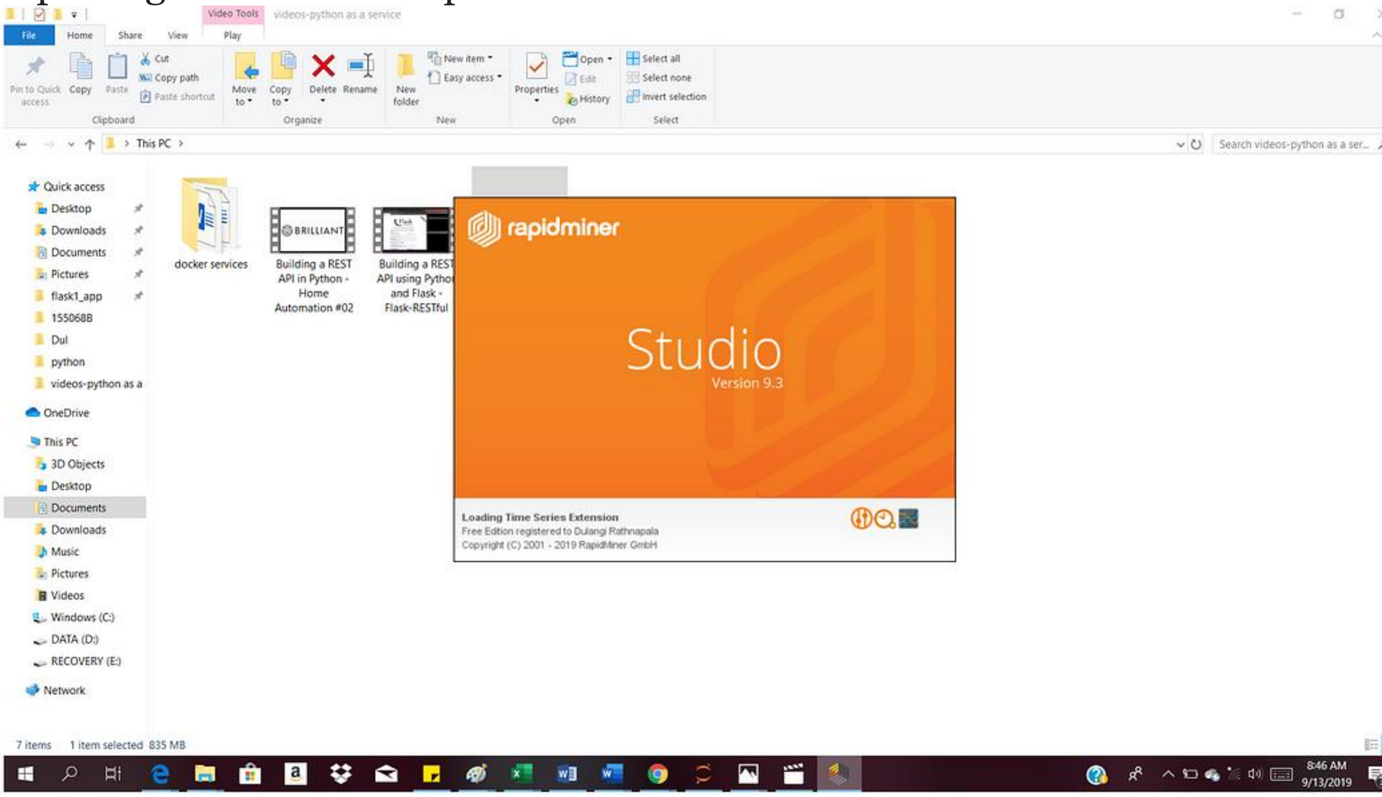
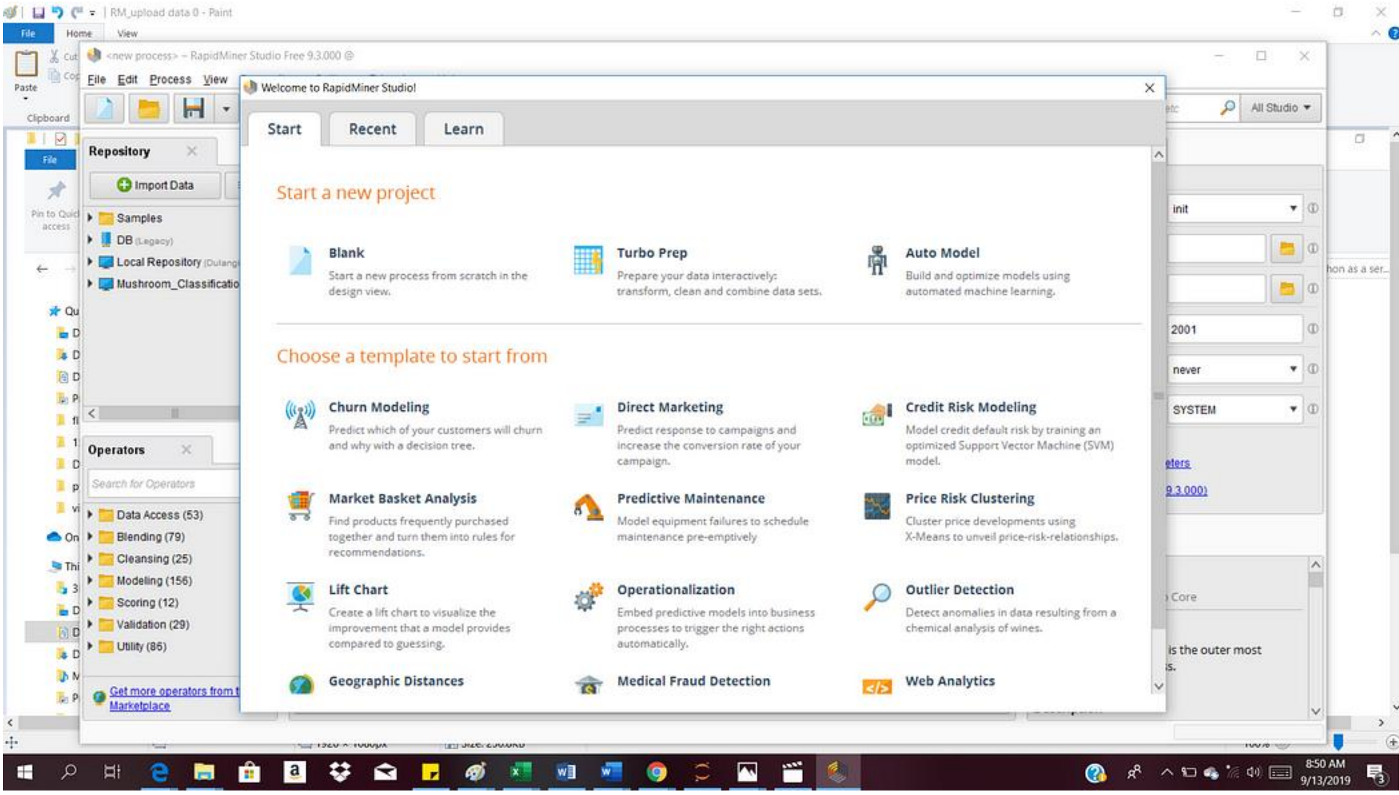
Importing a data set to Rapid Miner



Loading Time Series Extension
Free Edition registered to Dulangi Rathnapala
Copyright (C) 2001 - 2019 RapidMiner GmbH

When you first download RapidMiner using this url :https://rapidminer.com/get-started/, and opening of RapidMiner .



Start a new project

Blank Start a new process from scratch in the design view.

Turbo Prep Prepare your data interactively: transform, clean and combine data sets.

Auto Model Build and optimize models using automated machine learning.

Choose a template to start from

Churn Modeling Predict which of your customers will churn and why with a decision tree.

Direct Marketing Predict response to campaigns and increase the conversion rate of your campaign.

Credit Risk Modeling Predict credit default risk by training an optimized Support Vector Machine (SVM) model.

Market Basket Analysis Find products frequently purchased together and turn them into rules for recommendations.

Predictive Maintenance Model equipment failures to schedule maintenance pre-emptively.

Price Risk Clustering Cluster price developments using X-Means to unveil price-risk-relationships.

Lift Chart Create a lift chart to visualize the improvement that a model provides compared to guessing.

Operationalization Embed predictive models into business processes to trigger the right actions automatically.

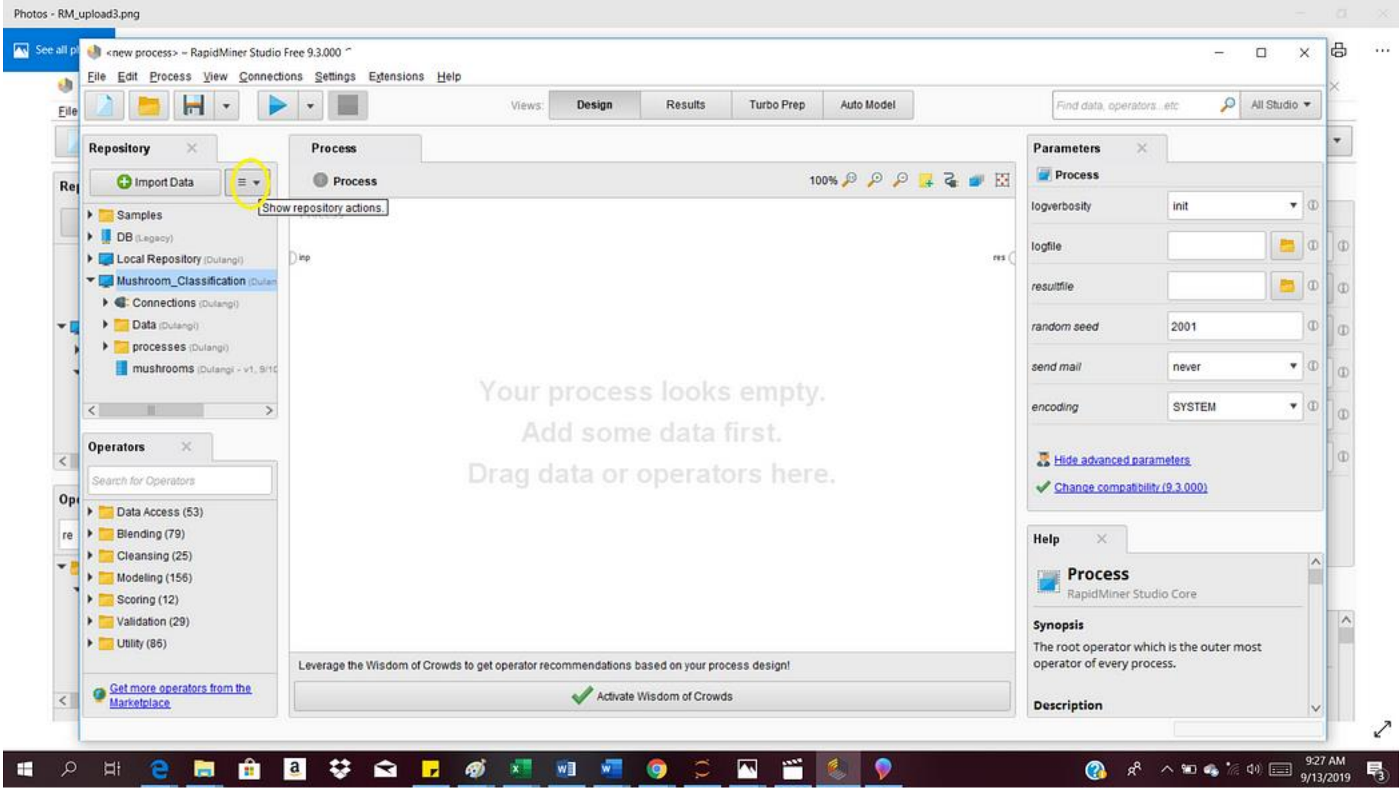
Outlier Detection Detect anomalies in data resulting from a chemical analysis of wines.

Geographic Distances

Medical Fraud Detection

Web Analytics

Now when this page appears, select blank project, to start work on a new project.



Your process looks empty.
Add some data first.
Drag data or operators here.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

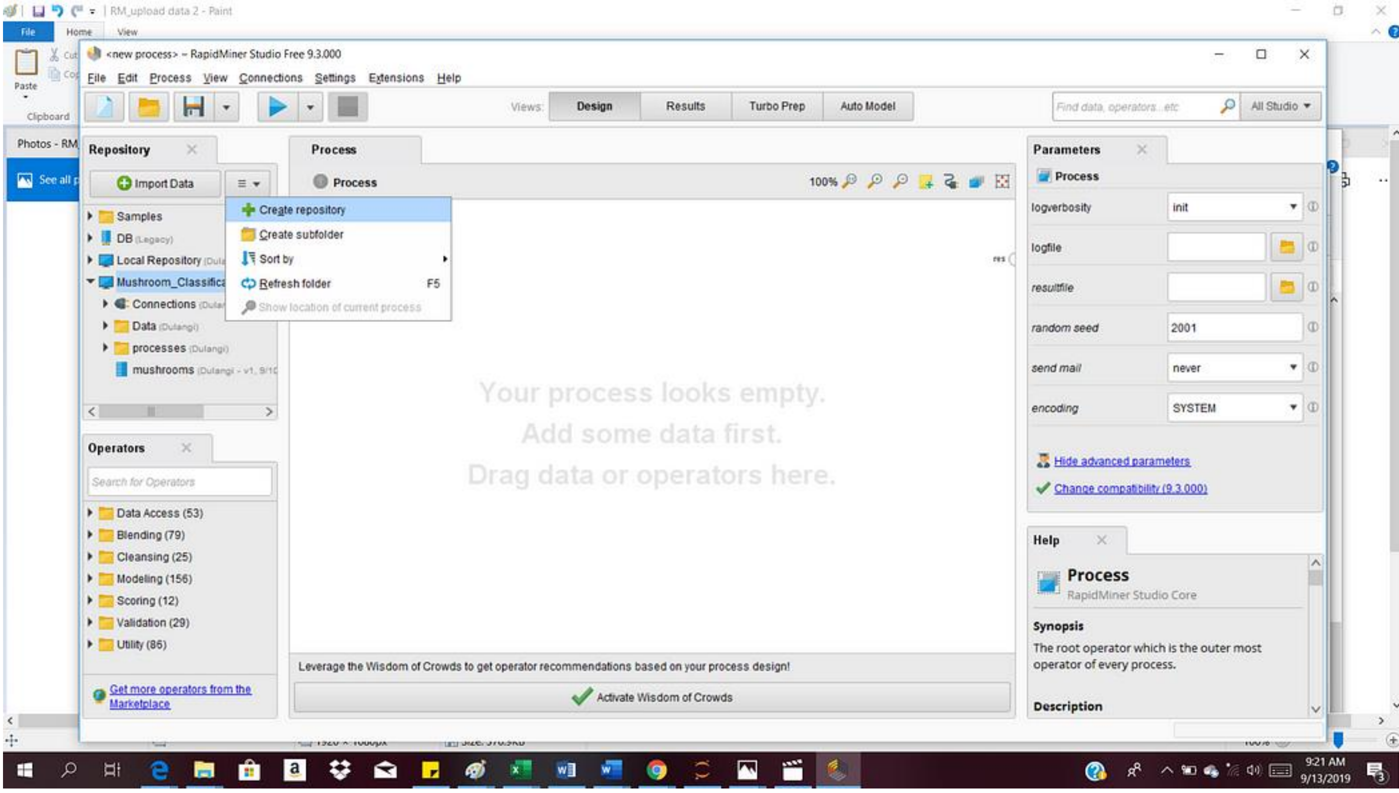
Process	
logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

Synopsis

The root operator which is the outer most operator of every process.

Description

Select the yellow highlighted drop down arrow.



<new process> – RapidMiner Studio Free 9.3.000
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Repository
Import Data
Create repository
Create subfolder
Sort by
Refresh folder F5
Show location of current process
Samples
DB (Legacy)
Local Repository
Mushroom_Classifica
Connections
Data (Dulangi)
processes (Dulangi)
mushrooms (Dulangi - v1, 9/1
Operators
Data Access (53)
Blending (79)
Cleansing (25)
Modeling (156)
Scoring (12)
Validation (29)
Utility (86)
Get more operators from the Marketplace
Process
Your process looks empty.
Add some data first.
Drag data or operators here.
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Hide advanced parameters
Change compatibility (9.3.000)
Help
Process
RapidMiner Studio Core
Synopsis
The root operator which is the outer most operator of every process.
Description
9:21 AM
9/13/2019

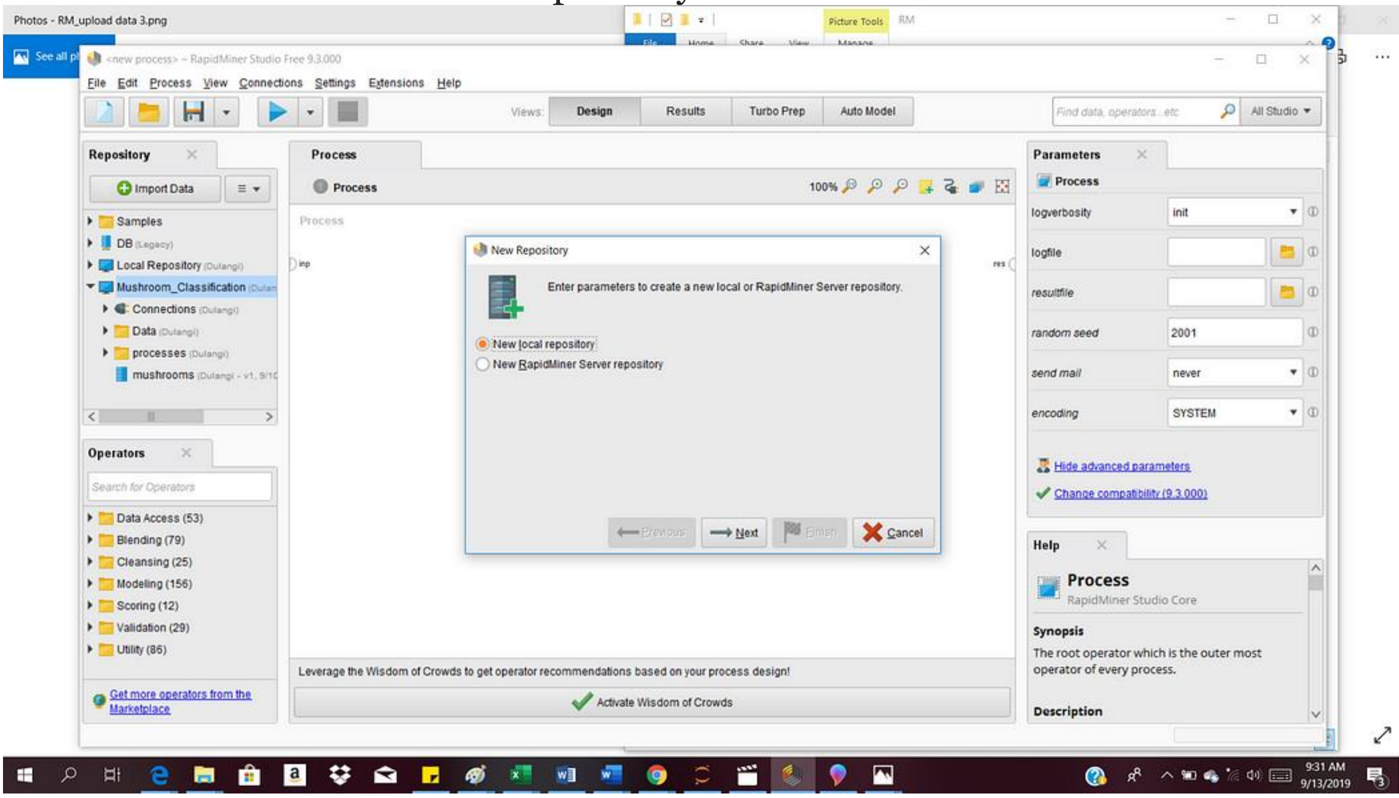
logverbosity: init

random seed: 2001

send mail: never

encoding: SYSTEM

First we have to create a new repository.



Photos - RM_upload data 3.png
<new process> – RapidMiner Studio Free 9.3.000
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Repository
Import Data
Samples
DB (Legacy)
Local Repository (Dulangi)
Mushroom_Classification (Dulan
Connections (Dulangi)
Data (Dulangi)
processes (Dulangi)
mushrooms (Dulangi - v1, 9/1
Operators
Data Access (53)
Blending (79)
Cleansing (25)
Modeling (156)
Scoring (12)
Validation (29)
Utility (86)
Get more operators from the Marketplace
New Repository
Previous
Next
Finish
Cancel
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Hide advanced parameters
Change compatibility (9.3.000)
Help
Process
RapidMiner Studio Core
Synopsis
The root operator which is the outer most operator of every process.
Description
9:31 AM
9/13/2019

Enter parameters to create a new local or RapidMiner Server repository.

● New local repository

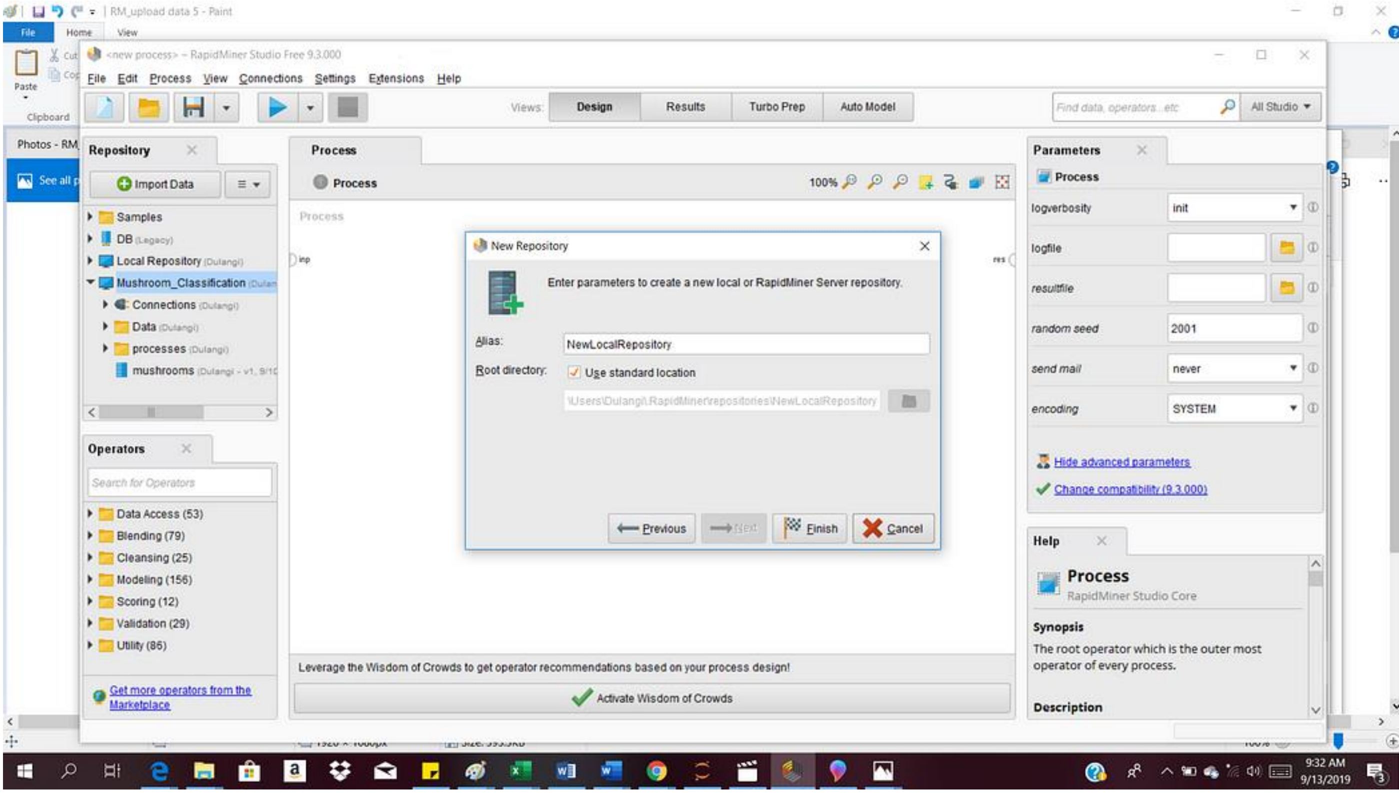
○ New RapidMiner Server repository

logverbosity: init

random seed: 2001

send mail: never

encoding: SYSTEM



New Repository

Enter parameters to create a new local or RapidMiner Server repository.

Alias: NewLocalRepository

Root directory: ☑ Use standard location

\Users\Dulangi\.RapidMiner\repositories\NewLocalRepository

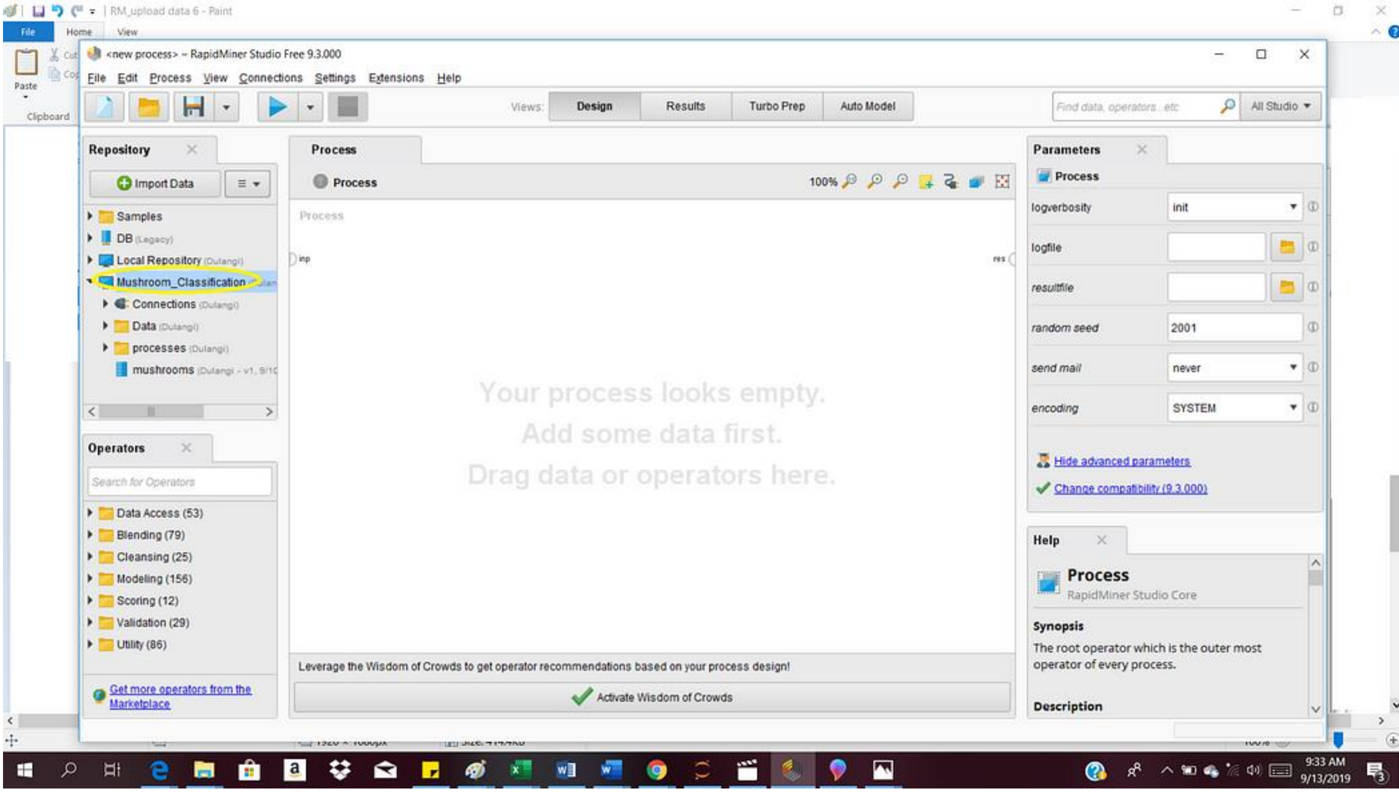
Parameters	
logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Synopsis

The root operator which is the outer most operator of every process.

Give a name to the repository, its advisable give the same name as name of your data set, as you work with many projects later to prevent difficulties to locate which repository.



Your process looks empty.
Add some data first.
Drag data or operators here.

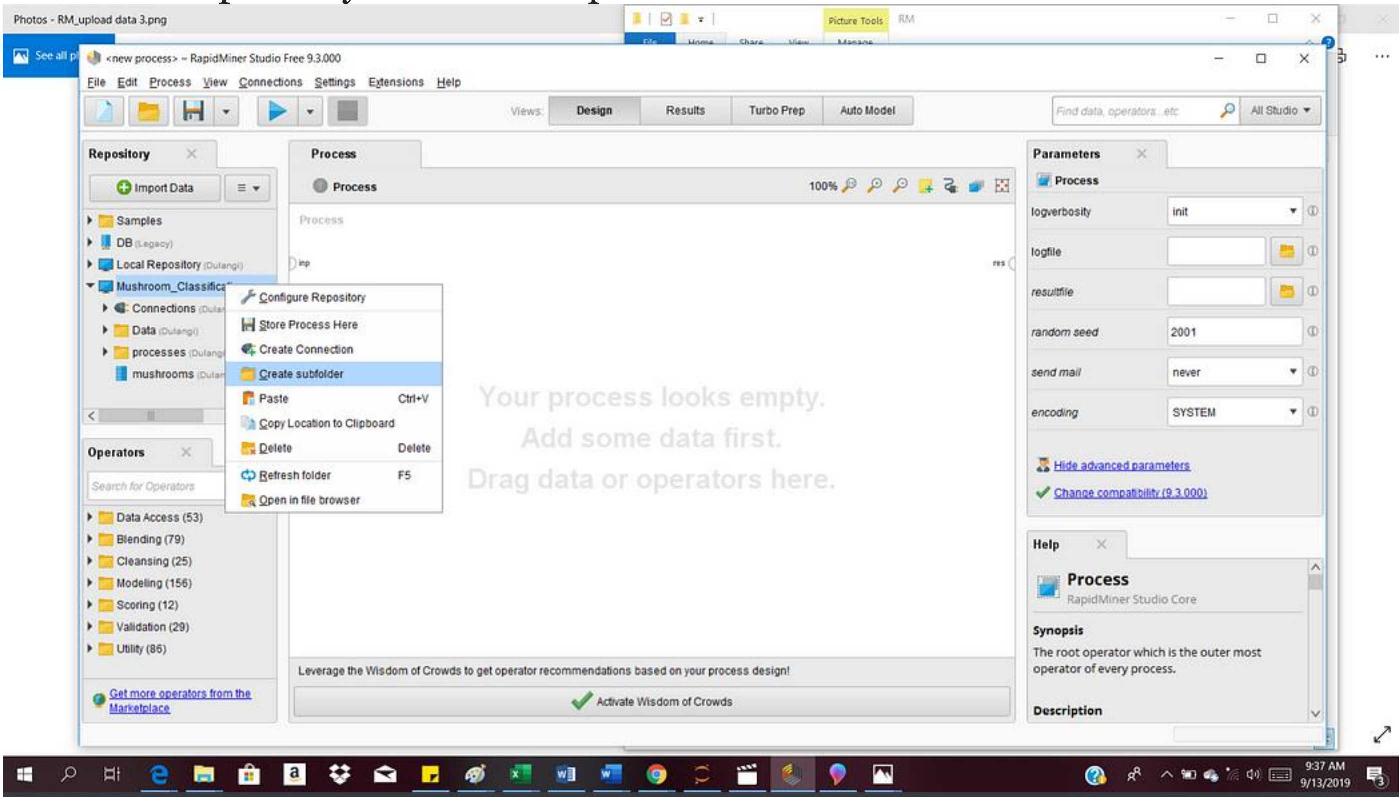
Parameters	
logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Synopsis

The root operator which is the outer most operator of every process.

Now select the repository you created, and create two sub folders under the repository as Data and processes.



Your process looks empty.
Add some data first.
Drag data or operators here.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

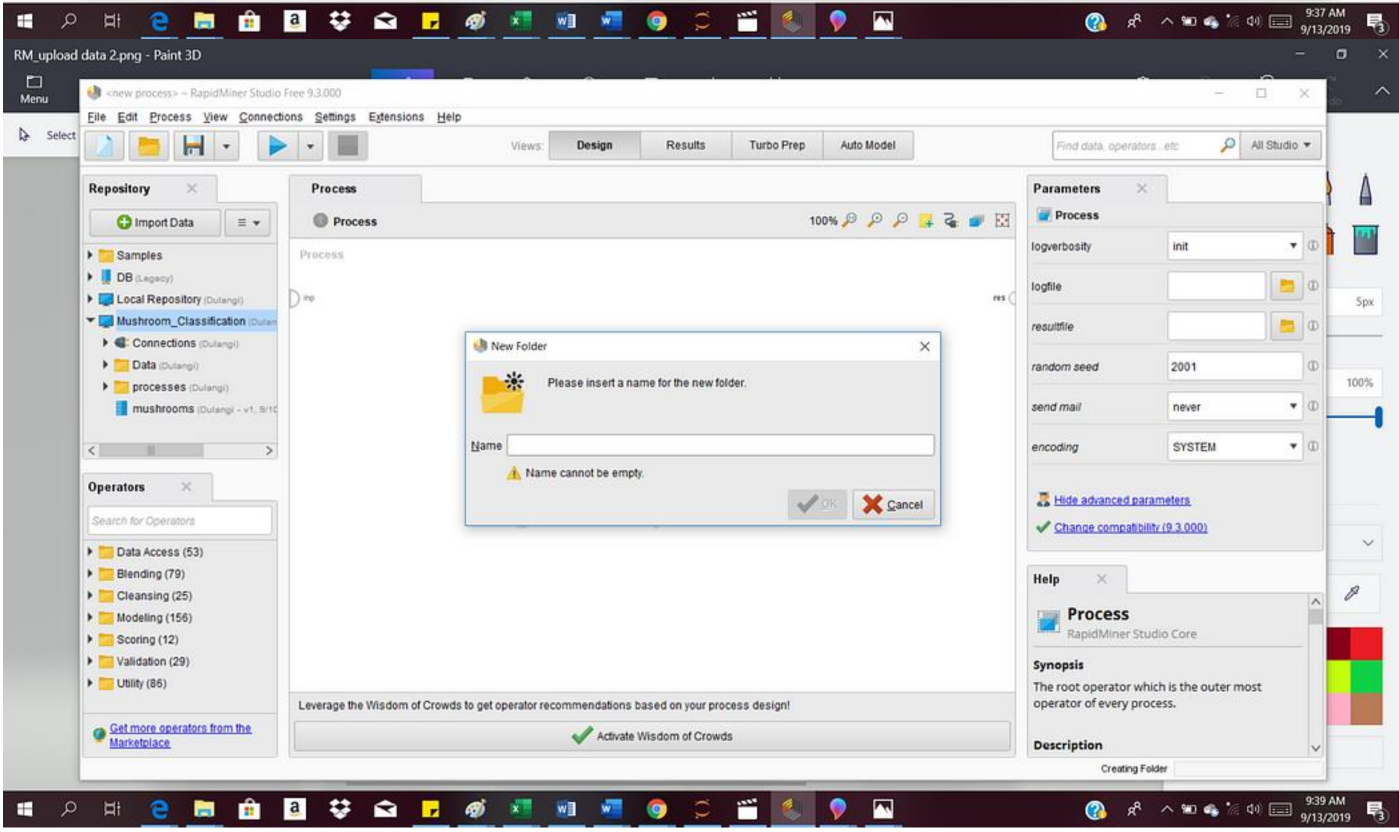
Process

logverbosity: init
logfile:
resultfile:
random seed: 2001
send mail: never
encoding: SYSTEM

Process
RapidMiner Studio Core

Synopsis
The root operator which is the outer most operator of every process.

Description


RM_upload data 2.png - Paint 3D
Menu
Select
<new process> – RapidMiner Studio Free 9.3.000
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Repository
Import Data
Samples
DB (Legacy)
Local Repository (Dulangi)
Mushroom_Classification (Dulan...
Connections (Dulangi)
Data (Dulangi)
processes (Dulangi)
mushrooms (Dulangi - v1, 9/1...
Operators
Search for Operators
Data Access (53)
Blending (79)
Cleansing (25)
Modeling (156)
Scoring (12)
Validation (29)
Utility (86)
Get more operators from the Marketplace
Process
100%
New Folder
OK
Cancel
Activate Wisdom of Crowds
Parameters
Hide advanced parameters
Change compatibility (9.3.000)
Help
Creating Folder
9:39 AM
9/13/2019

Please insert a name for the new folder.

Name:

Name cannot be empty.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Process

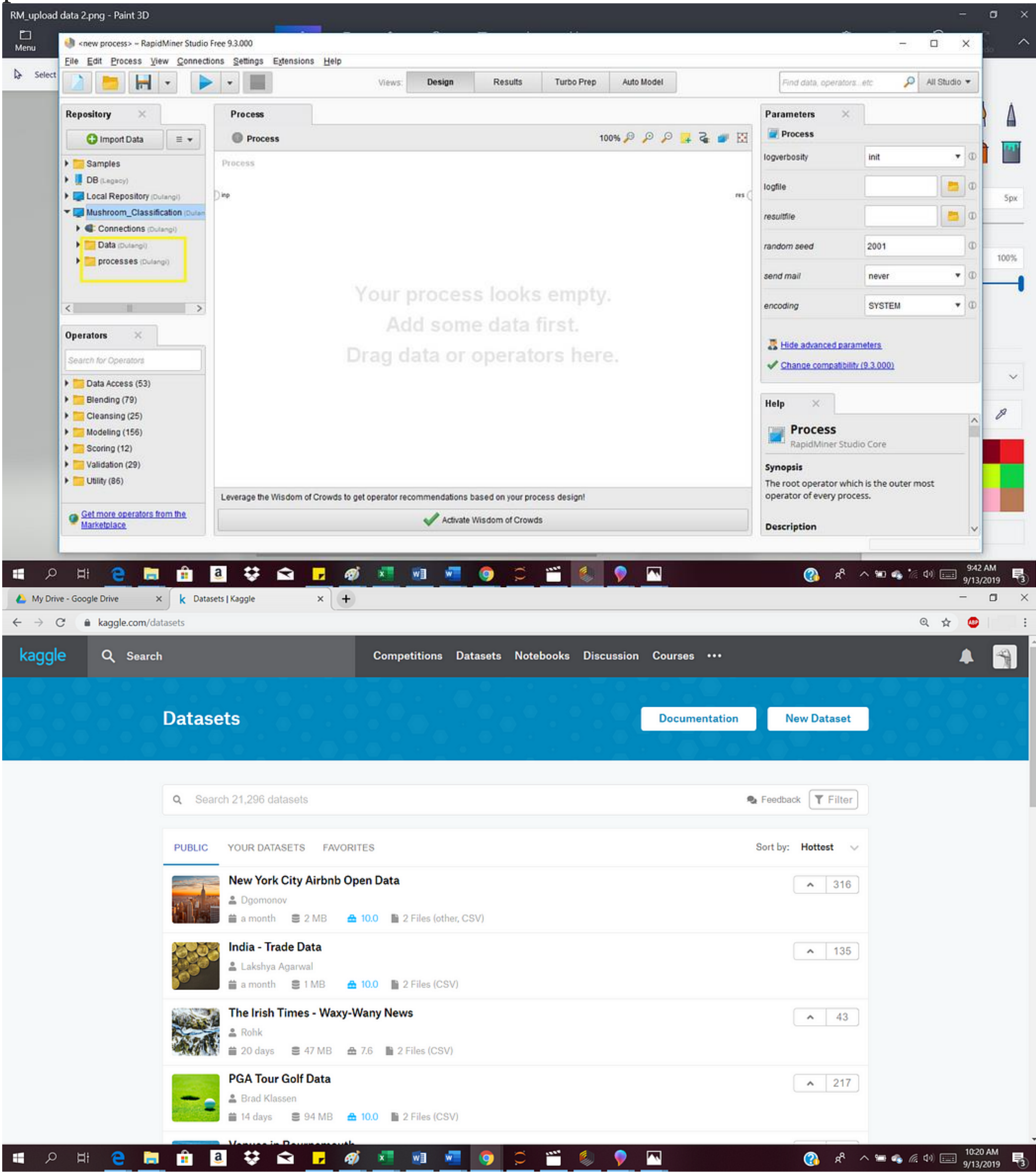
logverbosity: init
logfile:
resultfile:
random seed: 2001
send mail: never
encoding: SYSTEM

Process
RapidMiner Studio Core

Synopsis
The root operator which is the outer most operator of every process.

Description

Type in Data and repeat same method for creating sub folder for process.



Process

logverbosity: init

logfile:

resultfile:

random seed: 2001

send mail: never

encoding: SYSTEM

Your process looks empty.
Add some data first.
Drag data or operators here.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Process
RapidMiner Studio Core

Synopsis
The root operator which is the outer most operator of every process.

New York City Airbnb Open Data
Dgomonov
a month 2 MB 10.0 2 Files (other, CSV) 316

India - Trade Data
Lakshya Agarwal
a month 1 MB 10.0 2 Files (CSV) 135

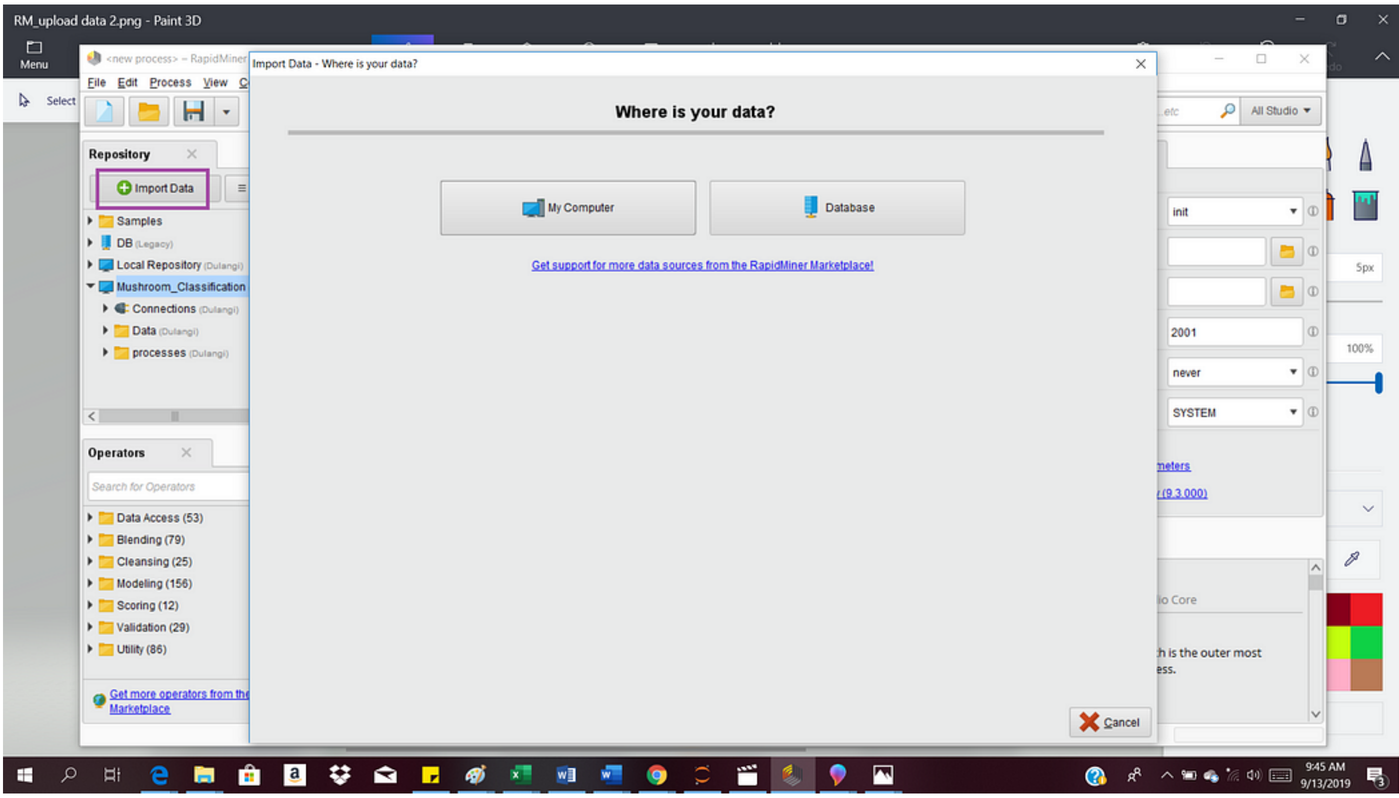
The Irish Times - Waxy-Wany News
Rohk
20 days 47 MB 7.6 2 Files (CSV) 43

PGA Tour Golf Data
Brad Klassen
14 days 94 MB 10.0 2 Files (CSV) 217

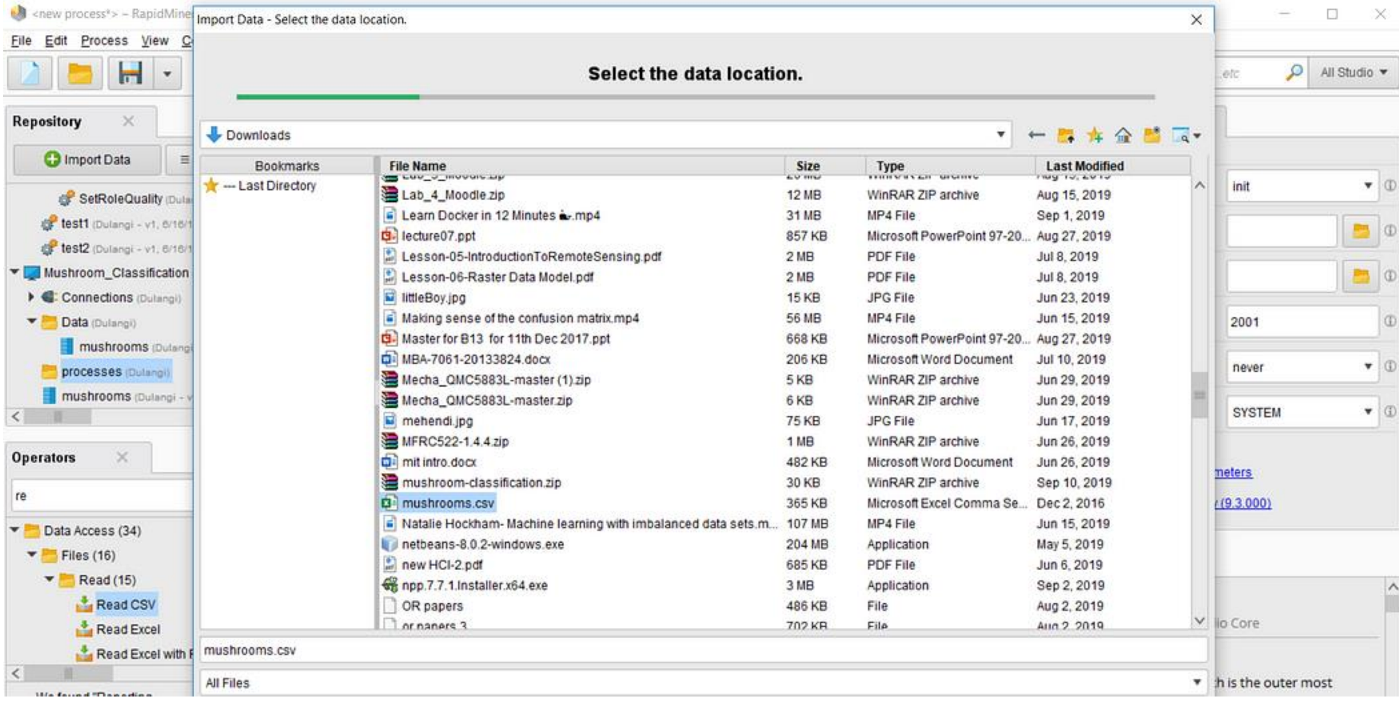
Get a data set for free from Kaggle. A good data set has at least four columns, and at least 1000 rows. Select an appropriate data set, there are two generally used types of classification as:

1. *For supervised learning, we need to chose a data set with a label class, and algorithms such as : Linear Regression, Decision Tree, Random Forest, Naive Bayes. For example lets say we are studying the factors responsible for heart attack via a data set, and in the data set studied are gender, age, occupation, income levels, cholesterol levels in blood and so on, and also there is a class label, column called has got heart attack, hence we can train using this data, and predict for a number of other people having similar factors the probability of them getting a heart attack.*

2. *In unsupervised learning, we don't have a class label for the training data set, methods such as clustering is used. For example lets say the amount of students enrolling for a particular course module, there is no classifier in that as class label.*



Now store first select the purple highlighted box, Import Data, and normally the data downloaded from kaggle will be in Downloads as a zip file. Hence select My Computer. Extract files to download.

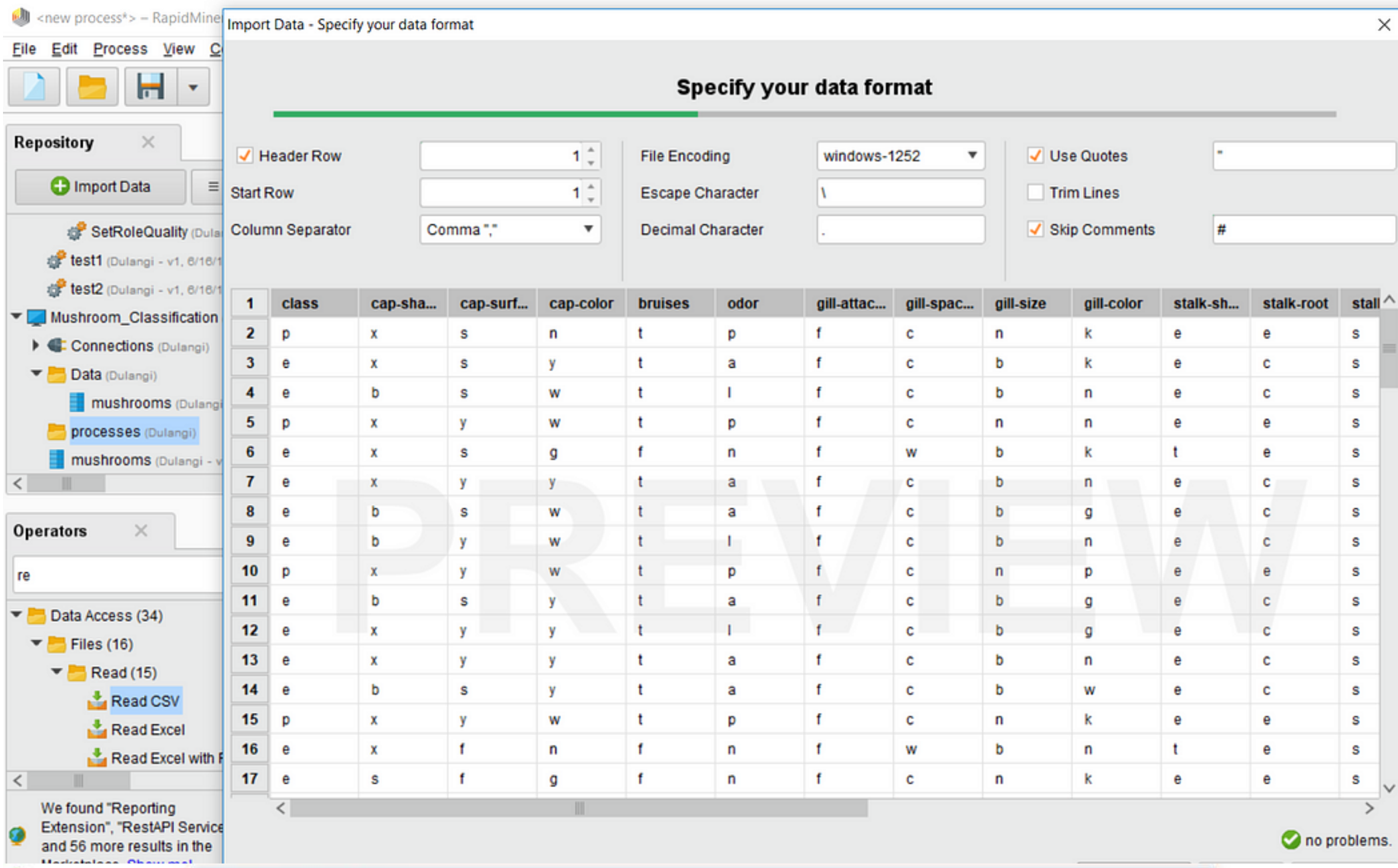


File Name	Size	Type	Last Modified
Lab_4_Moodle.zip	12 MB	WinRAR ZIP archive	Aug 15, 2019
Learn Docker in 12 Minutes 🐳.mp4	31 MB	MP4 File	Sep 1, 2019
lecture07.ppt	857 KB	Microsoft PowerPoint 97-20...	Aug 27, 2019
Lesson-05-IntroductionToRemoteSensing.pdf	2 MB	PDF File	Jul 8, 2019
Lesson-06-Raster Data Model.pdf	2 MB	PDF File	Jul 8, 2019
littleBoy.jpg	15 KB	JPG File	Jun 23, 2019
Making sense of the confusion matrix.mp4	56 MB	MP4 File	Jun 15, 2019
Master for B13 for 11th Dec 2017.ppt	668 KB	Microsoft PowerPoint 97-20...	Aug 27, 2019
MBA-7061-20133824.docx	206 KB	Microsoft Word Document	Jul 10, 2019
Mecha_QMC5883L-master (1).zip	5 KB	WinRAR ZIP archive	Jun 29, 2019
Mecha_QMC5883L-master.zip	6 KB	WinRAR ZIP archive	Jun 29, 2019
mehendi.jpg	75 KB	JPG File	Jun 17, 2019
MFRC522-1.4.4.zip	1 MB	WinRAR ZIP archive	Jun 26, 2019
mit intro.docx	482 KB	Microsoft Word Document	Jun 26, 2019
mushroom-classification.zip	30 KB	WinRAR ZIP archive	Sep 10, 2019
mushrooms.csv	365 KB	Microsoft Excel Comma Se...	Dec 2, 2016
Natalie Hockham- Machine learning with imbalanced data sets.m...	107 MB	MP4 File	Jun 15, 2019
netbeans-8.0.2-windows.exe	204 MB	Application	May 5, 2019
new HCI-2.pdf	685 KB	PDF File	Jun 6, 2019
npp.7.7.1.Installer.x64.exe	3 MB	Application	Sep 2, 2019
OR papers	486 KB	File	Aug 2, 2019

mushrooms.csv

All Files

Select the mushroom.csv file.



We found "Reporting Extension", "RestAPI Service and 56 more results in the

Import Data - Specify your data format

Specify your data format

☑ Header Row: 1

Start Row: 1

Column Separator: Comma ","

File Encoding: windows-1252

Escape Character: \

Decimal Character: .

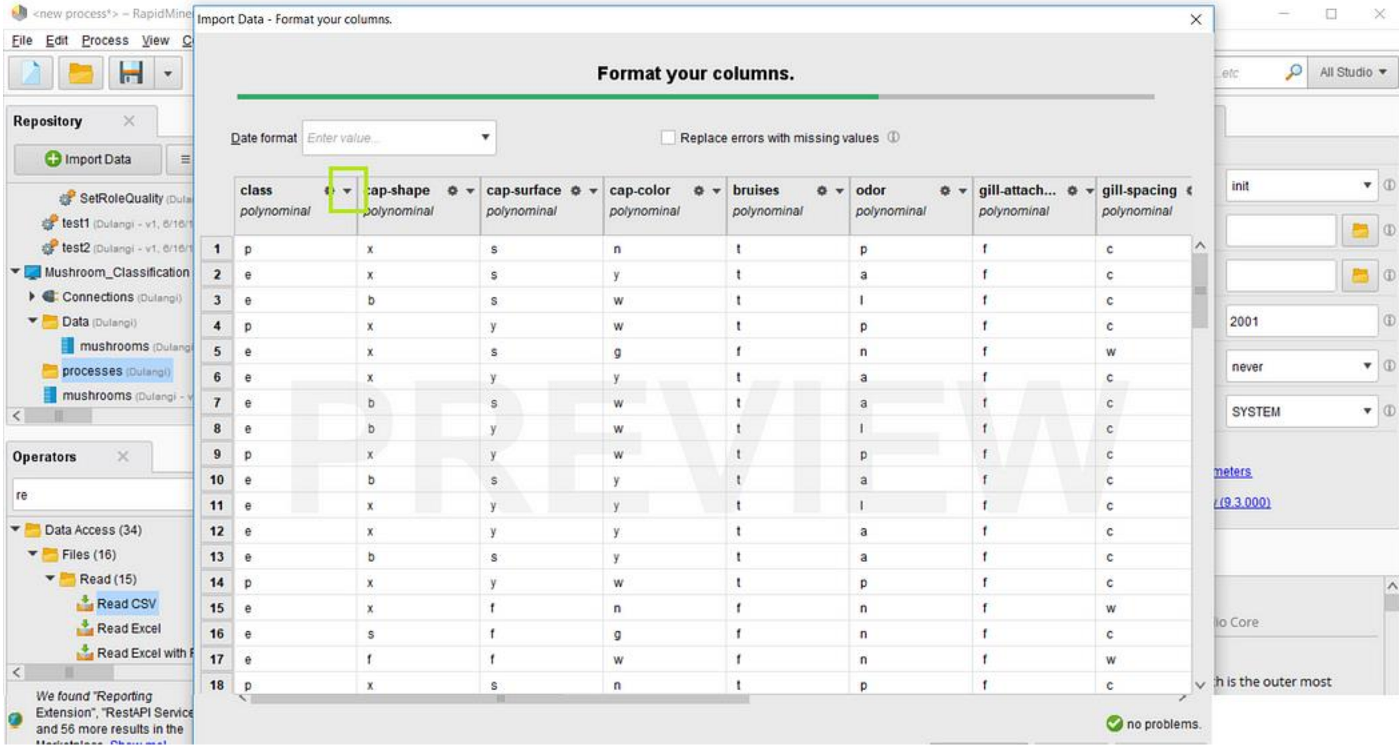
☑ Use Quotes: "

☐ Trim Lines

☑ Skip Comments: #

1	class	cap-sha...	cap-surf...	cap-color	bruises	odor	gill-attac...	gill-spac...	gill-size	gill-color	stalk-sh...	stalk-root	stall
2	p	x	s	n	t	p	f	c	n	k	e	e	s
3	e	x	s	y	t	a	f	c	b	k	e	c	s
4	e	b	s	w	t	l	f	c	b	n	e	c	s
5	p	x	y	w	t	p	f	c	n	n	e	e	s
6	e	x	s	g	f	n	f	w	b	k	t	e	s
7	e	x	y	y	t	a	f	c	b	n	e	c	s
8	e	b	s	w	t	a	f	c	b	g	e	c	s
9	e	b	y	w	t	l	f	c	b	n	e	c	s
10	p	x	y	w	t	p	f	c	n	p	e	e	s
11	e	b	s	y	t	a	f	c	b	g	e	c	s
12	e	x	y	y	t	l	f	c	b	g	e	c	s
13	e	x	y	y	t	a	f	c	b	n	e	c	s
14	e	b	s	y	t	a	f	c	b	w	e	c	s
15	p	x	y	w	t	p	f	c	n	k	e	e	s
16	e	x	f	n	f	n	f	w	b	n	t	e	s
17	e	s	f	g	f	n	f	c	n	k	e	e	s

✓ no problems.



We found "Reporting Extension", "RestAPI Service and 56 more results in the

Import Data - Format your columns.

Format your columns.

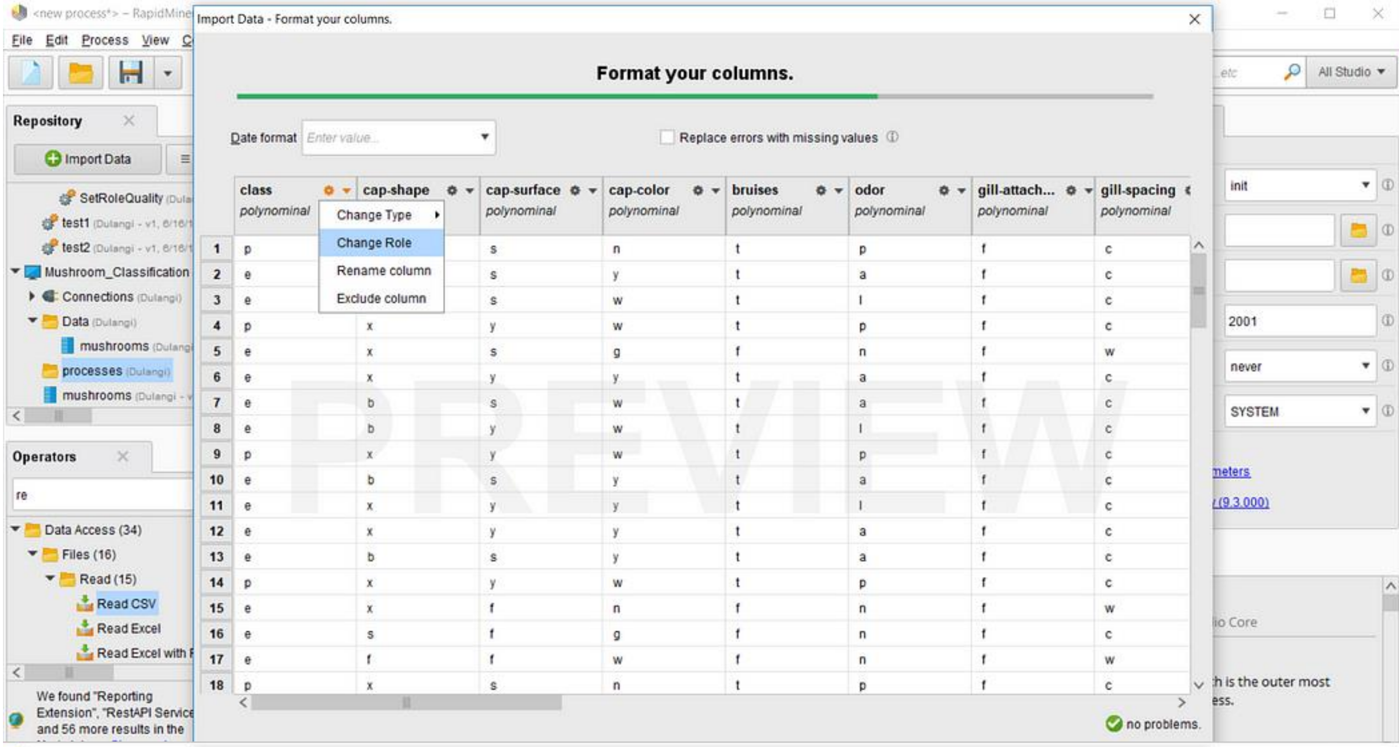
Date format: Enter value...

☐ Replace errors with missing values

	class polynominal	cap-shape polynominal	cap-surface polynominal	cap-color polynominal	bruises polynominal	odor polynominal	gill-attach... polynominal	gill-spacing polynominal
1	p	x	s	n	t	p	f	c
2	e	x	s	y	t	a	f	c
3	e	b	s	w	t	l	f	c
4	p	x	y	w	t	p	f	c
5	e	x	s	g	f	n	f	w
6	e	x	y	y	t	a	f	c
7	e	b	s	w	t	a	f	c
8	e	b	y	w	t	l	f	c
9	p	x	y	w	t	p	f	c
10	e	b	s	y	t	a	f	c
11	e	x	y	y	t	l	f	c
12	e	x	y	y	t	a	f	c
13	e	b	s	y	t	a	f	c
14	p	x	y	w	t	p	f	c
15	e	x	f	n	f	n	f	w
16	e	s	f	g	f	n	f	c
17	e	f	f	w	f	n	f	w
18	p	x	s	n	t	p	f	c

✓ no problems.

If the project you are working on is a supervised classification, then define class label by selecting the green highlighted drop down arrow.



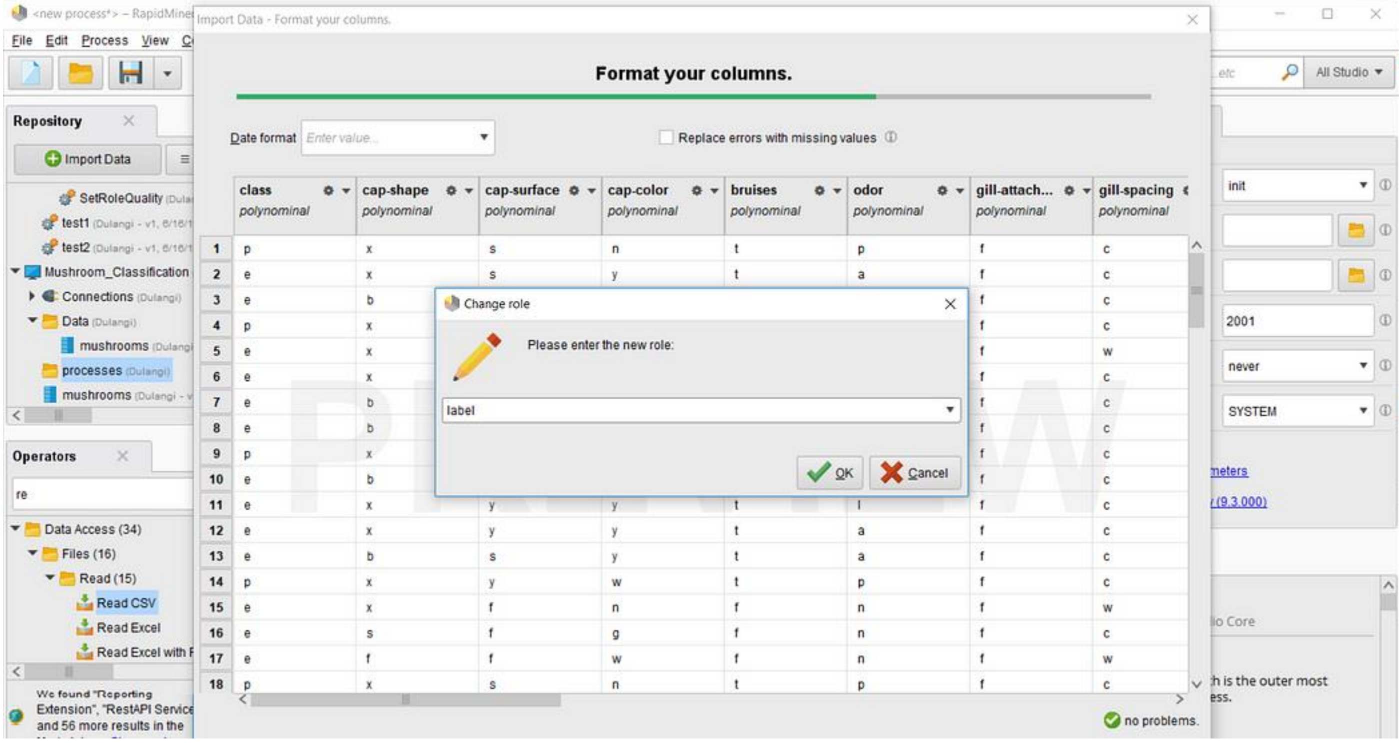
Format your columns.

Date format Enter value...

□ Replace errors with missing values

	class polynominal	cap-shape polynominal	cap-surface polynominal	cap-color polynominal	bruises polynominal	odor polynominal	gill-attach... polynominal	gill-spacing polynominal
1	p		s	n	t	p	f	c
2	e		s	y	t	a	f	c
3	e		s	w	t	l	f	c
4	p	x	y	w	t	p	f	c
5	e	x	s	g	f	n	f	w
6	e	x	y	y	t	a	f	c
7	e	b	s	w	t	a	f	c
8	e	b	y	w	t	l	f	c
9	p	x	y	w	t	p	f	c
10	e	b	s	y	t	a	f	c
11	e	x	y	y	t	l	f	c
12	e	x	y	y	t	a	f	c
13	e	b	s	y	t	a	f	c
14	p	x	y	w	t	p	f	c
15	e	x	f	n	f	n	f	w
16	e	s	f	g	f	n	f	c
17	e	f	f	w	f	n	f	w
18	p	x	s	n	t	p	f	c

✓ no problems.



Format your columns.

Date format Enter value...

□ Replace errors with missing values

Change role

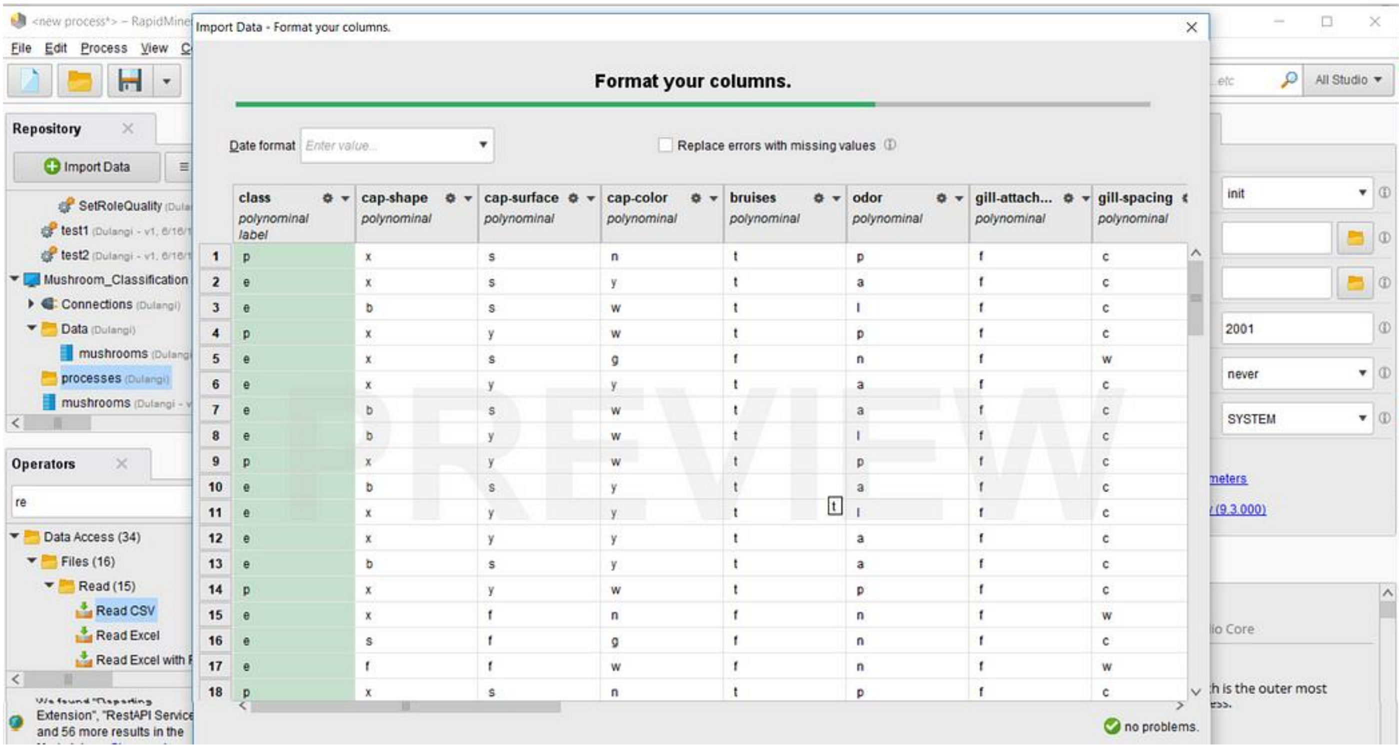
Please enter the new role:

label

	class polynominal	cap-shape polynominal	cap-surface polynominal	cap-color polynominal	bruises polynominal	odor polynominal	gill-attach... polynominal	gill-spacing polynominal
1	p	x	s	n	t	p	f	c
2	e	x	s	y	t	a	f	c
3	e	b					f	c
4	p	x					f	c
5	e	x					f	w
6	e	x					f	c
7	e	b					f	c
8	e	b					f	c
9	p	x					f	c
10	e	b					f	c
11	e	x	y	y	t	l	f	c
12	e	x	y	y	t	a	f	c
13	e	b	s	y	t	a	f	c
14	p	x	y	w	t	p	f	c
15	e	x	f	n	f	n	f	w
16	e	s	f	g	f	n	f	c
17	e	f	f	w	f	n	f	w
18	p	x	s	n	t	p	f	c

✓ no problems.

Chose label from drop down menu.



Import Data - Format your columns.

Format your columns.

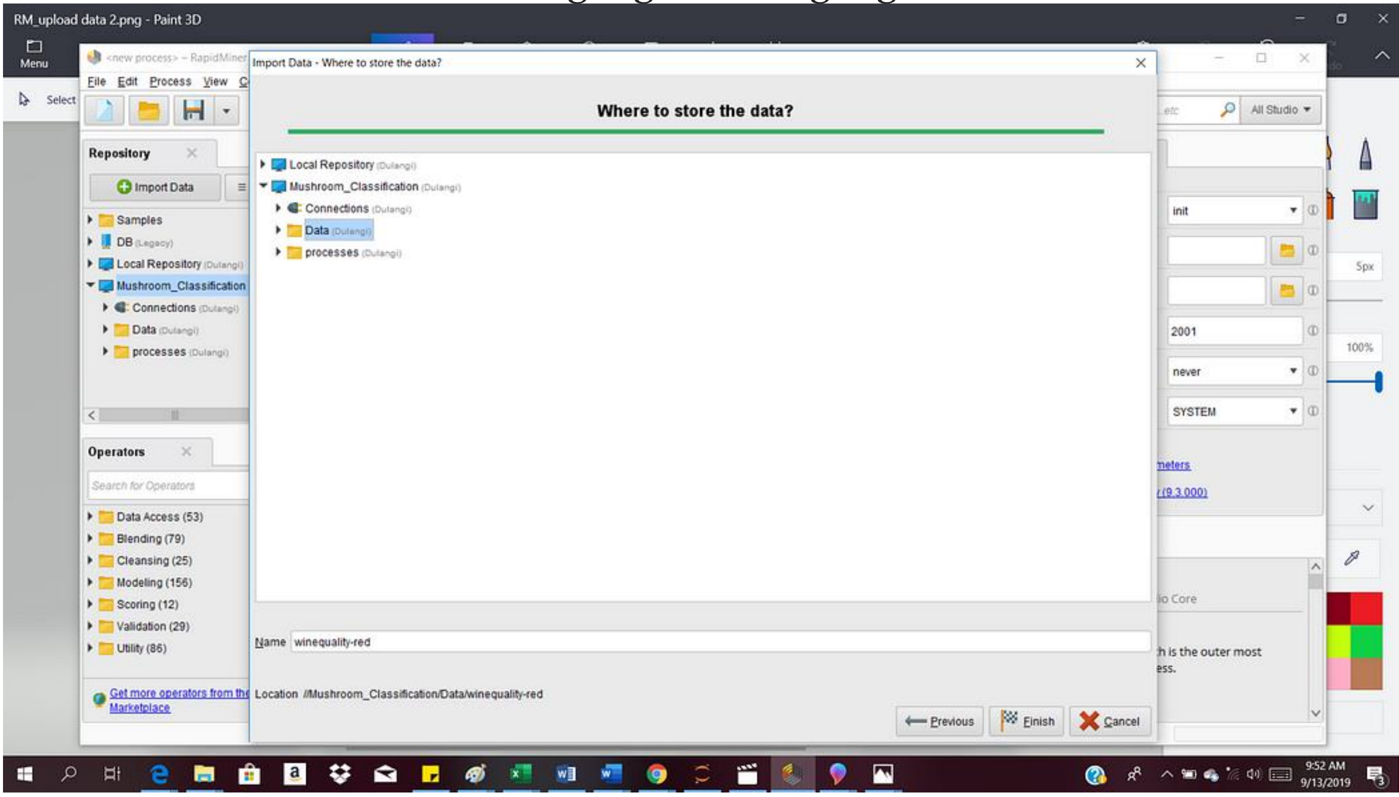
Date format: Enter value...

☐ Replace errors with missing values

	class polynominal label	cap-shape polynominal	cap-surface polynominal	cap-color polynominal	bruises polynominal	odor polynominal	gill-attach... polynominal	gill-spacing polynominal
1	p	x	s	n	t	p	f	c
2	e	x	s	y	t	a	f	c
3	e	b	s	w	t	l	f	c
4	p	x	y	w	t	p	f	c
5	e	x	s	g	f	n	f	w
6	e	x	y	y	t	a	f	c
7	e	b	s	w	t	a	f	c
8	e	b	y	w	t	l	f	c
9	p	x	y	w	t	p	f	c
10	e	b	s	y	t	a	f	c
11	e	x	y	y	t	l	f	c
12	e	x	y	y	t	a	f	c
13	e	b	s	y	t	a	f	c
14	p	x	y	w	t	p	f	c
15	e	x	f	n	f	n	f	w
16	e	s	f	g	f	n	f	c
17	e	f	f	w	f	n	f	w
18	p	x	s	n	t	p	f	c

no problems.

The class label column will be highlighted in light green as shown.



Import Data - Where to store the data?

Where to store the data?

- Local Repository (Dulangi)
- Mushroom_Classification (Dulangi)
  - Connections (Dulangi)
  - Data (Dulangi)
  - processes (Dulangi)

Name: winequality-red

Location: //Mushroom_Classification/Data/winequality-red

Now save the mushroom.csv file in Data sub folder. Do each of changes to the data such as remove duplicates, fill missing values with mean, and store the output data files under Data sub folder. Each of the processes done on data store in process sub folder.