MIS772

Predictive Analytics

Workshop: Setup and Intro

Start working with RapidMiner Studio (RM)

Workshop Plan

Objectives:
The task is to access RapidMiner, prepare the RapidMiner repository, load the data and explore it. Reflect on how to apply learnt knowledge to the assignment problem.

Data Set: *Titanic and student-mat*

Method:
Attend the workshop, follow the tutor's demo and instructions, take notes. Note that the class and online seminar will be recorded and their videos linked to the CloudDeakin topic for later access and study.

1. **Intro**
2. **Setup and explore RapidMiner Studio**
   (a) Install RM locally with an educational license
   (b) Create a RM repository folder
   (c) Start RM
   (d) Configure RM repository
   (e) Explore RM windows and panels
3. **Investigate a dataset in RM**
   (a) Access the Titanic data set
   (b) View the data table, stats and plots
   (c) Investigate and interpret correlations
   (d) Save your RM process
4. **Investigate and reflect**
   (a) Access the Titanic data set
   (b) Explore stats and plots
   (c) Save the RM process to your repository
   (d) *Optional*: Investigate the Read CSV operator to load external data into RapidMiner (in CSV format)
5. **Experiment (at home)**

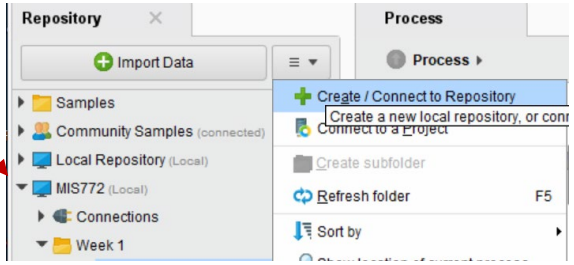
DEAKIN BUSINESS SCHOOL

Getting started

- Download the most recent version of Altair AI Studio (RM) from: https://altair.com/altair-rapidminer-free-trials
- As a Deakin student, you can register for an educational license of Altair AI Studio (RapidMiner) using your Deakin student email address.
- ote: the free trial edition has limited functionality (not suitable for our assignments) so check your edition under elp- About
- **Install the following extensions (their latest versions) from the marketplace:**
  - Text Processing
  - Anomaly Detection
  - Operator Toolbox
  - Series Extension
  - Weka Extension

- Create an RM repository folder to store your workshop projects.
- For each workshop you will be working on a new project, create its folder, e.g. create a folder "**...\MIS772\Week 1**" for this workshop.
- Data files can be downloaded from the unit site.

The created repository

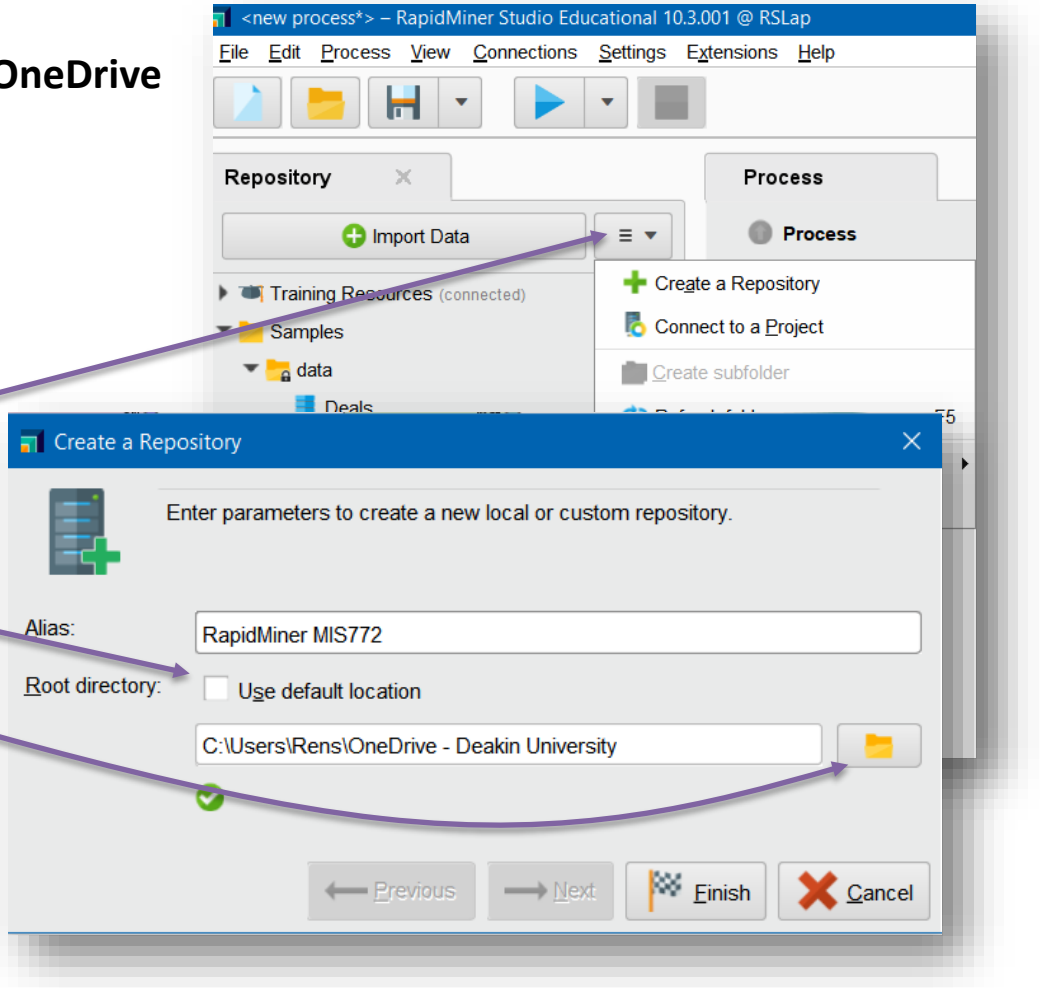


- You are now ready to start using RM Studio.
- Create the first RM process and explore the data.

DEAKIN BUSINESS SCHOOL

Create a RapidMiner repository on your Deakin OneDrive

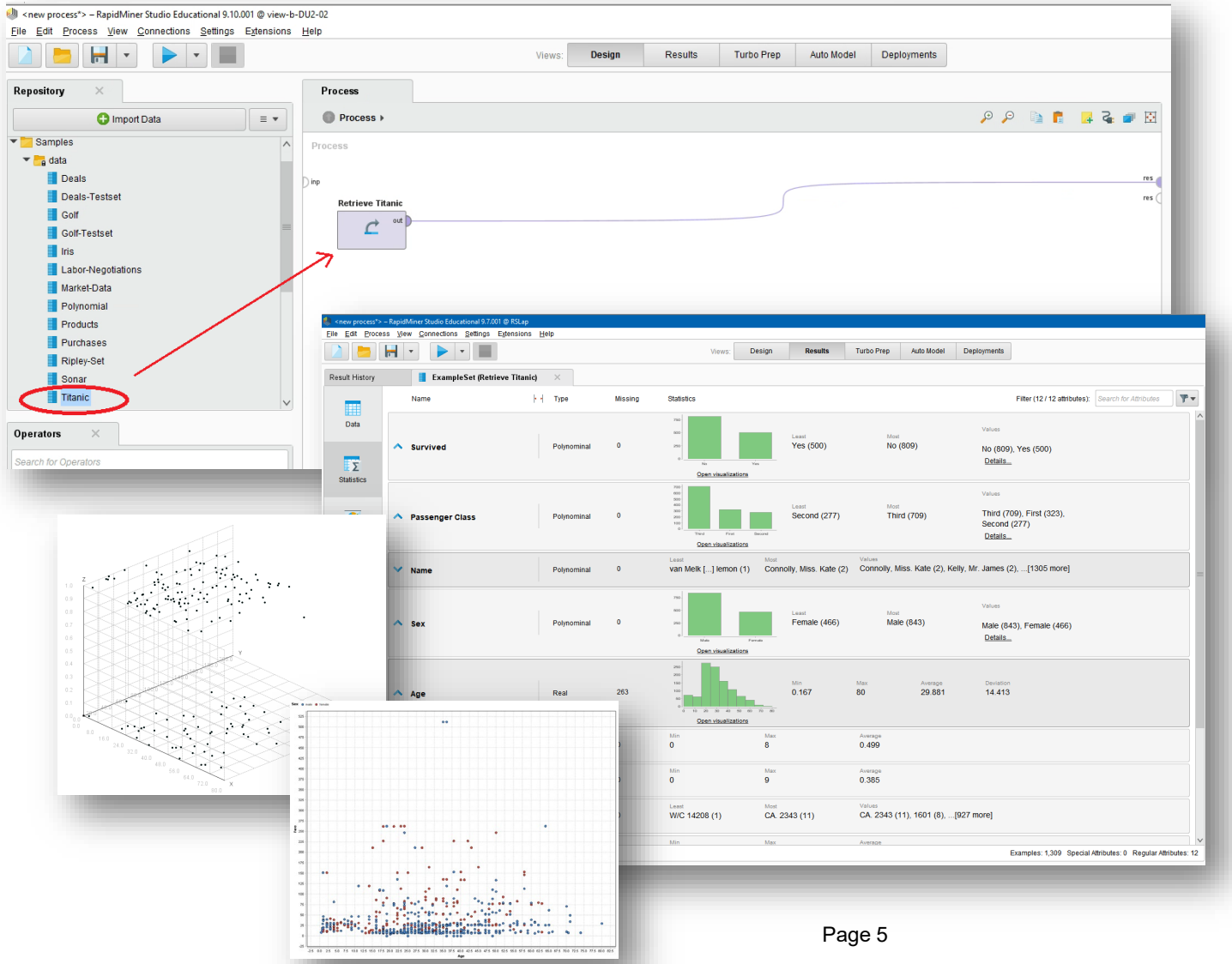
- Ensure you have Deakin OneDrive accessible on your computer. If it is not accessible, search Deakin IT Help for “Using OneDrive on your personal device”
- Start RapidMiner Studio
- On the Repository Panel, click the arrow box, then click “Create a Repository”
- Uncheck “Use default location”
- Navigate to Deakin OneDrive
- Enter an alias for your repository, e.g. “RapidMiner MIS772”
- Click Finish
- **Note: Do NOT save your RapidMiner processes to your local computer. Why?**



DEAKIN BUSINESS SCHOOL

Investigate the first data set in RM

- In RapidMiner, select “Blank Process” to create a new RM process in the “Design” perspective.
- Open the “Samples” data folder, locate the “Titanic” data set and drag it to the blank process.
- In the Design window, connect the ‘out’ port of the operator to the ‘res’ (Results) port.
- Run the process and view the Results.
- Check attribute statistics and distributions.
- Create and describe a scatter plot of Age vs. Passenger Fare and use Sex as the color code.
- Create and describe a 3D scatter plot of Age vs. Cabin vs. Survived on the X, Y, and Z axes.
- Save your process in the folder for this workshop, under your repository.
- Right click your Repository. Note the file location. Using File explorer, navigate to this folder and view the saved process file (e.g. processW1.rmp).
- Optional: Experiment with Store, Sample, Read CSV (for student-mat) and other visualisations.
- Optional: Take screenshots of what you could use in a report.



Challenge: Find attribute correlations

- Open the process you saved in the previous step.
- Find the operator “Correlation Matrix” and add it to the process.
- Connect the output of the data repository or Read CSV operator to the input example set “exa” of the correlation matrix operator “Correlation Matrix”.
- Connect the “mat” output of “Correlation Matrix” operator to the result port “res”.
- Run the process and explore the correlation matrix in the “Results” perspective.
  - How do you interpret the matrix?
  - Why there are some cells with a question mark in them?