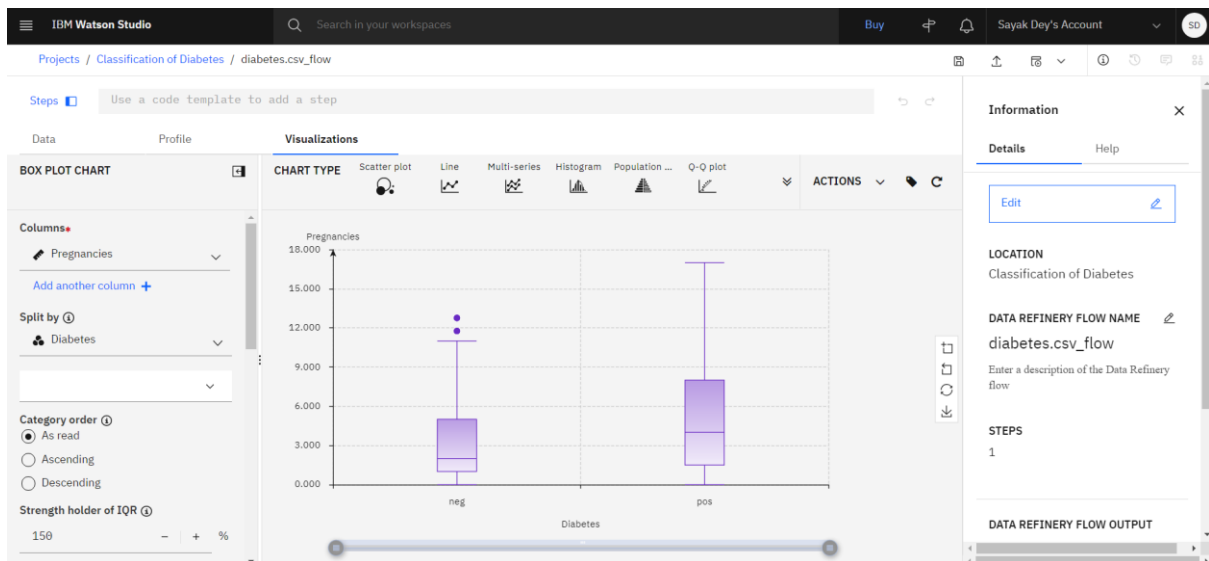
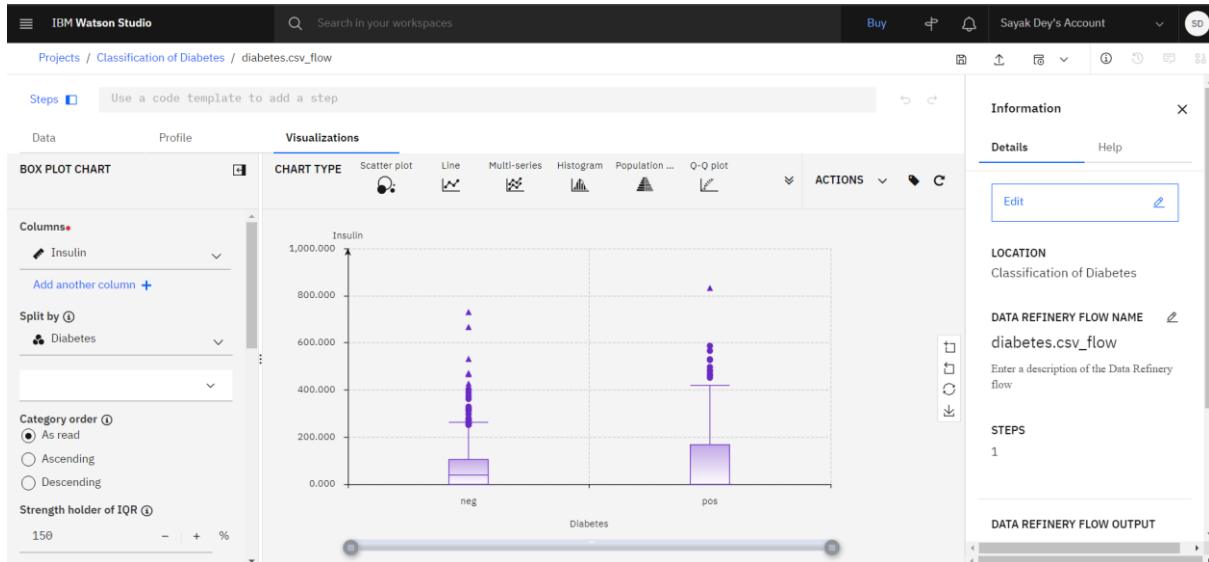


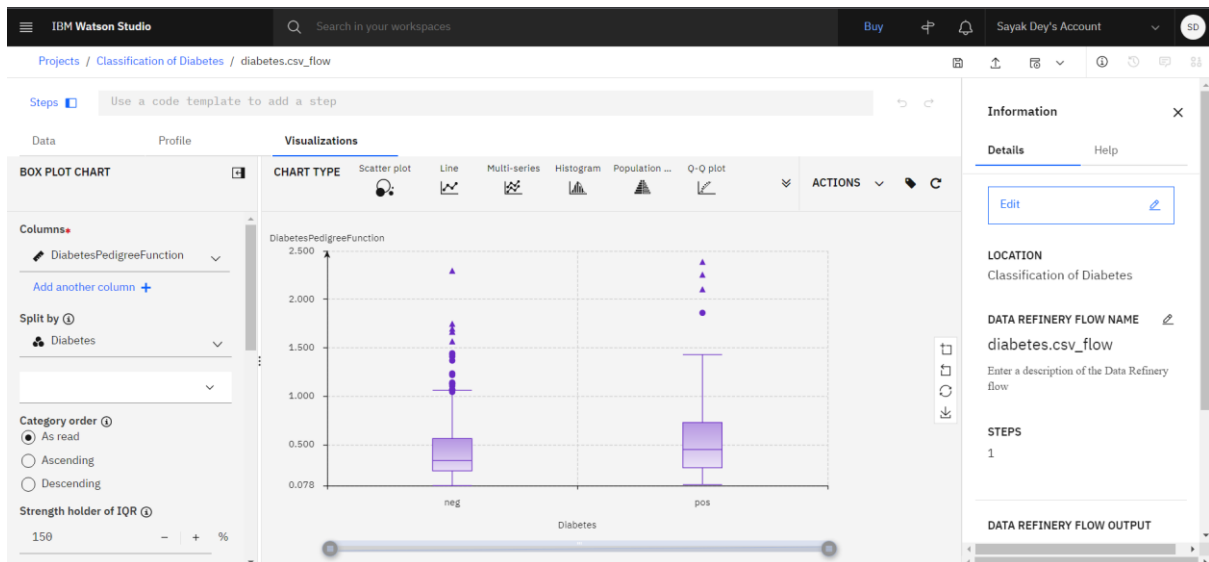
Data Profile:



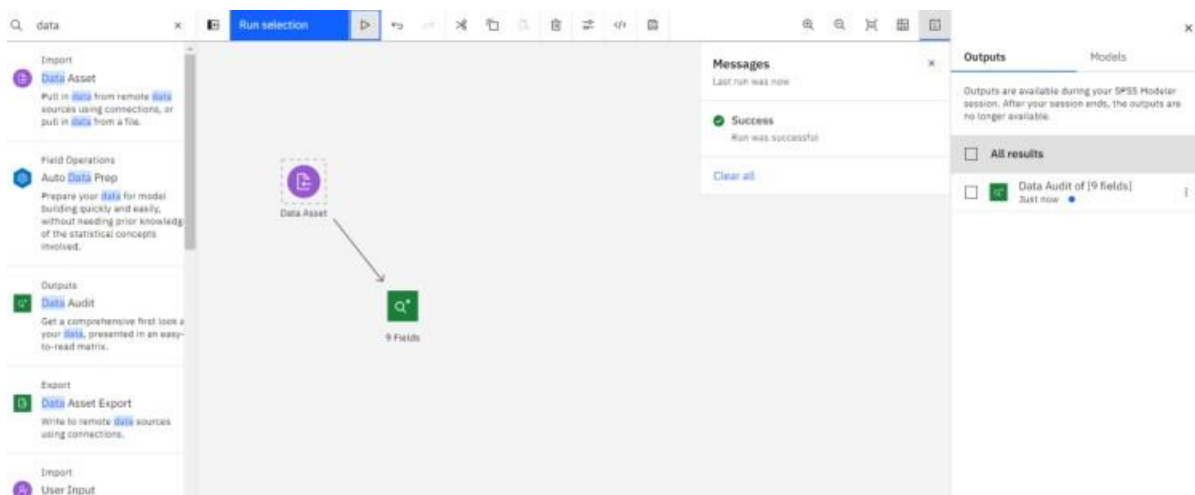


Data Visualization:





Import Data to SPSS Modeler and Build Model:









Audit Output:

Projects / Classification of Diabetes / Classification Of Diabetes

View Output: Data Audit of [9 fields]

Compare

	Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
1	Pregnancies		Continuous	0	17	3.845	3.370	0.902	--	768
2	Glucose		Continuous	0	199	120.895	31.973	0.174	--	768
3	BloodPressure		Continuous	0	122	69.105	19.356	-1.844	--	768
4	SkinThickness		Continuous	0	99	20.536	15.952	0.109	--	768
5	Insulin		Continuous	0	846	79.799	115.244	2.272	--	768
6	BMI		Continuous	0.000	67.100	31.993	7.884	-0.429	--	768

Setting the Target:

Projects / Classification of Diabetes / Classification Of Diabetes

type

Field Operations

- Type: Specify field metadata and properties that are invaluable to modeling.
- Filler: Replace field values and change storage. Often used with a Type node to replace missing values.
- Multiplot: A special type of plot that displays multiple Y fields over a single X field.
- Distribution: Shows the occurrence of symbolic (non-numeric) values, such as mortgage type or gender, in a dataset.

Data Asset → Type → 8 Fields

Type

Settings

Read values Clear values

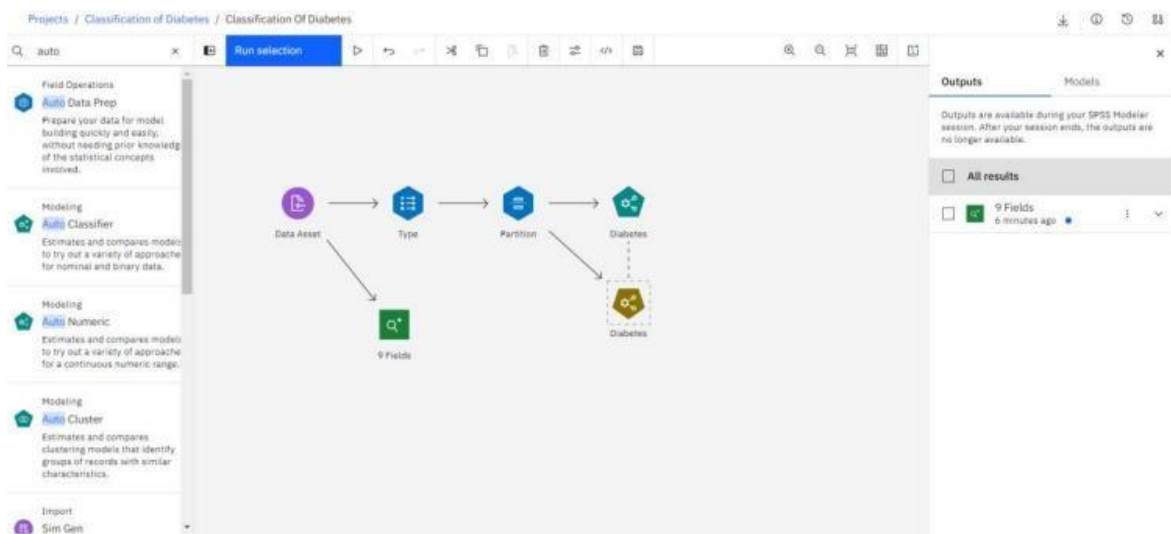
Find in column Field

Field	Measure	Role	Value mode	Values
<input type="checkbox"/> # Pregnancies	Continuous	Input	Read	
<input type="checkbox"/> # Glucose	Continuous	Input	Read	
<input type="checkbox"/> # BloodPressure	Continuous	Input	Read	
<input type="checkbox"/> # SkinThickness	Continuous	Input	Read	
<input type="checkbox"/> # Insulin	Continuous	Input	Read	
<input type="checkbox"/> # BMI	Continuous	Input	Read	
<input type="checkbox"/> # DiabetesPedigree	Continuous	Input	Read	
<input type="checkbox"/> # Age	Continuous	Input	Read	

Default mode: ☒ Read metadata ☐ Pass (do not scan)

Cancel Save

Auto Classifier Model:



Checking Which Model will Give the most accurate Prediction:

View Model: Diabetes

Auto Classifier ①

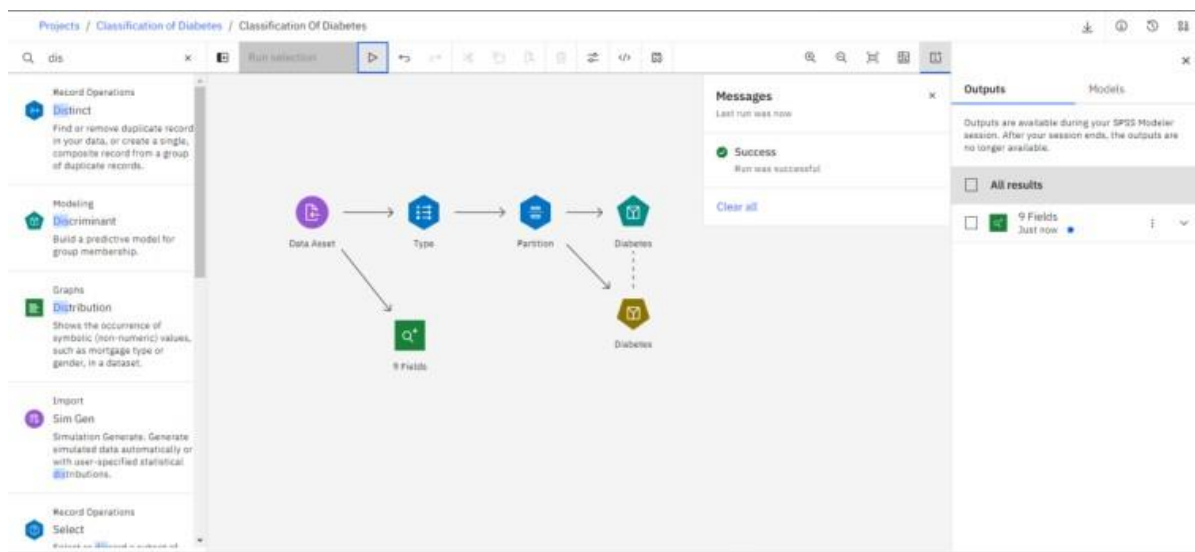
Models

Auto Classifier - Models ②

TARGET: 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-A5.1	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	0.748	0.748	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

Discriminant Model:



Feature Importance according to Discriminant Model:

