Experiencing RapidMiner: Classification

Predictive Analytics for the Uninitiated

Concepts, Decisions and Classification

- What is Predictive Analytics?
- Data mining and model building
- Predictive applications
- Tools and technology
- Analytic process and its design
- What is classification
- Classification models
  - k-NN Nearest Neighbour
  - Decision trees
- Applying predictive model
- Model evaluation
  - Training performance
  - Hold-out validation
  - Cross-validation
- Model optimisation
- Summary and conclusion

Based on notes by Jacob Cybulski
Also some examples and models
are based on the publically available
YouTube videos by ironfrown (Jacob in the free)

DEAKIN UNIVERSITY

Predictive Analytics
Definition - Methods - Terms

Predictive analytics
encompasses techniques that help analysing current and historical facts to make predictions about future or otherwise unknown events
(Wikipedia)

Predictive analytics
relies on methods and techniques from many disciplines, to include:

- **Mathematics**
- **Statistics**
- **Operations research**
- **Information science**
- **Computer science**
- **Artificial intelligence**
- **Data visualisation**
- **Databases**
- **Data warehousing**
- **High performance computing**

Predictive analytics
provides the foundation of methods to build models useful in:

- **explaining the past,**
- **acting in the present, and**
- **predicting the future.**

Some inter-related terms

- **Data science**
- **Data analytics**
- **Text analytics**
- **Data mining**
- **Data wrangling**
- **Pattern recognition**
- **Machine learning**
- **Cognitive computing**
- **Stream analytics**
- **Descriptive analytics**
- **Predictive analytics**
- **Prescriptive analytics**
- **Decision analytics**
- **Business analytics**
- **Statistics**

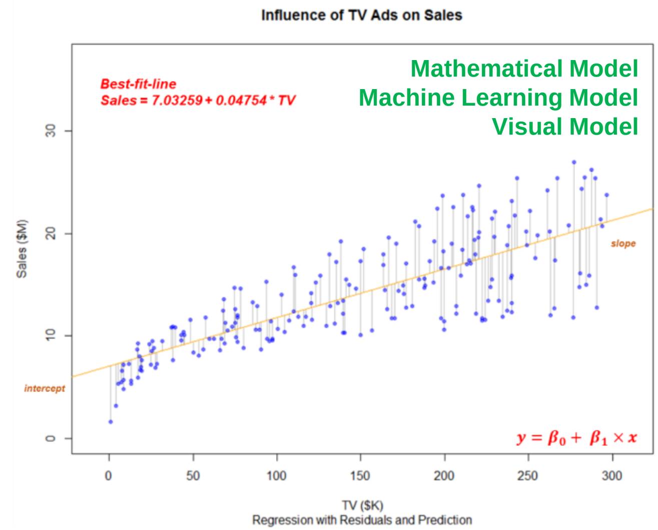
Predictive Analytics Model Building

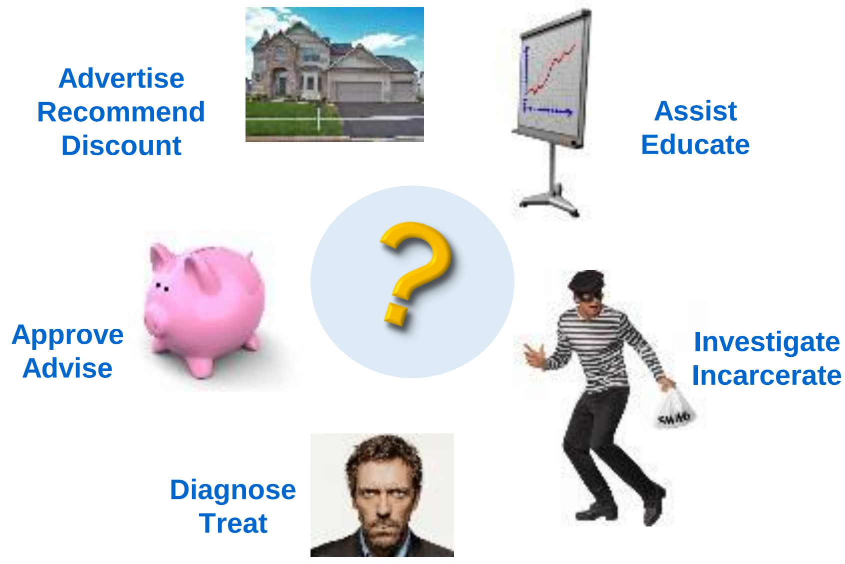
Predictive analytics gives you an analytics process to analyse data over time, leading to more refined outcomes and corrective actions.

The process allows analysts to observe real world entities and then *estimate* their unknown or hidden values, identify their *classification*, and establish their *ranking* or *grouping* in relationship to each other.

Commonly, the same model can be used for *explanation*, *decision support* and *prediction*.

The data sets used in model building are often very large, as it may include data of their own or collected by other organisations, also obtained from open data repositories.





Data → Model

"Who controls the past controls the future - who controls the present controls the past"
(George Orwell, *1984*)

Gartner 2017:
Magic Quadrant for Data Science Platforms

Tools and Approaches

Open Source Tools:

- R / MRO with R Studio
- Python / Anaconda with Spyder
- Orange (for Python)
- WEKA

O/S Deep Learning Tools:

- Tensorflow, Keras, Caffe, CNTK, Torch, Theano, MXNet, H2O.ai

Commercial / Community Tools:

- RapidMiner Studio
- KNIME Analytics Platform

Commercial Tools:

- SAS Enterprise Miner
- IBM SPSS Modeler
- SAP BusinessObjects
- Microsoft Azure ML Studio
- Oracle BI
- Alteryx

Approaches:

- Statistical methods
  - Linear regression
  - Logistic regression
  - General linear models
  - Naïve Bayes models
  - Bayesian modelling
  - Association analysis
  - Time series analysis
- Machine Learning
  - Lazy Learners (k-NN)
  - Decision trees
  - Neural networks
  - Cluster analysis
  - Text mining
  - Support vector machines
  - Anomaly analysis
  - Genetic algorithms
  - Induction and deduction

Classifiers

(Adapted from Wikipedia)

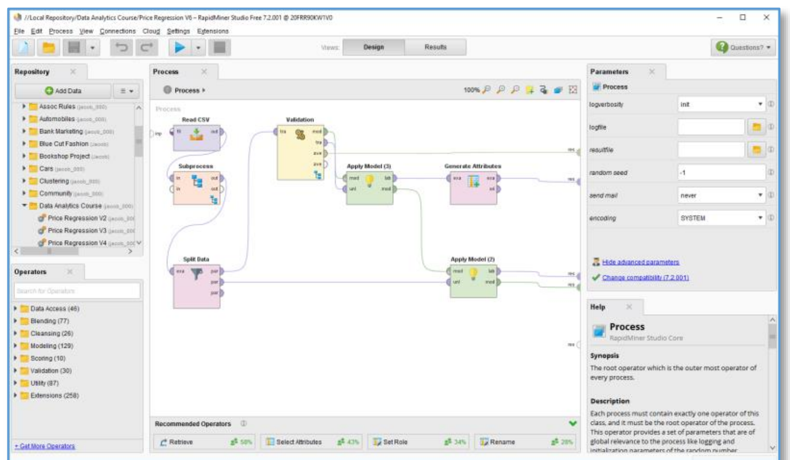
In Case You Wanted to Install RapidMiner Studio

Install RapidMiner

- At Deakin AppsOnDemand, we have RM 6. However, at home install the most up-to-date version of RM Studio 7.xxx
- Install RM Studio 7.xxx from: https://rapidminer.com/
- You need a Laptop or PC running Windows, Mac OS or Linux (e.g. Ubuntu); 8-16Gb RAM; 64 bit OS preferred
- Once installed, for free and unrestricted use of RapidMiner Studio you will need to be registered as "educational"

Some great RapidMiner extensions

1. Run RapidMiner and then select Help > Marketplace, now go to…
2. Updates Tab: install any updates to RapidMiner (e.g. a newer version)
3. Top Downloads Tab: Text Processing, Web Mining, Weka Extensions, Anomaly Detection.
4. Search Tab: type in SOM and install Self-Organising Map; type in Recommender and install Recommender Extension.
5. Restart RapidMiner
6. You are now ready to use RapidMiner Studio to do some serious data mining and data analytics.



In case you wanted to use RapidMiner on your own computer!

Thinking Point on Classification

- (Nearly) everything in the world can be described with a set of unique *classes of labels*, such as:
  - Toyota Prado 2012 in the car yard will be "Sold" today, or
  - James' final mark in MIS171 will be "HD"
- Objects of the same class can be considered *similar*.
- Can we rely on the past observations to *predict classes* (or labels) of objects yet to be observed and events likely to happen in the future?
- The answer is Yes and the method is called *classification*!

Classifying Observations Using Attribute Values

The world and our naive understanding of its complexity dictate the rules!

Data analytics systems help building (quality) decision trees.

Trucks have at least 3 axles and/or are red
Cars have 2 axles and/or are small

Decision Tree

Axles > 2

yes

Truck!

no

Red?

yes

Truck!

I am not a truck!

no

Small?

no

Truck!

yes

Car!

Clearly, it is possible to build a better decision tree or select better variables!

What am I? By my looks!

The world is imperfect!

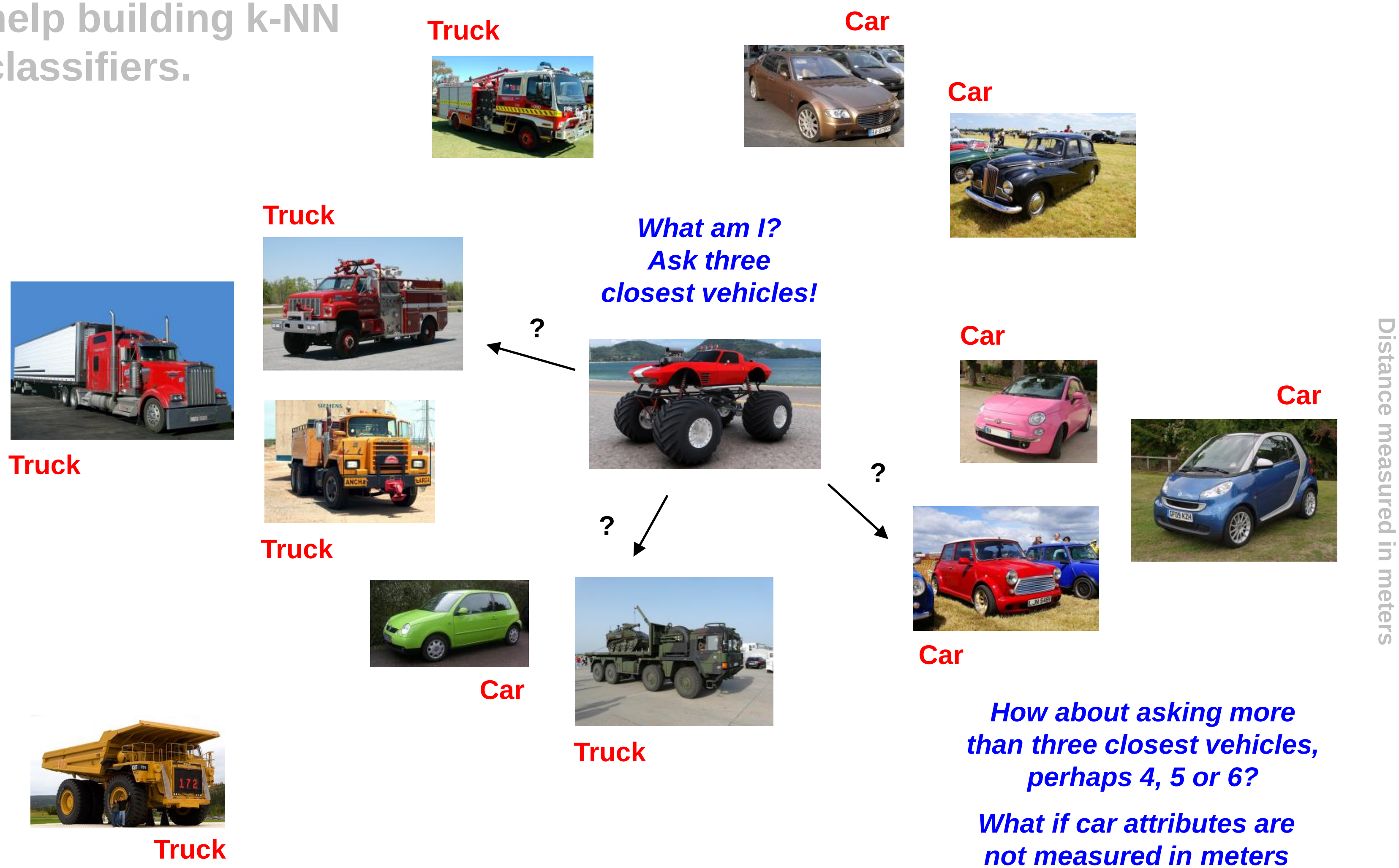
Classifying Observations by Measuring Distance

The world and our naive understanding of its complexity dictate the rules!

Data analytics systems help building k-NN classifiers.

Trucks are parked near trucks and cars are parked near cars

k Nearest Neighbours



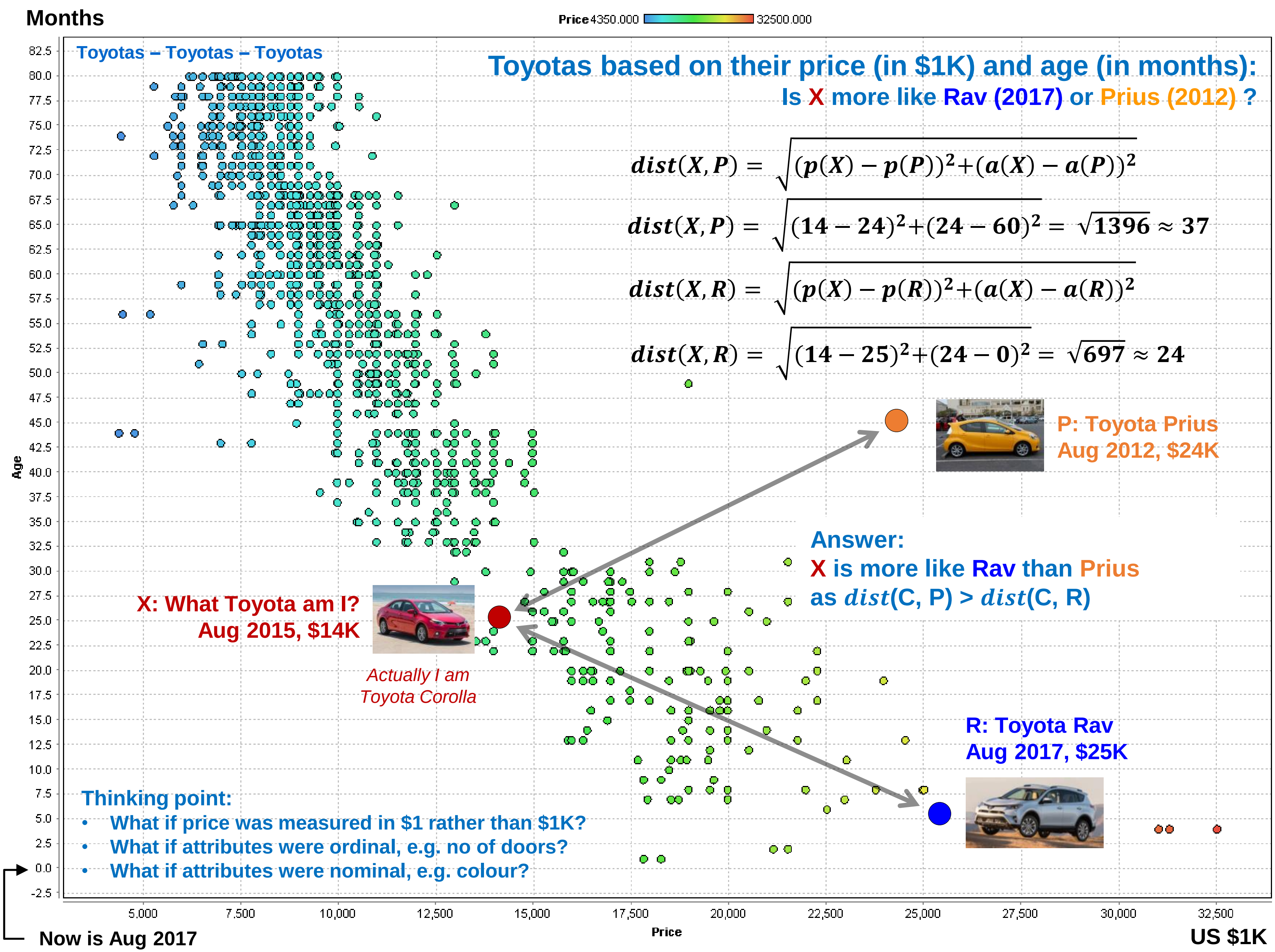
How about asking more than three closest vehicles, perhaps 4, 5 or 6?

What if car attributes are not measured in meters but in their age and price?

The world is never perfect!

Measurement is Key to Analytics, e.g. in k-NN

In this example, data points consist of pairs of attributes, i.e. the car's age and its price. Can we measure "similarity" between based on their "distance"?



$$dist(X,P) = \sqrt{(p(X) - p(P))^2 + (a(X) - a(P))^2}$$

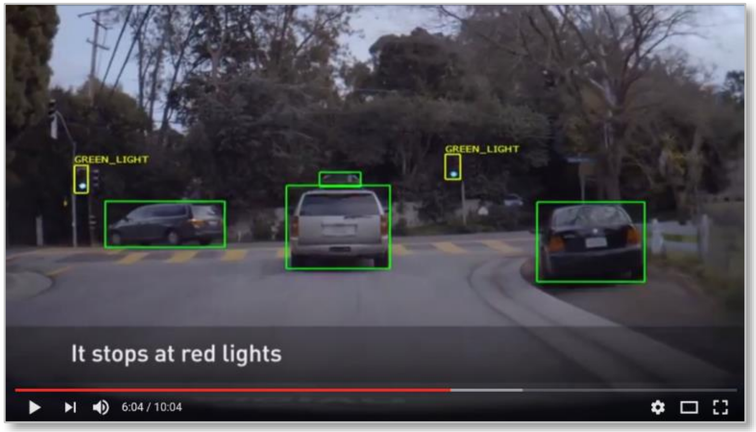
$$dist(X,P) = \sqrt{(14-24)^2 + (24-60)^2} = \sqrt{1396} \approx 37$$

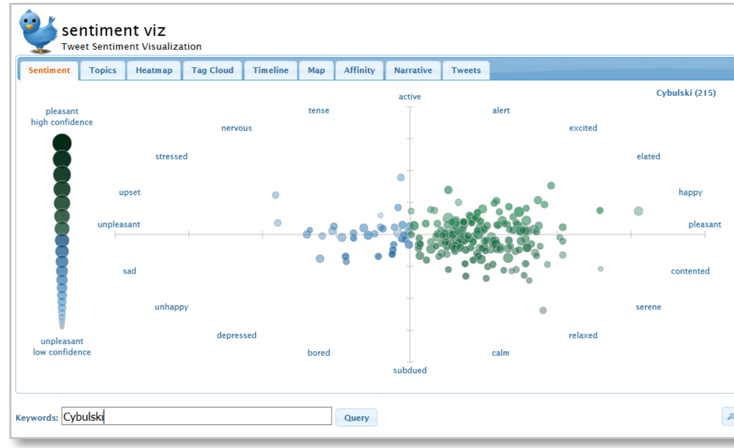
$$dist(X,R) = \sqrt{(p(X) - p(R))^2 + (a(X) - a(R))^2}$$

$$dist(X,R) = \sqrt{(14-25)^2 + (24-0)^2} = \sqrt{697} \approx 24$$

DEAKIN UNIVERSITY

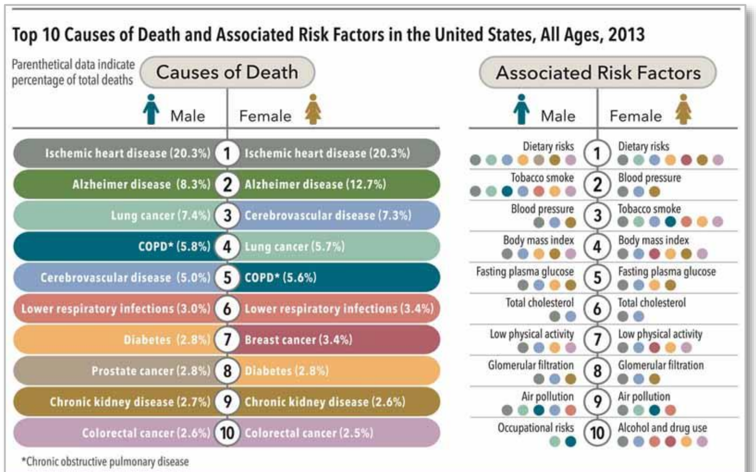
What else can be classified and how analytics helps





NVIDIA self-driving cars:
Should I drive or stop?
https://www.youtube.com/watch?v=MF9NwOTLLgE

Christopher Healey:
Is this tweet positive or negative towards the lecturer?
https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/



USA Institute of Health Metrics and Evaluation:
What are my health risks?
http://www.healthdata.org/infographic/when-and-why-people-die-united-states-1990-2013

IBM Watson – Morgan movie trailer:
Is this movie clip sufficiently scary to be included in a trailer?
https://www.youtube.com/watch?v=gJEzuYynaiw

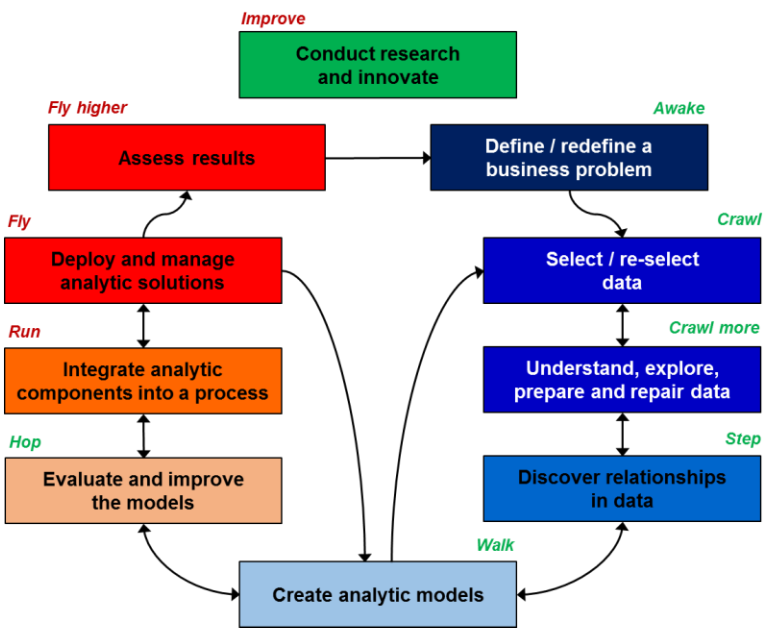
Different kinds of classifications, i.e. to detect obstacles in front of self-driving cars, sense emotion of movie viewers, explain health risks and identify sentiment of Twitter messages. Once we classify past examples, we can then apply such classification to future individuals (predict their class / decision).

Classification and its Process

Applications

Credit approval
Target marketing
Medical diagnosis
Fraud detection
Sentiment analysis

- **Classification**
  - The process of organizing a set of observations (data samples) into classes, each identified by a label (nominal)
- **Model (Classifier)**
  - Existing data is used to create a classification model
  - The model is (usually) simpler than a collection of samples used in its creation
- **Prediction**
  - The model is subsequently used to classify new data, i.e. predict missing or unknown class labels



The Entire Process is More Complex !!!

- **Define a business problem**
- **Select data**
  - Structured and/or unstructured
  - What to predict (label)
  - What are the predictors (attributes)
- **Explore and understand data**
  - Statistics
  - Distribution
  - Relationships
- **Build the model**
- **Evaluate model performance**
  - Training performance
  - Hold-out validation
  - Cross-validation
- **Integrate the model with enterprise systems**
- **Deploy validated model**
  - Use the validated model
  - Predict labelled attribute
  - Account for possible error
- **As the world changes assess the model results and its performance – a new model may be needed!**

Define a Business Problem & Select Data

The management of the legal firm Righteous Compensation Lawyers asked you to develop a computer-assisted method of analysing Worker's Compensation claims, capable of identifying:

Subrogation potential, i.e. possibility of insurance company to recover all its costs due to the fault of the parties involved;

Motor-vehicle injuries, detected in claims that are likely to involve motor-vehicle accidents and which should be processed within a different jurisdiction; and finally,

Fraudulent claims.

The firm provided you with a sample of over 3000 examples of claims described in terms of injured body part, the nature and cause of injury, as well as adjustor notes taken by insurance employees when in contact with the claimants, their employer or representatives. After the lengthy process, each of the claims has been verified and annotated with the flags indicating if the injury involved a vehicle (whether or not stated in the claim), whether it ended in the recovery of all payed entitlements and costs, and whether or not fraudulent claims have been detected and the applicant eventually sued.

Claim Number	Adjustor Notes	Body Part	Nature of Injury	Cause of Injury	Vehicle Flag (...	Subrogation...	Fraud Flag (...
4487308	Strained neck trying to catch falling product.	Neck	Sprain/Strain	Slip/Fall	0	1	0
309831108	Fingers caught in machine.	Finger	Contusion	Caught in Machine	0	0	0
1301185908	Claimant caught left hand between two machine so...	Hand	Laceration	Equipment/Machinery	0	0	0
1716965808	Claimant states that while he and coworker were dri...	Multiple	Contusion	Struck Object	1	1	0
1924817308	Smashed right second finger, was using a drill pres...	Finger	Contusion	Struck Object	0	0	0
2500385808	Claimant alleges that he injured his right knee. Thre...	Knee	Sprain/Strain	Unknown	0	1	0
2525865808	Left ankle pain due to getting in and out of a truck re...	Ankle	Repetitive Motion	Repetitive Motion	1	0	0
2601381908	While trying to avoid hitting a car out of control, came...	Neck	Contusion	MVA	1	1	0
2613478908	Fell in blast freezer, injured back and side.	Back	Contusion	Struck Object	0	0	0
2614936508	Employee was struck by automobile --- contusion to ...	Knee	Contusion	MVA	1	1	0
2701592908	Employee alleges while letting a machine down into...	Shoulder	Contusion	Struck Object	0	1	0
2714742208	Employee failed to yield and was hit by an oncoming...	Multiple	Contusion	MVA	1	1	0
2829016508	Claimant states he was loading a patio door onto a t...	Knee	Sprain/Strain	Lifting	1	0	0

Sample claims records

Classification Modelling to Identify Subrogation Cases

What we have: data from the past, e.g. information about the previously processed claims classified by subrogation flag (Data).

ExampleSet (2126 examples, 1 special attribute, 4 regular attributes)

Row No.	Subrogation...	Body Part	Nature of Inj...	Cause of Inj...	Vehicle Flag ...
1	1	Neck	Sprain/Strain	Slip/Fall	0
2	0	Finger	Contusion	Caught in Ma...	0
3	1	Multiple	[illegible]	[illegible]	[illegible]
4	0	Finger	[illegible]	[illegible]	[illegible]
5	1	Knee	Sprain/Strain	Unknown	0

Training – Create a predictive model able to classify existing data

How are we going to do this: we will use training data to create a model capable of claim classification by subrogation flag (Training).

ExampleSet (911 examples, 4 special attributes, 4 regular attributes)

Row No.	Subrogation...	prediction(S...	confidence(1)	confidence(0)	Body Part	Nature of Inj...	Cause of Inj...	Vehicle Flag ...
1	0	0	0.200	0.800	Hand	Laceration	Equipment/M...	0
2	0	0	0.200	0.800	Ankle	[illegible]	[illegible]	[illegible]
3	1	1	1	0	Knee	[illegible]	[illegible]	[illegible]
4	1	1	1	0	Shoulder	[illegible]	[illegible]	[illegible]
5	0	1	0.600	0.400	Hand	[illegible]	[illegible]	[illegible]

Validation – Classify existing cases and compare against known classification

How would we know if it worked: we will use the validation data not used in training to test accuracy of predictions (Validation).

ExampleSet (1432 examples, 0 special attributes, 6 regular attributes)

Row No.	Claim Number	Adjustor Not...	Body Part	Nature of Inj...	Cause of Inj...	Vehicle Flag ...
1	160122408	Laceration to ...	Finger	Laceration	Struck Object	0
2	360784508	While unloadi...	Back	Sprain/Strain	Lifting	0
3	860564608	Claimant wa...	Hand	[illegible]	[illegible]	[illegible]
4	3060046708	Claimant wa...	Head	[illegible]	[illegible]	[illegible]
5	3160698008	Alleges carpa...	Wrist	Repetitive Mo...	Repetitive Mo...	0

New Cases – Collect new unclassified data, i.e. with missing information

How about future cases: we will collect new cases of insurance claims without subrogation flag (Deployment).

What we want: we will apply the validated model to predict whether or not the claim could end in subrogation (Application).

ExampleSet (1432 examples, 3 special attributes, 6 regular attributes)

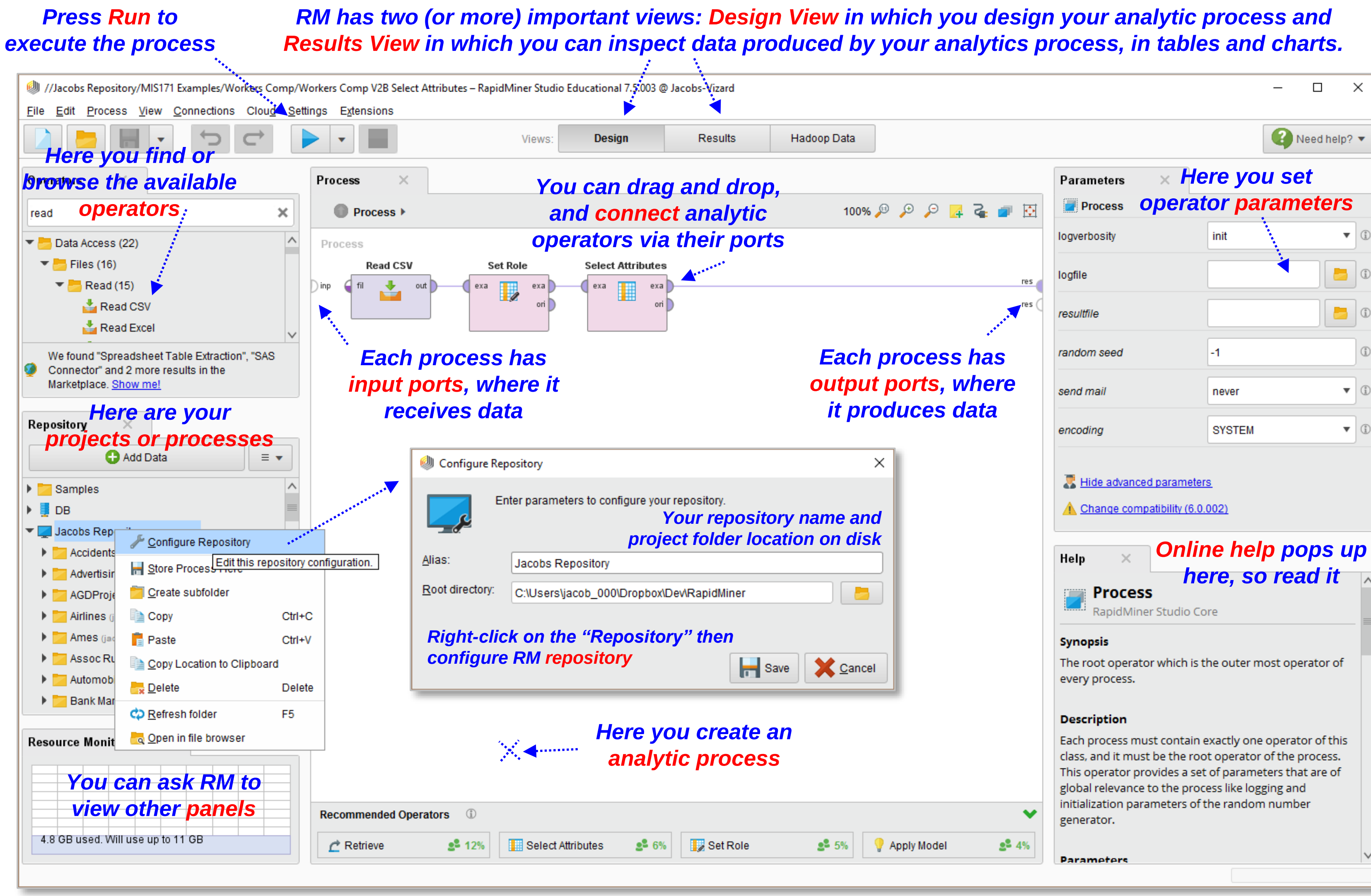
Row No.	prediction(S...	confidence(1)	confidence(0)	Claim Number	Adjustor Not...	Body Part	Nature of Inj...	Cause of Inj...	Vehicle Flag ...
1	0	0.200	0.800	160122408	Laceration to ...	Finger	Laceration	Struck Object	0
2	0	0.200	0.800	360784508	While unloadi...	Back	Sprain/Strain	Lifting	0
3	0	0.200	0.800	860564608	Claimant wa...	Hand	[illegible]	[illegible]	[illegible]
4	1	0.800	0.200	3060046708	Claimant wa...	Head	[illegible]	[illegible]	[illegible]
5	1	0.800	0.200	3160698008	Alleges carpa...	Wrist	Repetitive Mo...	Repetitive Mo...	0

Application - Use the predictive model to classify all new cases

Demonstration Using k-NN Classifier

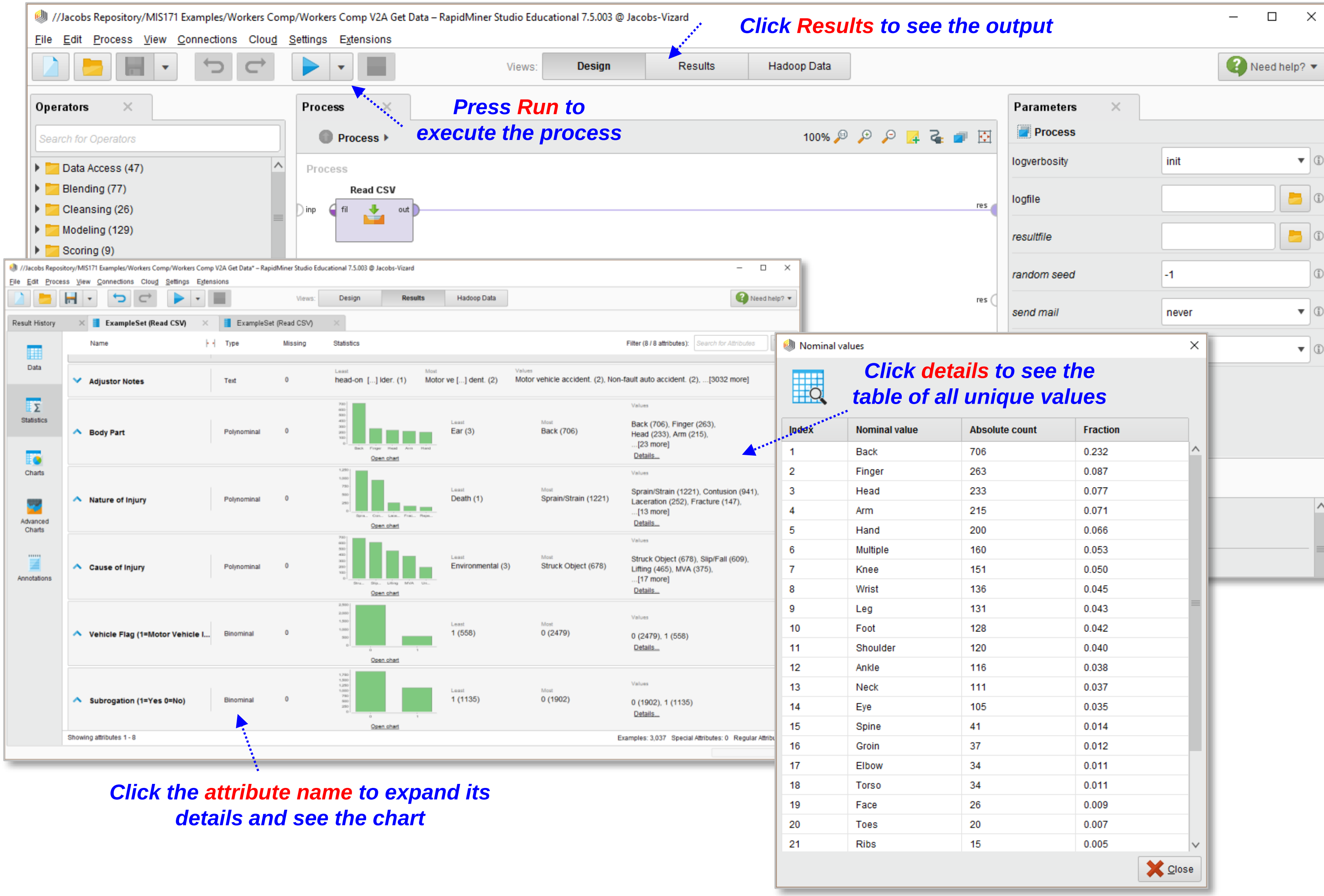
- **All slides from this point on in the lecture are for your reference only**
- **Watch and learn from the hands-on demonstration of building a classification model in RapidMiner Studio**
- **You can also view the demonstration in a pre-recorded video**
- **Observe that all modelling is done in very little steps**
- **You do a little modelling, run and test, then do a little analysis**
- **Try doing the same: watch a little, work with RapidMiner Studio a little and learn a little**
- **Enjoy the experience**

RapidMiner Studio In 10 Easy Steps



(1) Install RapidMiner. **(2)** Create a Project folder on your disk drive. **(3)** Create a Data folder inside the Project folder. **(4)** Get CSV files and place them in the Data folder. **(5)** Start RapidMiner. **(6)** Configure RapidMiner repository to point to your Project folder and be named as you like. **(7)** Start a new RapidMiner process. **(8)** Save it inside your Project folder (or its sub-folder). **(9)** Run and Explore the Results. **(10)** Enjoy RM Analytics!

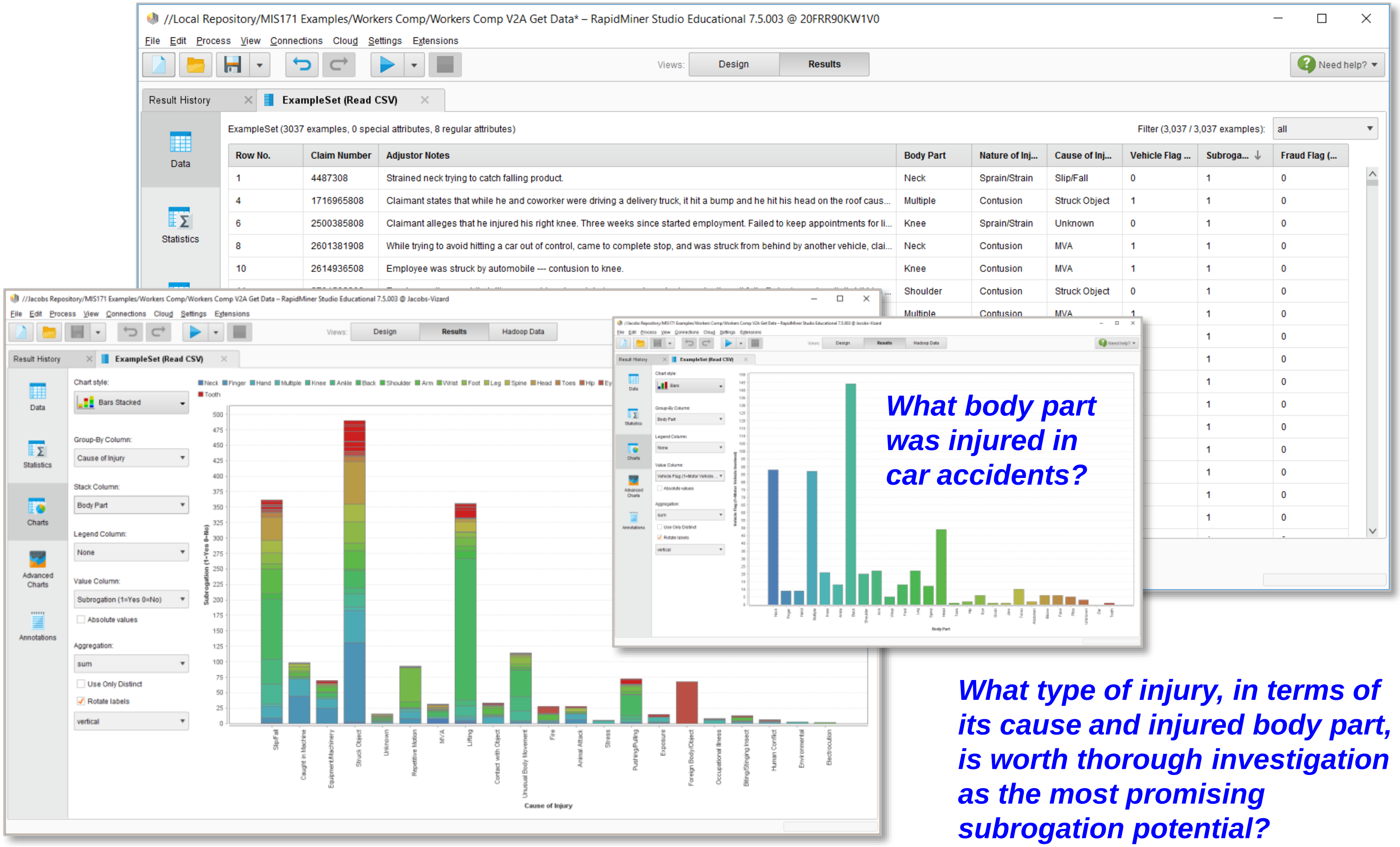
Read Data
Data Overview



Read the data – ensure all attributes (variables) are defined correctly. Work in small steps, so execute this “mini” model and see the results.

Conduct a quick overview of the data set. Check the basic statistics (min, max, mean, median, mode), distribution and values of all attributes.

In-depth Exploration



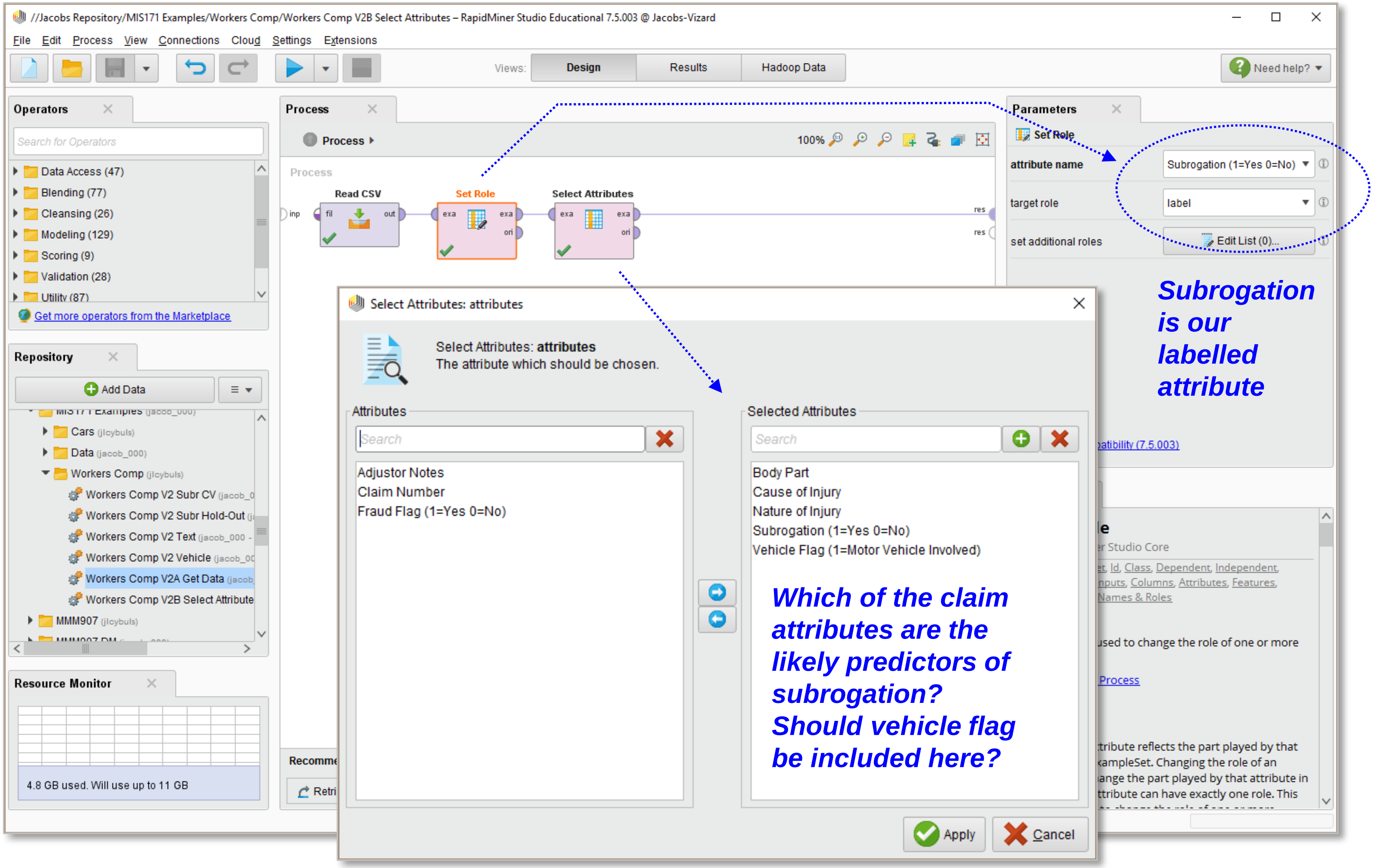
What type of injury, in terms of its cause and injured body part, is worth thorough investigation as the most promising subrogation potential?

Analyse each attribute in more detail, e.g. use column charts to investigate prevalence of certain attribute values.

Also look for possible relationships between variables, e.g. use stacked column charts and seek insights emerging from data.

Inspect raw data. If nominal / binomial variables were coded numerically, check how it was done. Is coding meaningful and correct?

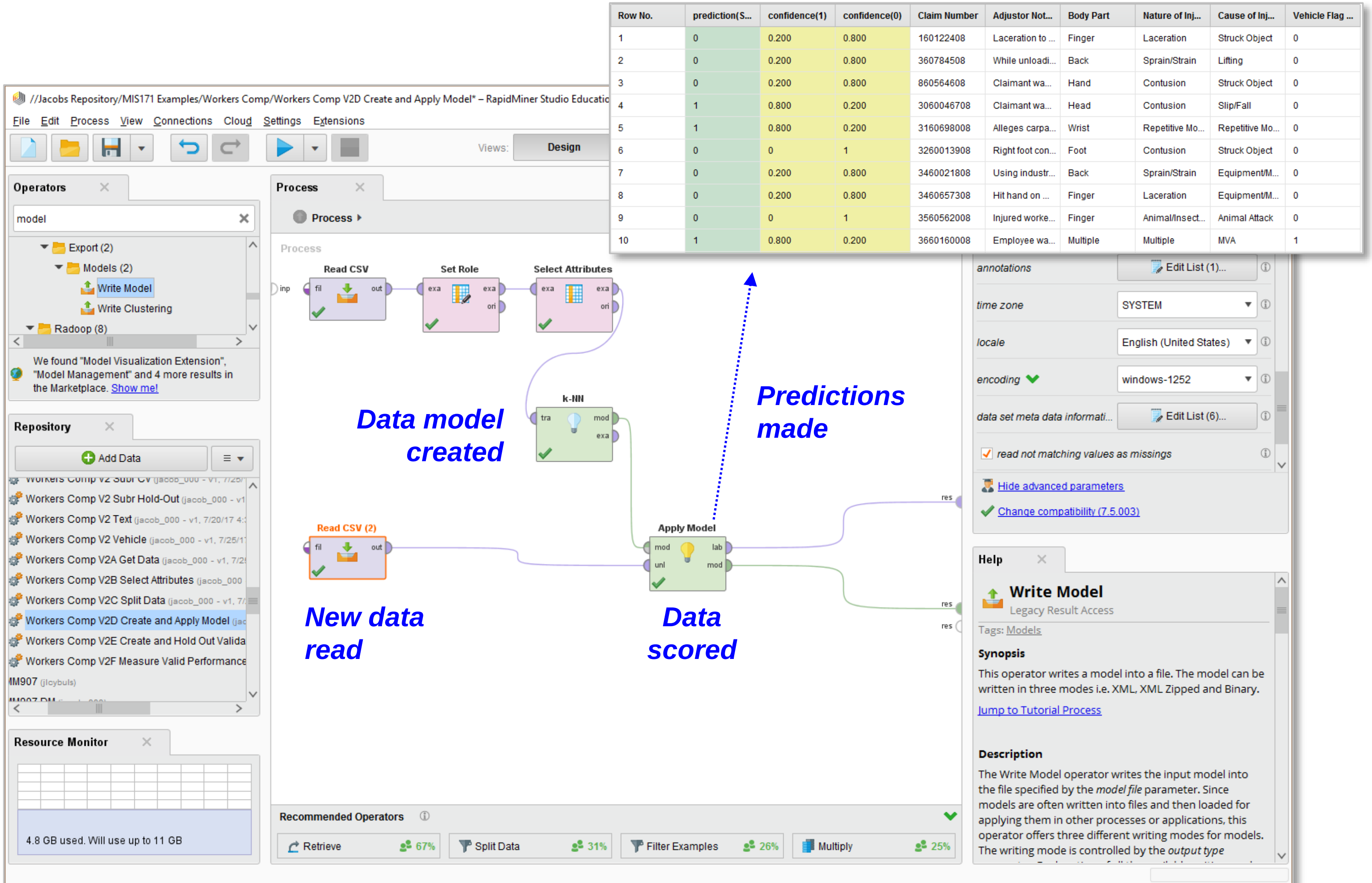
Select Attributes for Model Construction



Decide what aspect is to be predicted (something we cannot control), which becomes the target of your investigation, and define it as a “label” attribute.

Select those attributes, which are the potential predictors of the labelled attribute, and define them as “regular”. Execute the model and check!

Create a Model and Score New Data (Predict)



Row No.	prediction(S...	confidence(1)	confidence(0)	Claim Number	Adjustor Not...	Body Part	Nature of Inj...	Cause of Inj...	Vehicle Flag ...
1	0	0.200	0.800	160122408	Laceration to ...	Finger	Laceration	Struck Object	0
2	0	0.200	0.800	360784508	While unloadi...	Back	Sprain/Strain	Lifting	0
3	0	0.200	0.800	860564608	Claimant wa...	Hand	Contusion	Struck Object	0
4	1	0.800	0.200	3060046708	Claimant wa...	Head	Contusion	Slip/Fall	0
5	1	0.800	0.200	3160698008	Alleges carpa...	Wrist	Repetitive Mo...	Repetitive Mo...	0
6	0	0	1	3260013908	Right foot con...	Foot	Contusion	Struck Object	0
7	0	0.200	0.800	3460021808	Using industr...	Back	Sprain/Strain	Equipment/M...	0
8	0	0.200	0.800	3460657308	Hit hand on ...	Finger	Laceration	Equipment/M...	0
9	0	0	1	3560562008	Injured worke...	Finger	Animal/Insect...	Animal Attack	0
10	1	0.800	0.200	3660160008	Employee wa...	Multiple	Multiple	MVA	1

The selected data can now be used to develop a predictive model, e.g. we could use k-NN (k=5). The new model is then created. The model can then be written into a file and deployed. It can also be instantly applied to a new data to predict subrogation opportunities. However, how would we know if the model produces results that can be trusted?

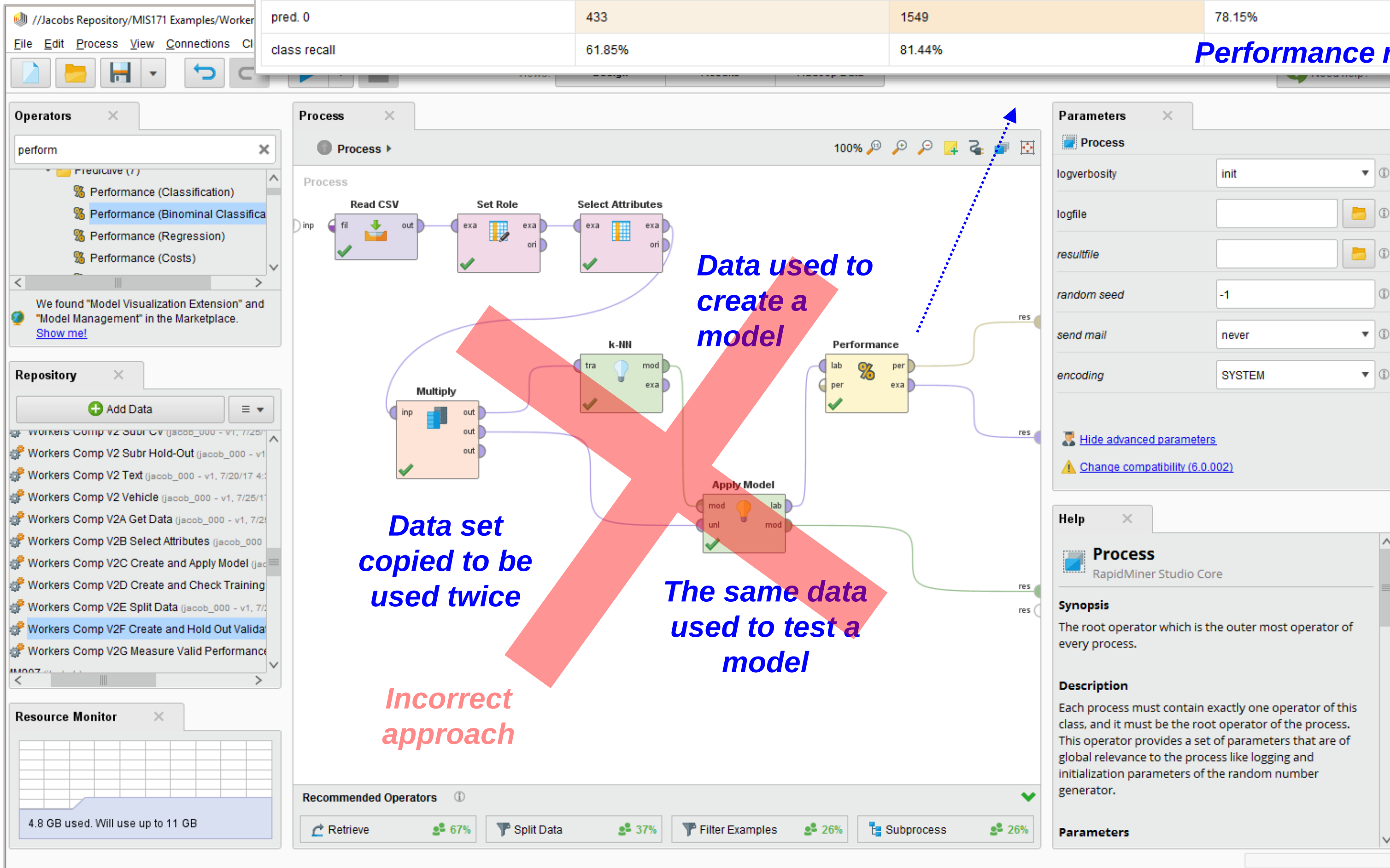
Create a Model and Score New Data (Predict)

Confusion Matrix

accuracy: 74.12%

	true 1	true 0	class precision
pred. 1	702	353	66.54%
pred. 0	433	1549	78.15%
class recall	61.85%	81.44%	

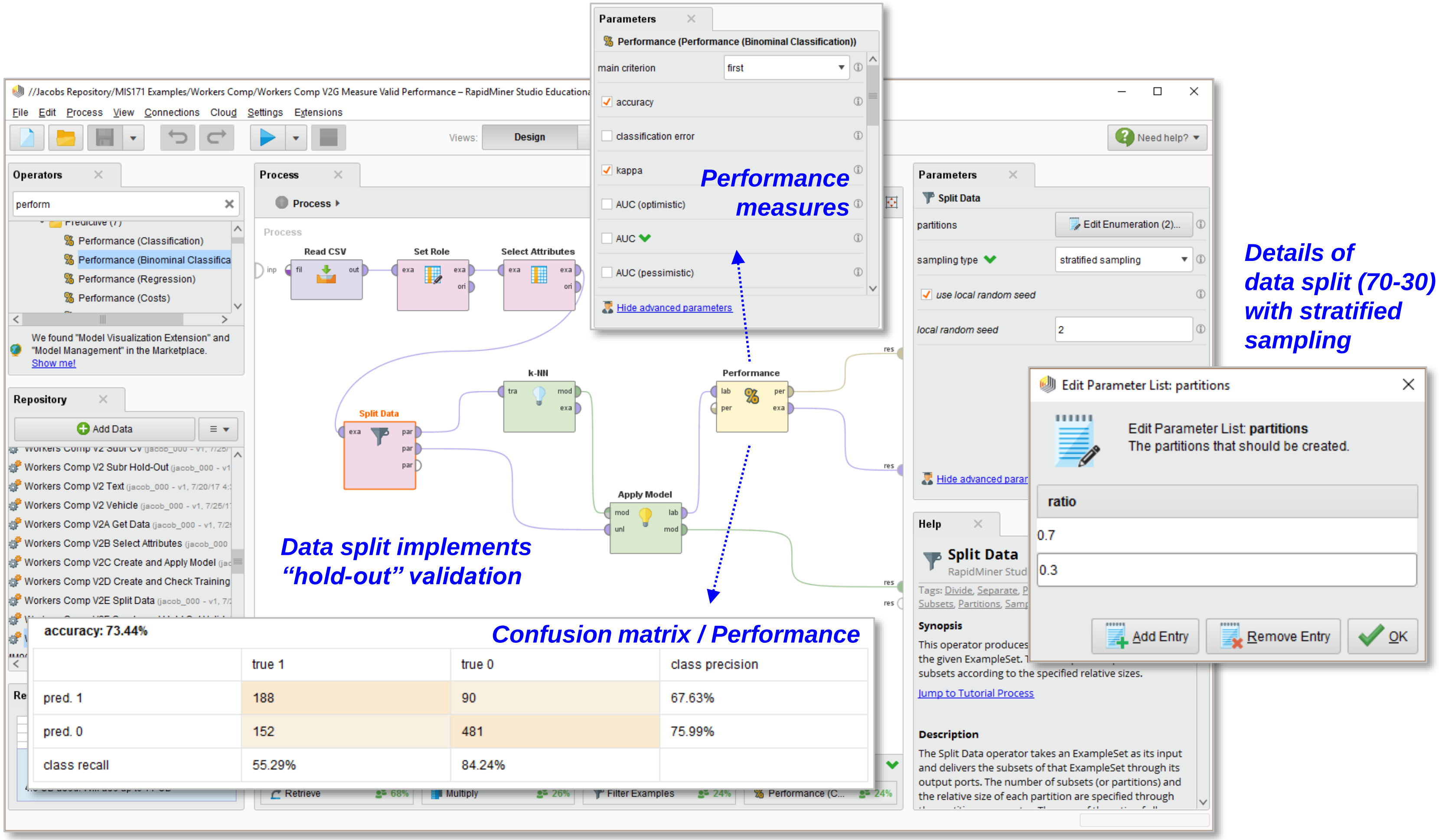
Performance reported



Once the model is created it can be tested on the same data that was used to create it. The model accuracy (the proportion of correct predictions) can then be reported. This result, however, cannot be trusted – all it tells us is how much of training data the model can remember!

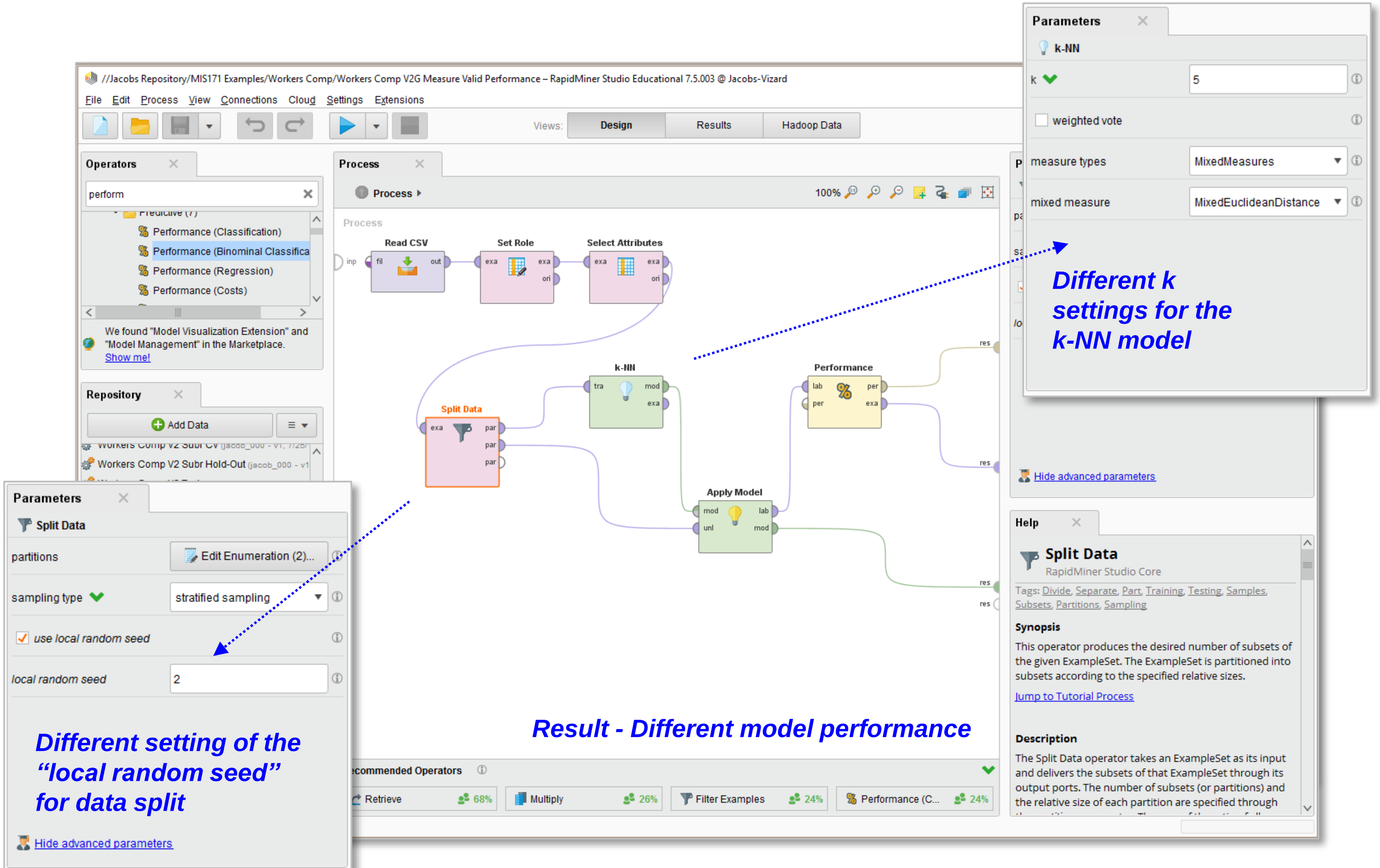
DEAKIN UNIVERSITY

Create and Validate a Data Model (k-NN)



- Instead, we can split the data into two parts, one to train the model (70%) and one to validate it (30%)
- To ensure the label values are distributed evenly between these two parts, we use stratified sampling
- We then create a k-NN model (k=5)
- The new model can then be applied to validation data, which was held out from training
- Performance statistics are calculated and reported

Experiment with Different Splits of Data



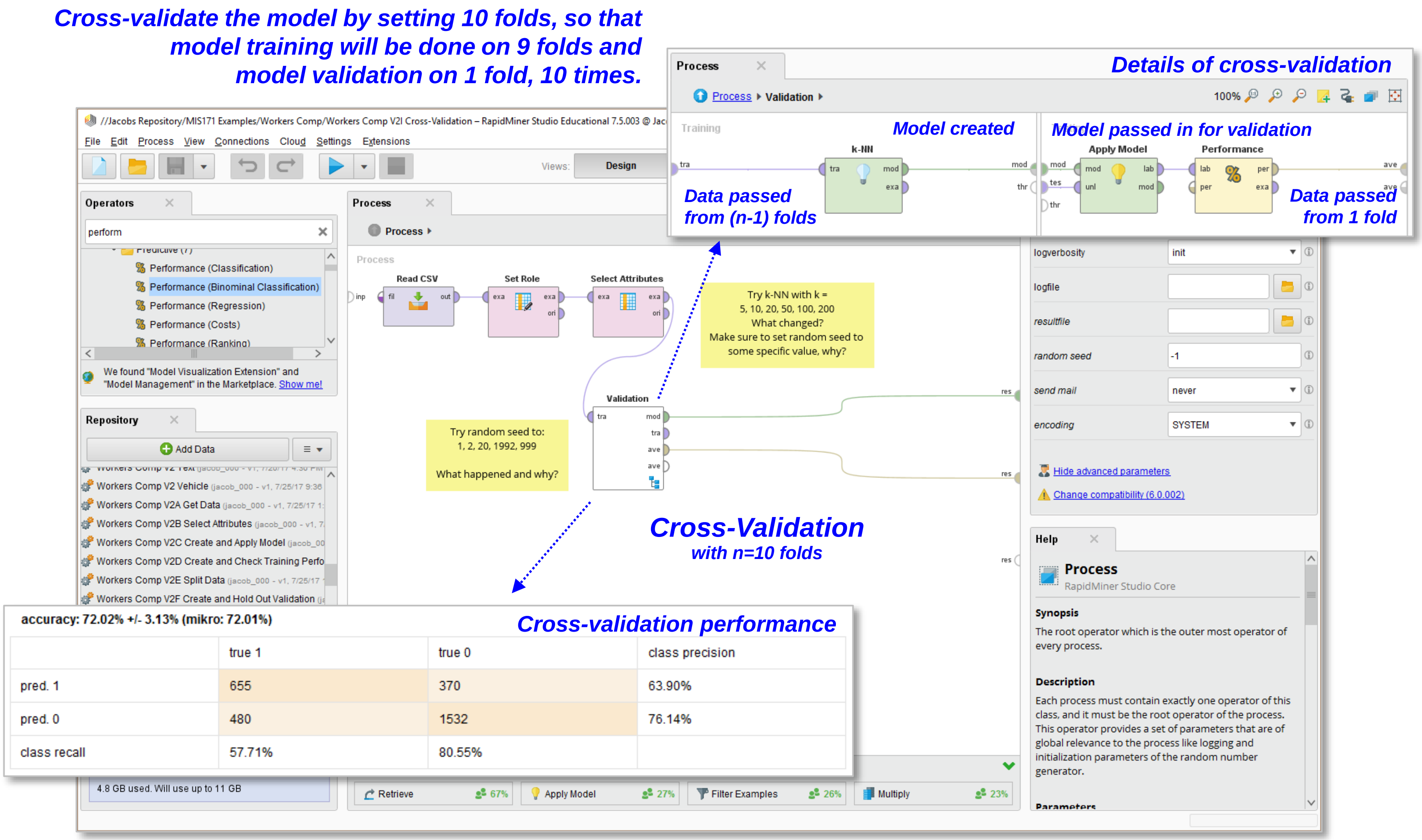
Experiment with different splits of data between training and validation sets, set the "local random seed" to values:

– 1, 2, 20, 1992, 999
– What have you observed?
– Why is this happening?

Experiment with different settings of the k-NN model, while keeping the same random split, set "k" to values:

– 5, 10, 20, 50, 100, 200
– What have you observed?
– Why is this happening?

Create and Cross-Validate a Data Model (k-NN)



	true 1	true 0	class precision
pred. 1	655	370	63.90%
pred. 0	480	1532	76.14%
class recall	57.71%	80.55%	

- When we have a small data sample, the model performance in hold-out validation is a lot of luck (good or bad)
- We may get vastly different model accuracy depending on the data split
- Cross-validation is thus used to determine a more realistic model performance
- We split data into n folds (n=10)
- We use n-1 folds to train the model and 1 fold to validate the model
- We repeat it n times, each time using a different fold for validation
- The model performance is then given as an average performance of n runs

Thinking Point

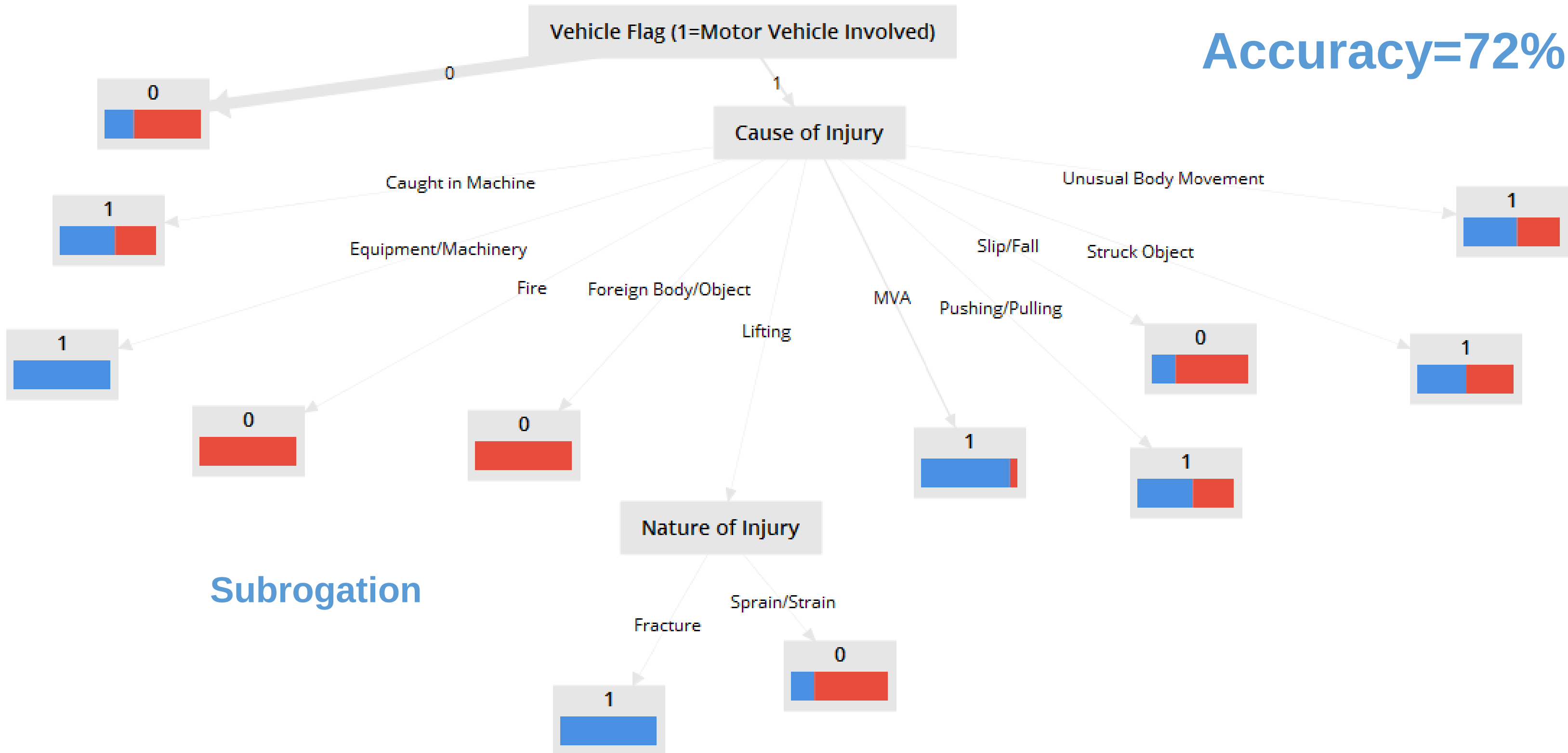
- Can the k-NN model be improved by optimizing its accuracy in respect of many possible “k” values?
- Can the prediction be improved by replacing k-NN with a different model, e.g.
  - Decision Tree, Random Forest or Gradient Boosted Trees
  - Logistic Regression or Neural Networks?
- Can the model be improved by processing the claims’ adjustor notes? Perhaps by relying on the text analytics methods, which are well supported in RapidMiner, and which can significantly enhance the model prediction.
- All such extensions, however, require more advanced analytic techniques, which are explained in further business analytics units.

Demonstration
Using Decision Trees

- There are many other methods of data classification
- One of the best performing are Decision Trees (or their Forests and Ensembles) which can be used for classification and estimation
- They can use both categorical (nominal) and continuous (interval) target variables
- Decision tree structure can be used to generate (business) rules able to classify or predict target variable based on the observation attributes
- Decision tree structure (sometimes) can be used to explain the prediction process to the client

DEAKIN UNIVERSITY

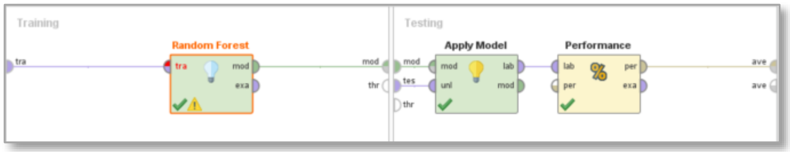
Decision Trees in RapidMiner

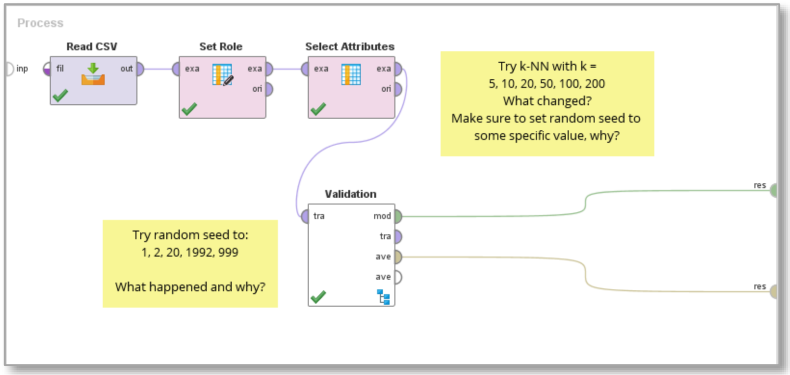


Accuracy=72%

Subrogation

In the previously developed model, all we have to change is to replace k-NN with one of many "decision tree" models





Summary

- Classification is the process of organizing a set of observations into a collection of labelled classes (categorical / nominal)
- Classification process commonly includes the model creation, its improvement and the subsequent use
- Similarity models compare observations to prototypes – well-known example or representatives, which may include all past observations – this could facilitate non-parametric classification or regression
- k Nearest Neighbour (or k-NN) models classify new observations by considering values of k closest matching prototypes
- There are many other modelling approaches to classification, e.g. Decision Trees, Gradient Boosted Trees, Random Forest (which are all decision trees or their collections)
- Decision trees often give the best predictive performance
- Performance of classification models is often measured in terms of their accuracy
- When the class we want to predict is not balanced, i.e. there are great many values of one type than others, then a simple accuracy will not work – there are many advanced approaches to deal with this. A simpler method is to check a "kappa" statistic which gives a more conservative assessment of accuracy
- For classification purposes, it is also possible to use models commonly used for estimation, e.g. (logistic) regression and neural networks.