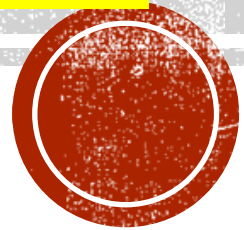
LECTURE NO : 13

LOGISTIC REGRESSION:



WHAT IS K-MEDOID CLUSTERING

Logistic regression is a statistical method used for binary classification that can be extended to multiclass classification under certain conditions. It's used to predict the probability of a categorical dependent variable—typically a binary outcome based on one or more predictor variables (features).

LOGISTIC REGRESSION WITH BINARY CLASSIFICATION

Logistic Regression is a machine learning algorithm based on supervised learning.

It is a statistical method that is used for predicting probability of target variable.
Logistic Regression makes probability for classification problems that are discrete in nature.

Example: English or Hindi, True or False, 1 or 0, Right or Wrong, cat or dog or goat.

- Binary Classification Ex: win or loss, dead or alive
- Multiclass Classification Ex: Onion or Potato or Sweet Potato, Lily or Sunflower or Rose

BASICS OF LOGISTIC REGRESSION:

- **Outcome:** Logistic regression deals with scenarios where the outcome is binary (yes/no, true/false, success/failure). The classic example is predicting whether an email is spam or not.

- **Probability Estimation:** Unlike linear regression that outputs a continuous number, logistic regression estimates the probability that a given input point belongs to a certain class.

APPLICATIONS OF LOGISTIC REGRESSION:

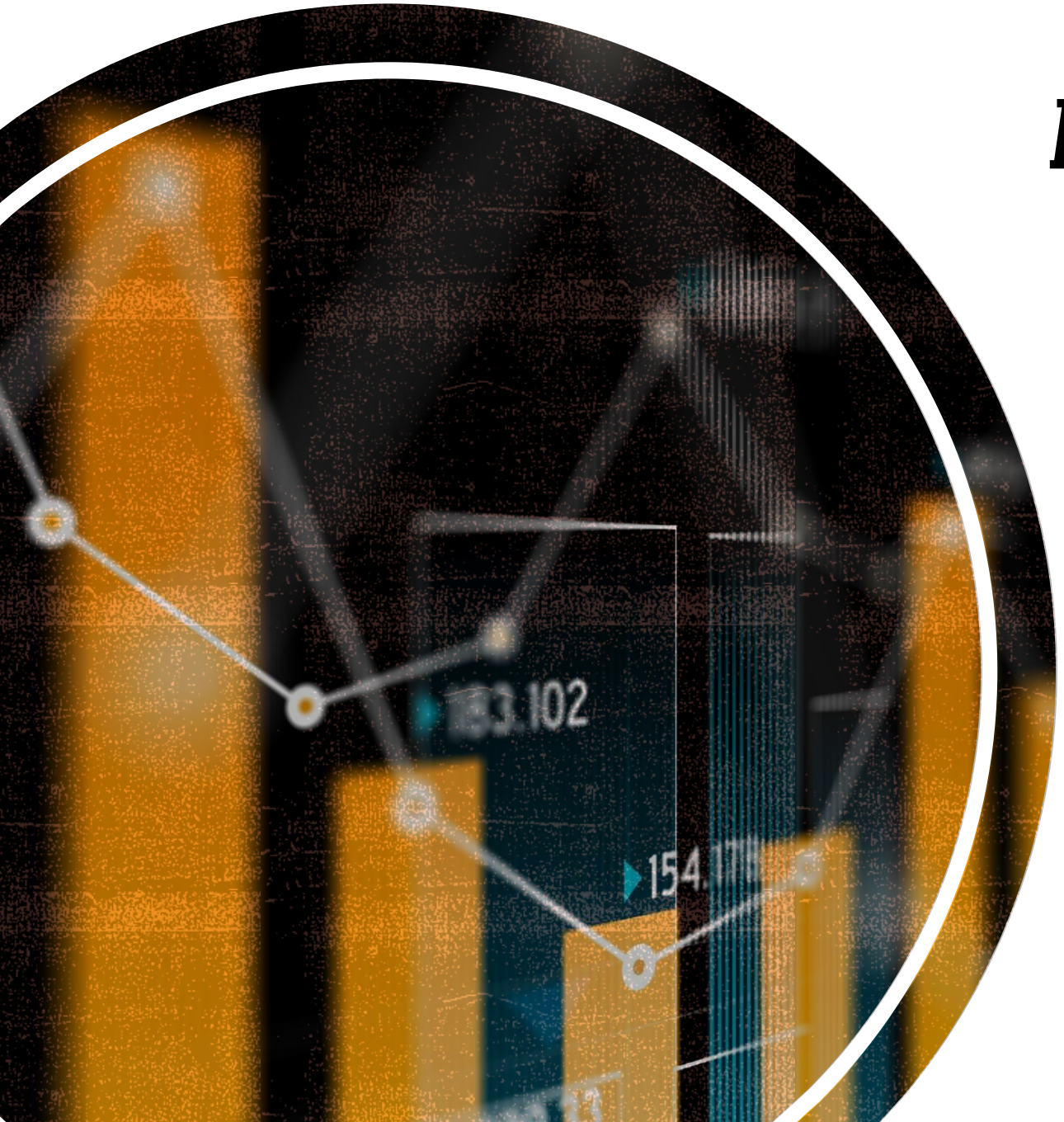
- **Medical Fields:** Predicting whether a patient has a disease (yes/no) based on observed characteristics of the patient (age, sex, body measurements, etc.).
- **Financial Sectors:** Predicting whether someone will default on a loan based on their financial history.
- **Marketing:** Predicting whether a customer will purchase a product or subscribe to a service.

APPLICATIONS OF LOGISTIC REGRESSION

Heart Attack Prediction
Tumour Prediction

Credit-Card Fraud

Spam Detection



ADVANTAGES

Efficiency: Logistic regression is computationally inexpensive to run, easy to implement, and provides highly interpretable results.

EXPRESSION FOR LOGISTIC REGRESSION:

Sigmoid Function: The probabilities predicting yes/no are calculated with a logistic function, which outputs a value between 0 and 1. This function is also known as the sigmoid function. The formula for the sigmoid function is:

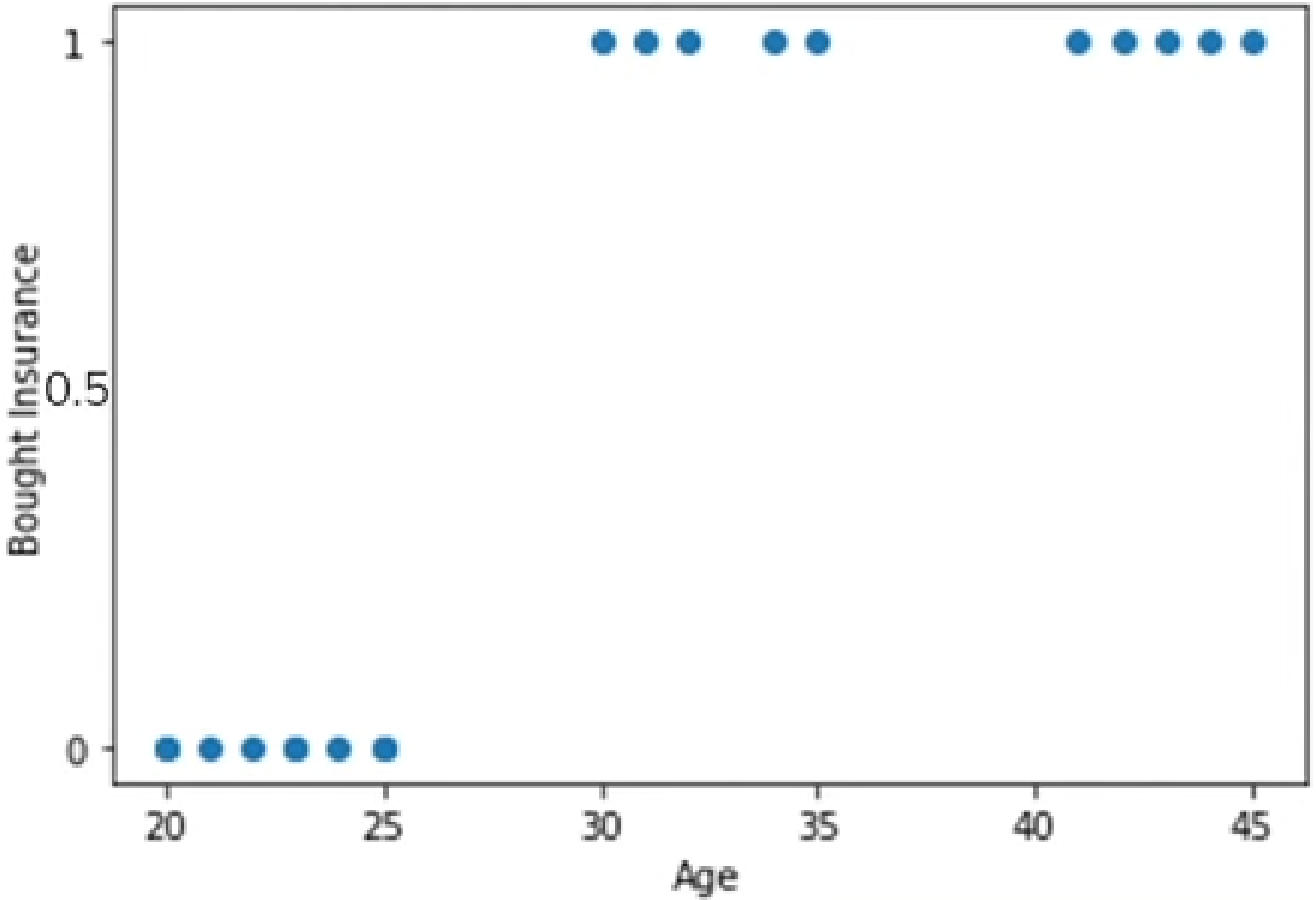
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

EXAMPLE:

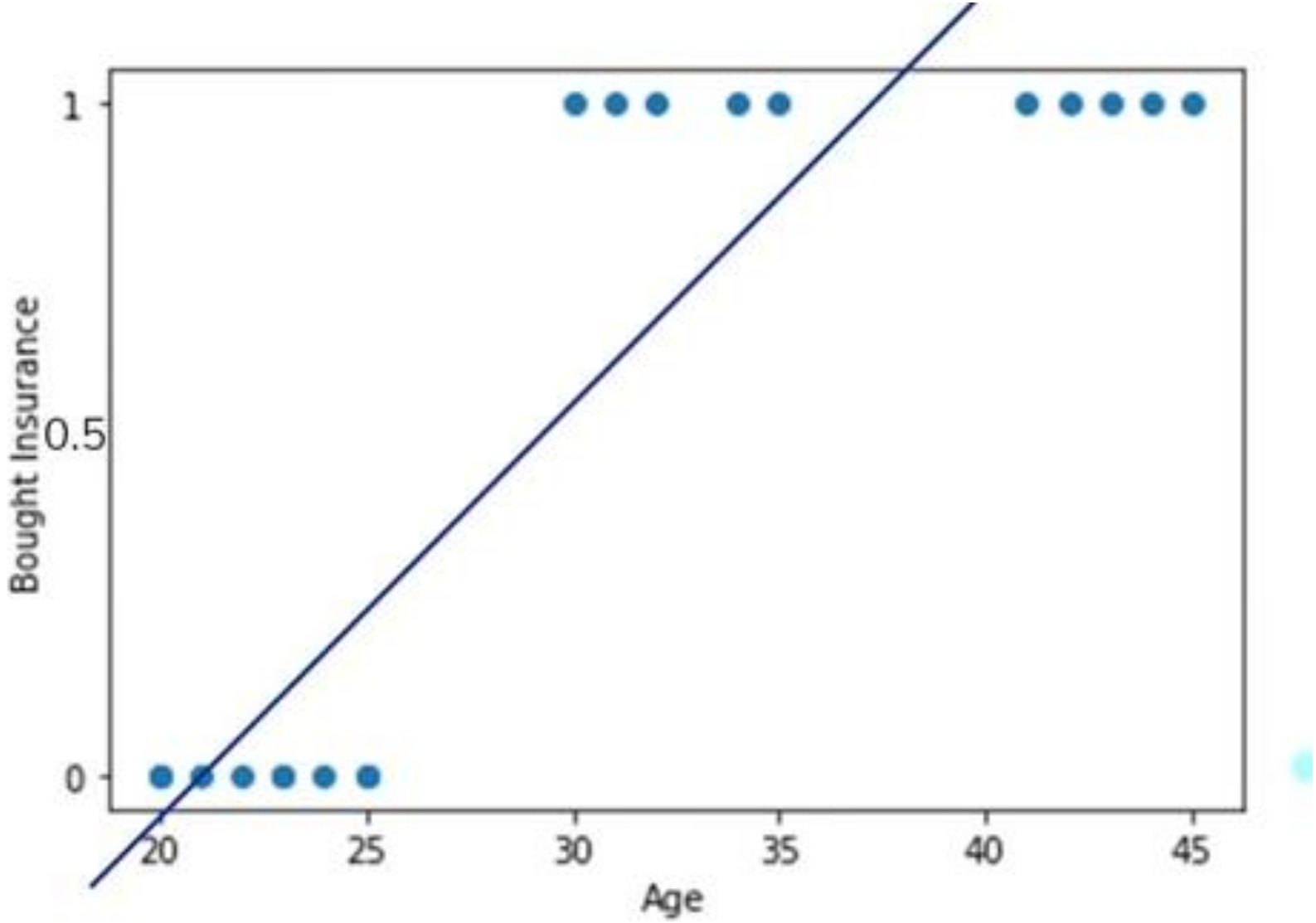
Age	Bought Insurance
21	no
48	yes
32	yes
41	yes
20	no
35	yes
20	no
23	no

Bought Insurance
0
1
1
1
0
1
0
0

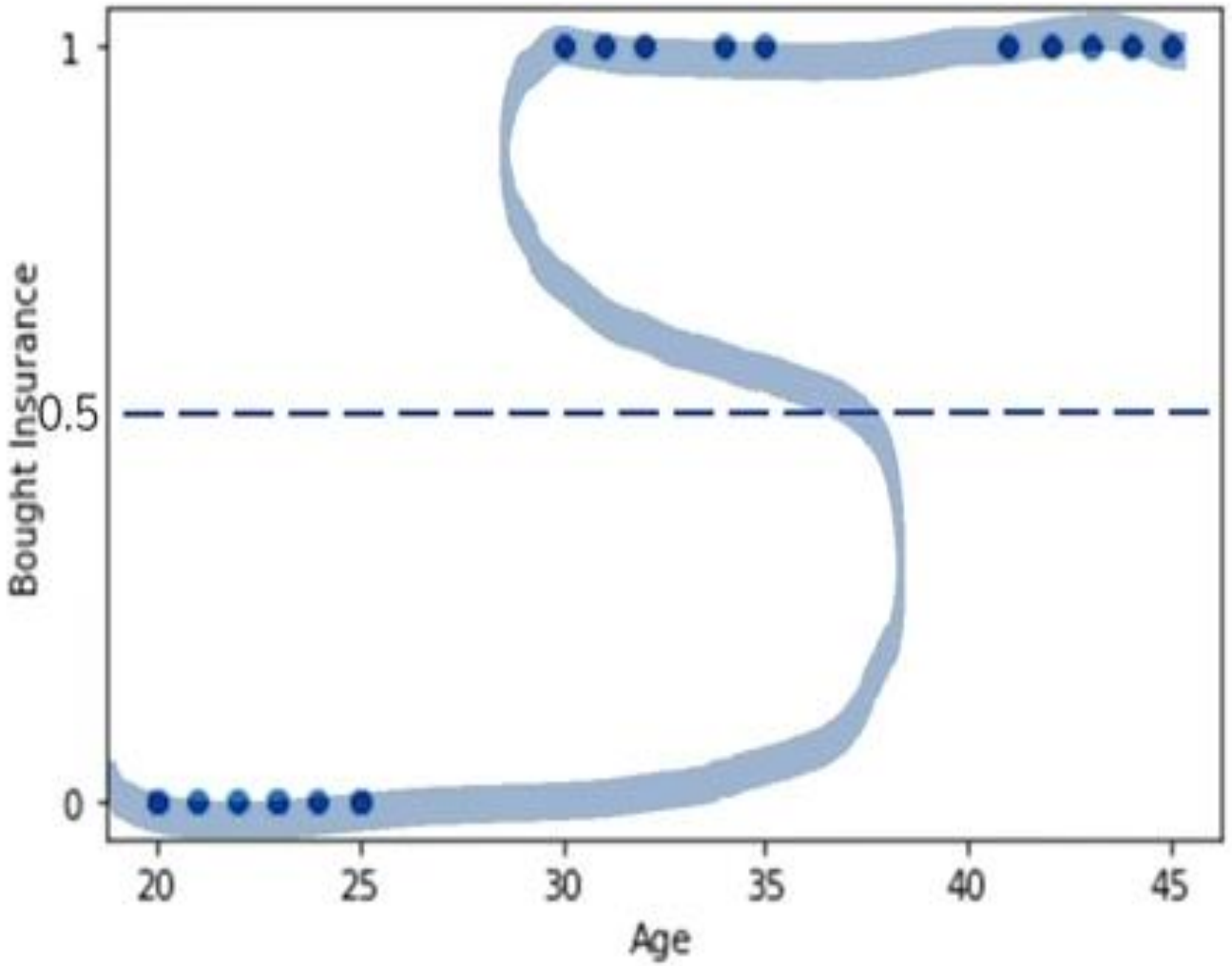
SCATTER PLOT VISUALIZATION:



PROBLEM WITH LINEAR REGRESSION:



SOLUTION:



$$y = 1 / 1+e^{-x}$$

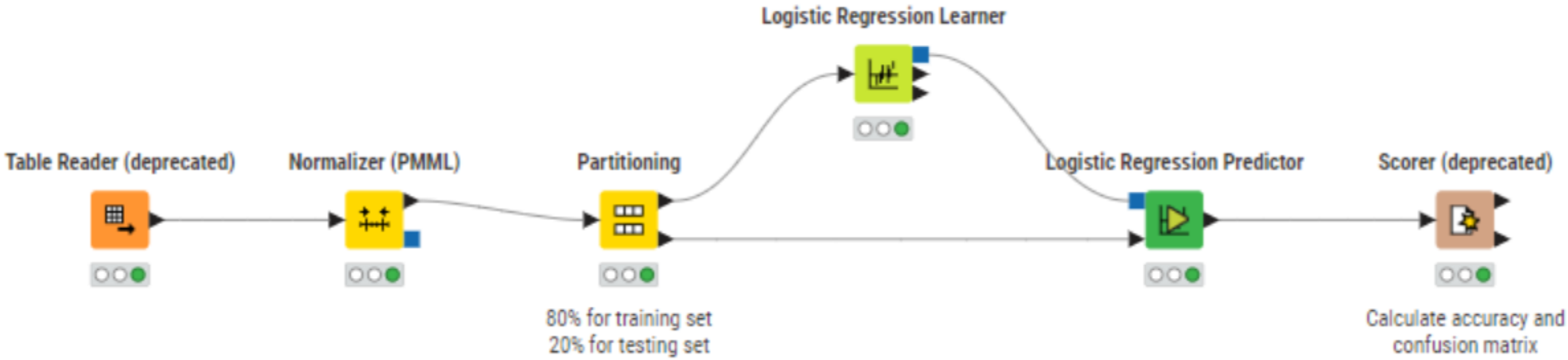
Sigmoid Function
Converts input (x) into range 0 or 1

y - dependent variable (Bought Insurance)

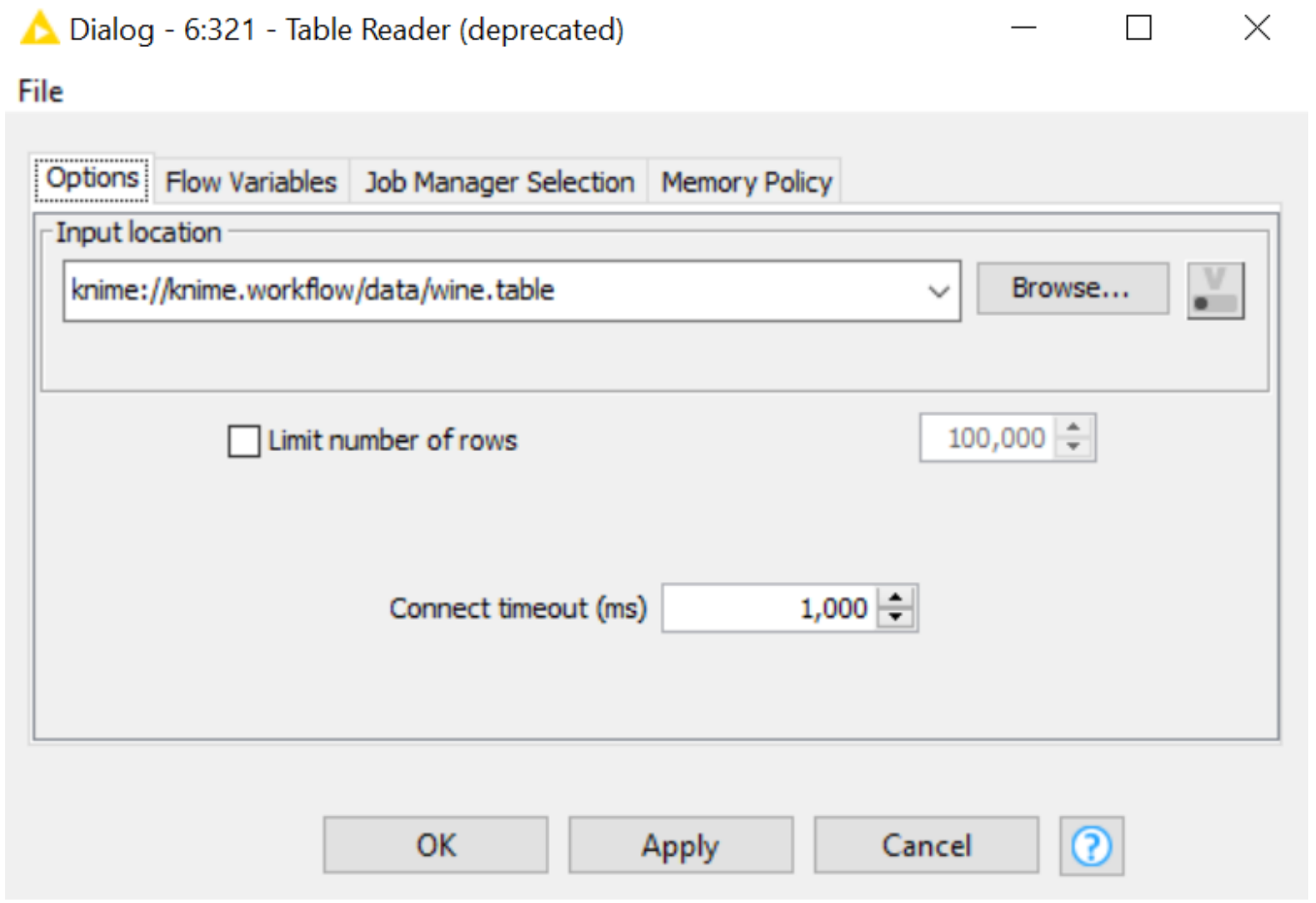
x - independent variable (age)

e - Euler's Constant ~ 2.71828

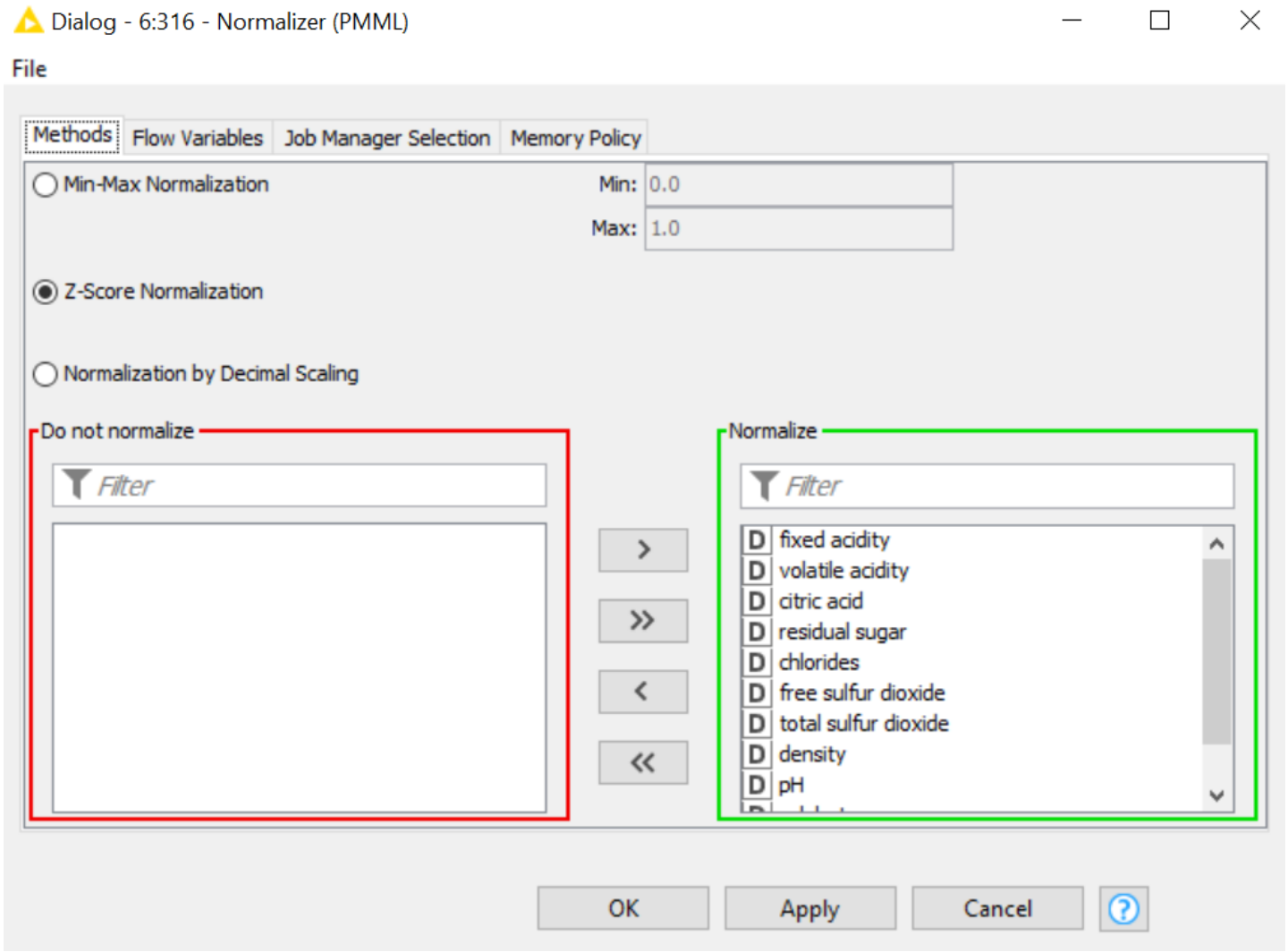
LOGISTIC REGRESSION IN KNIME:



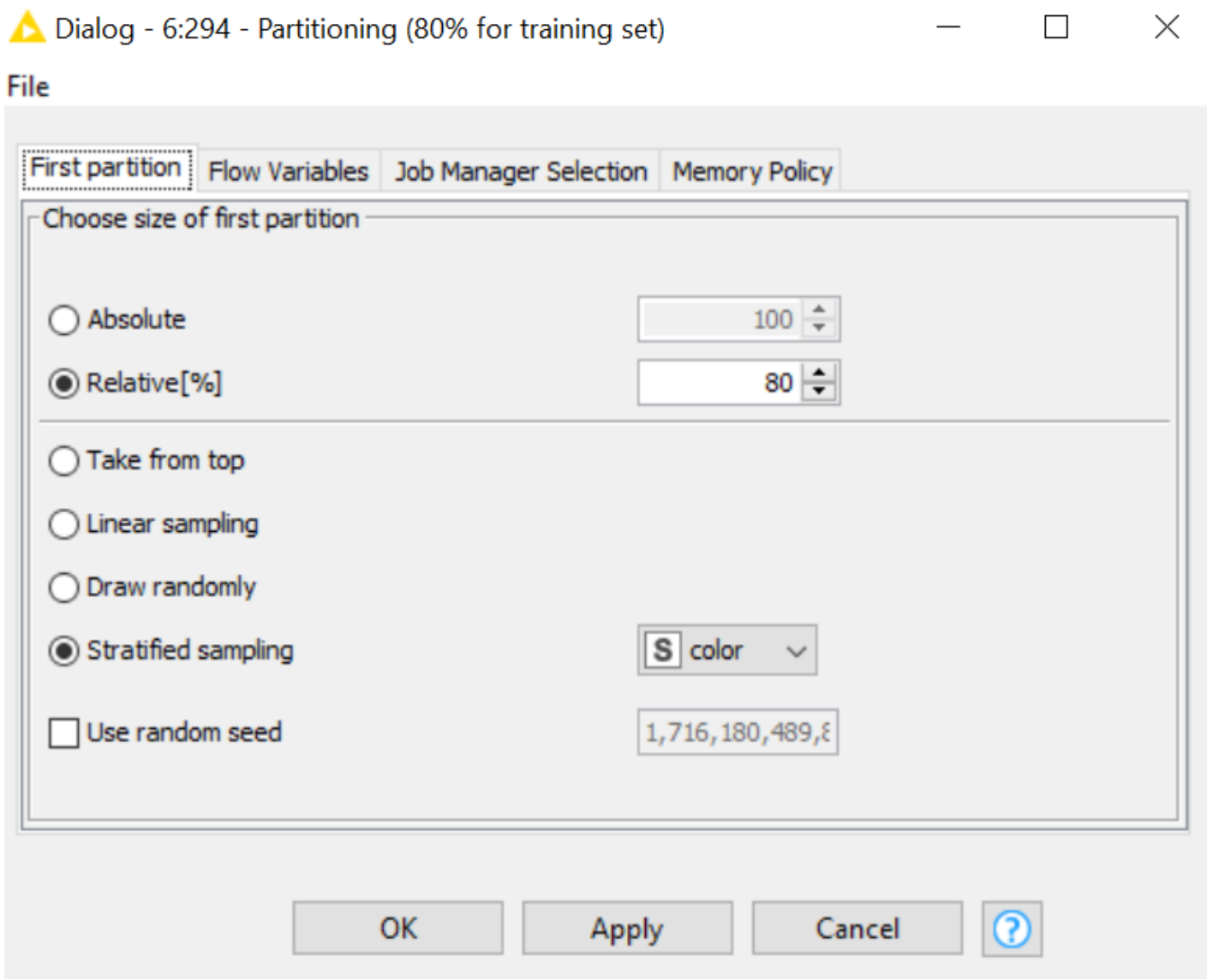
CONTINUE:


Dialog - 6:321 - Table Reader (deprecated)
File
Options
Flow Variables
Job Manager Selection
Memory Policy
Input location
knime://knime.workflow/data/wine.table
Browse...
Limit number of rows
100,000
Connect timeout (ms)
1,000
OK
Apply
Cancel

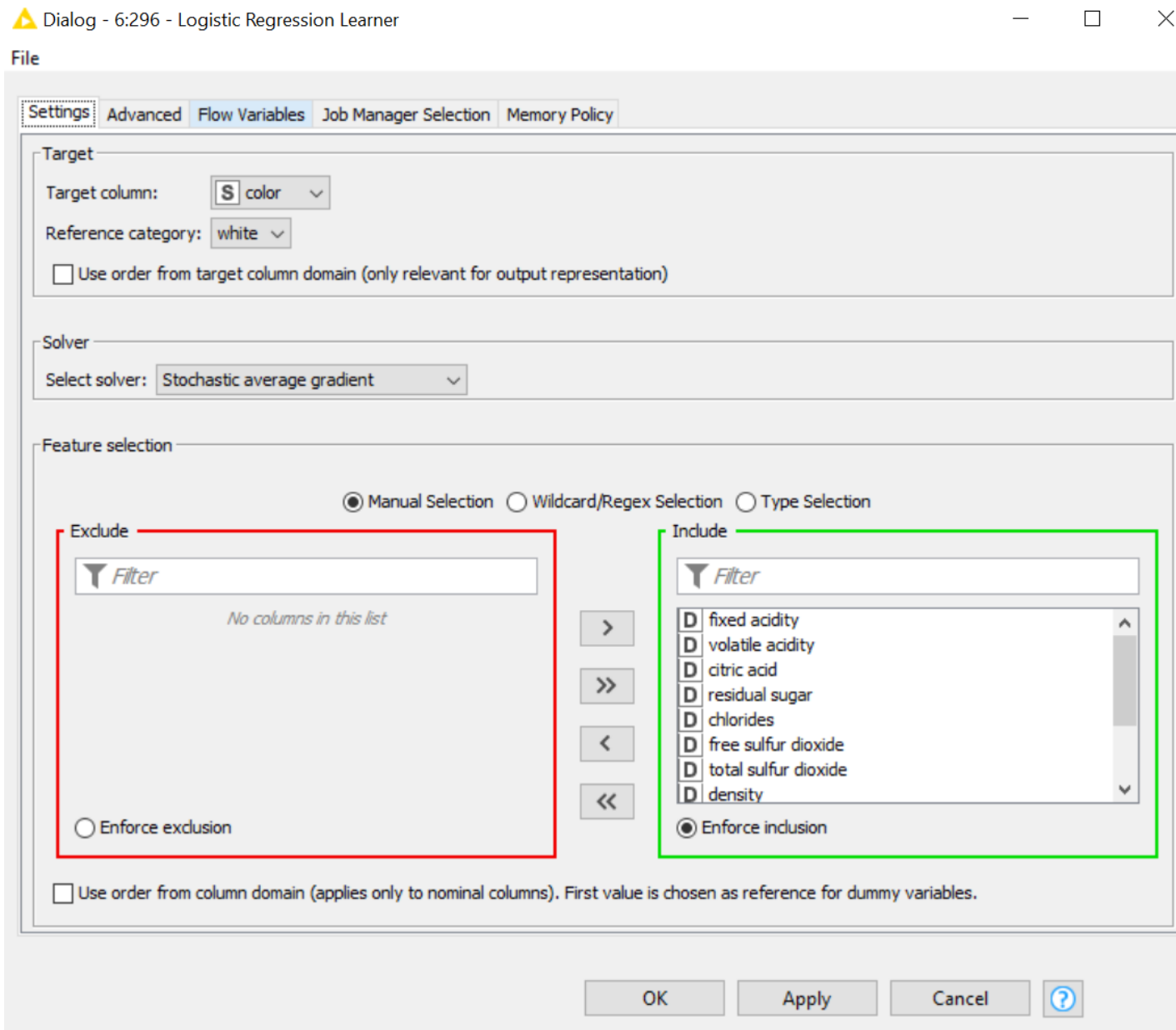
NORMALIZER:


Dialog - 6:316 - Normalizer (PMML)
File
Methods
Flow Variables
Job Manager Selection
Memory Policy
Min-Max Normalization
Min: 0.0
Max: 1.0
Z-Score Normalization
Normalization by Decimal Scaling
Do not normalize
Filter
Normalize
Filter
D fixed acidity
D volatile acidity
D citric acid
D residual sugar
D chlorides
D free sulfur dioxide
D total sulfur dioxide
D density
D pH
OK
Apply
Cancel

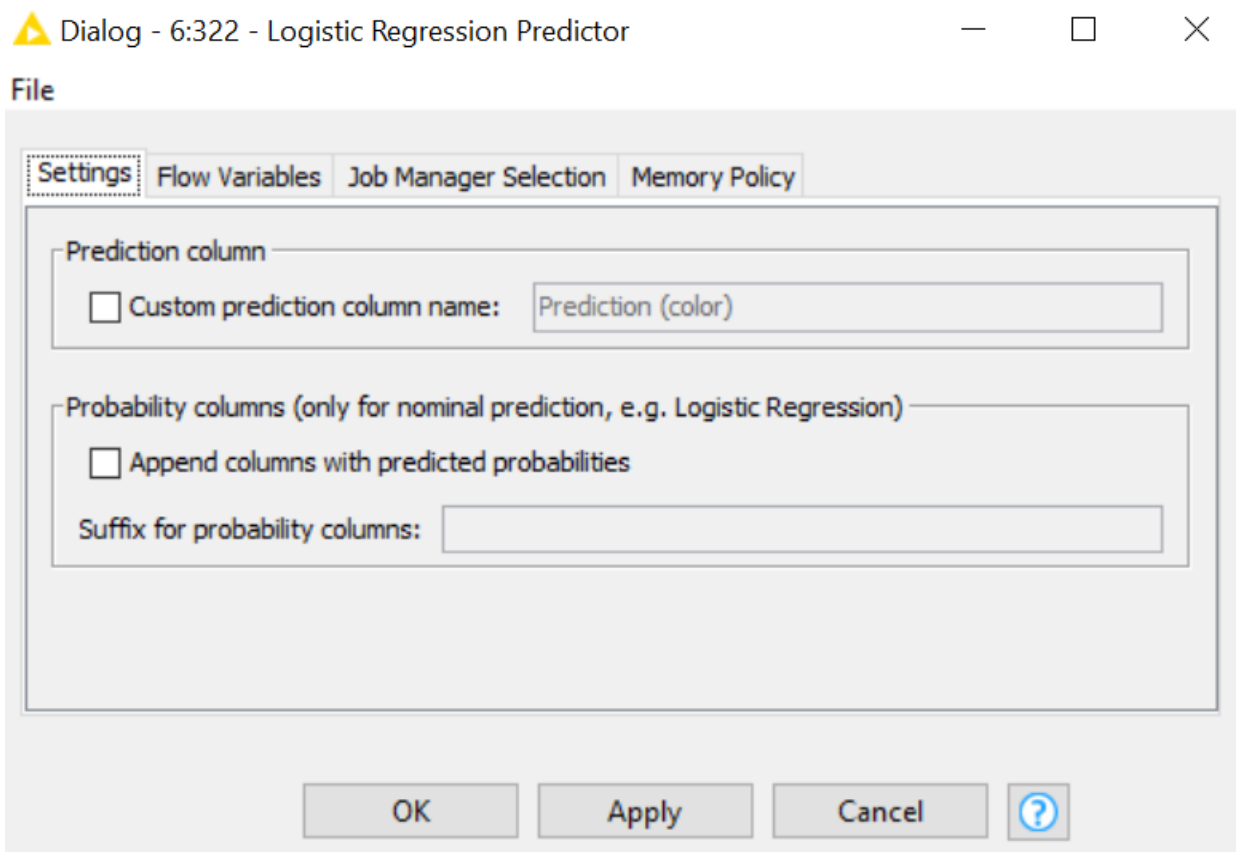
PARTITIONING:



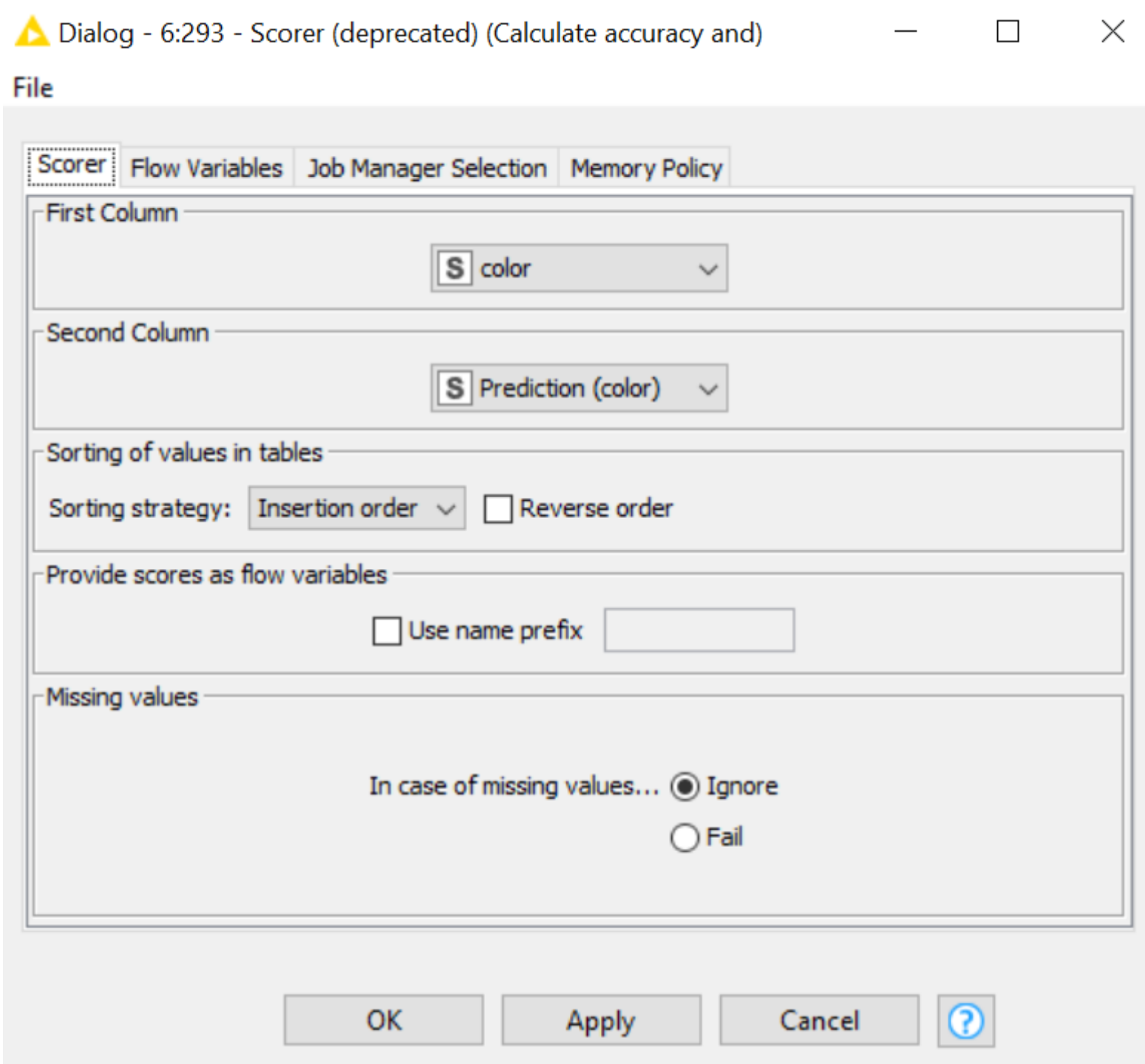
LOGISTIC REGRESSION LEARNER:



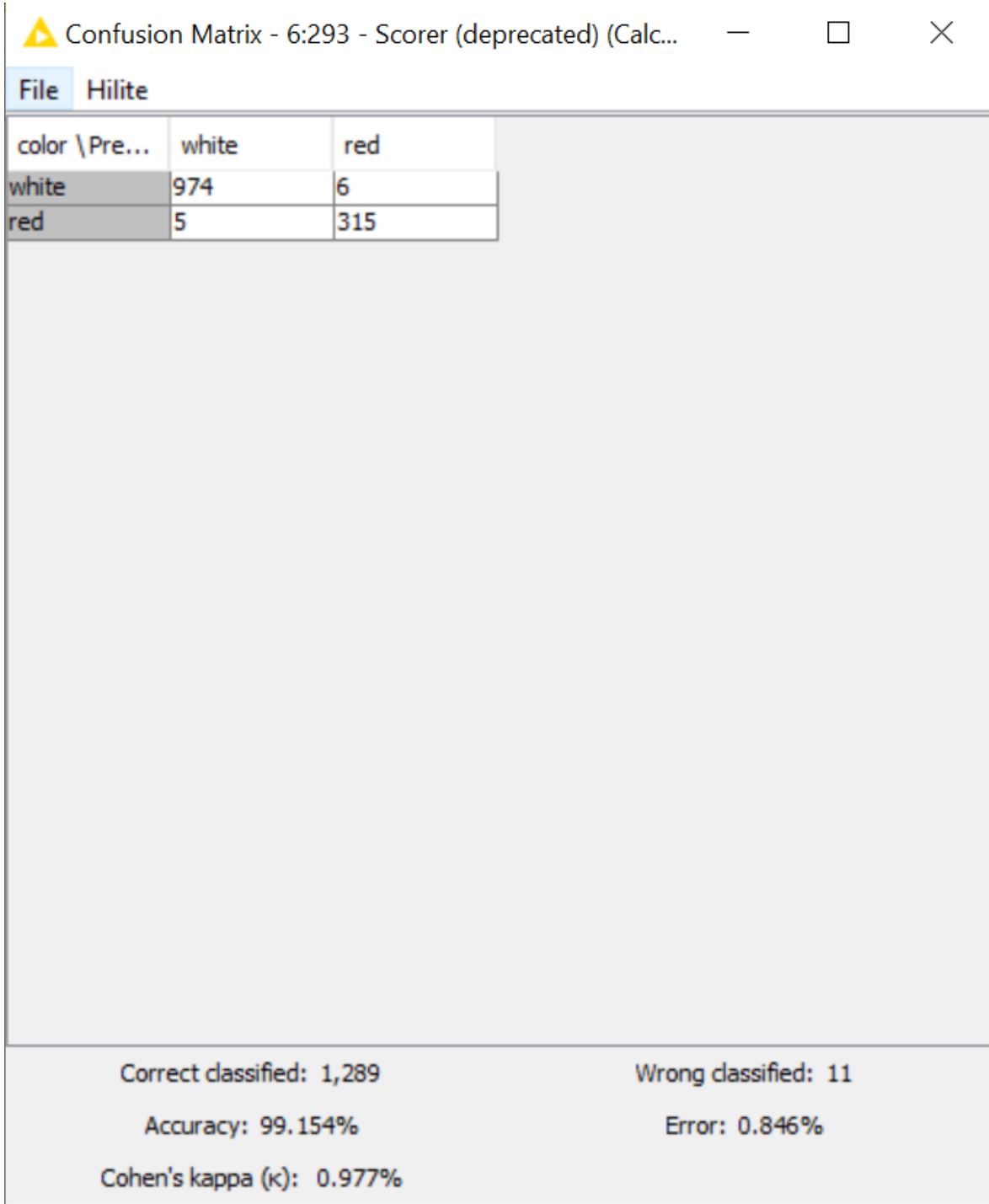
LOGISTIC REGRESSION PREDICTOR:



SCORER:


Dialog - 6:293 - Scorer (deprecated) (Calculate accuracy and)
File
Scorer
Flow Variables
Job Manager Selection
Memory Policy
First Column
S color
Second Column
S Prediction (color)
Sorting of values in tables
Sorting strategy:
Insertion order
Reverse order
Provide scores as flow variables
Use name prefix
Missing values
In case of missing values...
Ignore
Fail
OK
Apply
Cancel

CONFUSION MATRIX:




CODE IMPLEMENTATION
IN PYTHON:

LOGISTIC REGRESSION IN PYTHON

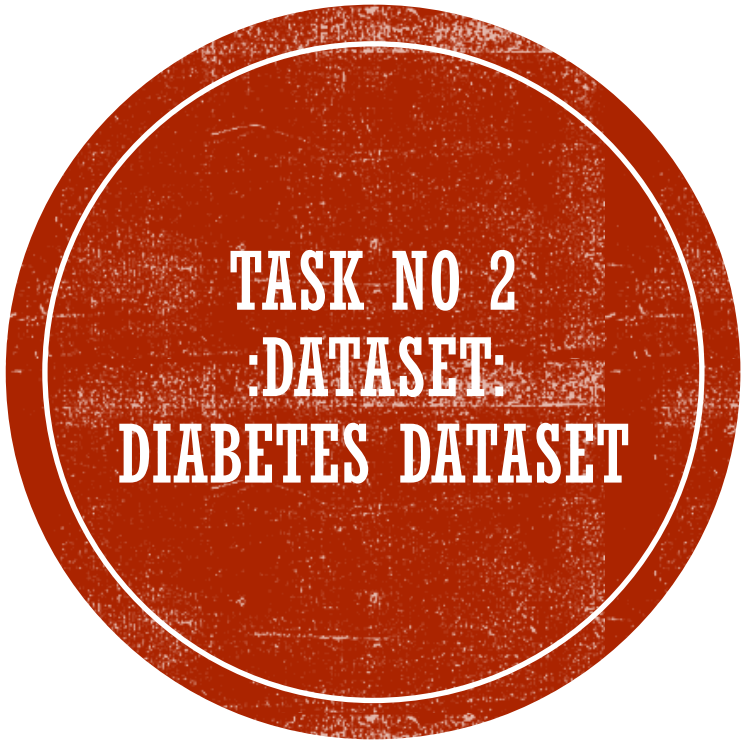
```python
import numpy as np
import pandas as pd
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Load the Iris dataset
iris = load_iris()
X = iris.data
y = (iris.target == 0).astype(int)  # 1 if Setosa, 0 otherwise

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Train a LogisticRegression model
clf = LogisticRegression(max_iter=100)
clf.fit(X_train, y_train)

# Make predictions
y_pred = clf.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred, target_names=['Non-Setosa', 'Setosa'])

# Print the results
print(f"Accuracy: {accuracy:.2f}")
print("Confusion Matrix:")
print(conf_matrix)
print("Classification Report:")
print(class_report)
```

TASKS:

TASK 1: TITANIC DATASET TASK

As a data scientist, you are part of a team that aims to analyze the Titanic dataset to predict passenger survival. The dataset includes attributes like class, sex, age, and fare. Logistic regression is the chosen method because of its suitability for binary classification tasks and its ability to provide interpretable results.

Develop a logistic regression model using KNIME that accurately predicts whether a passenger survived the Titanic disaster. The model will be evaluated based on its accuracy, precision.



You are a data analyst working for a healthcare organization, and your task is to predict the likelihood of diabetes in patients using logistic regression in Python. The goal is to identify high-risk individuals to enable early intervention and better management of diabetes.