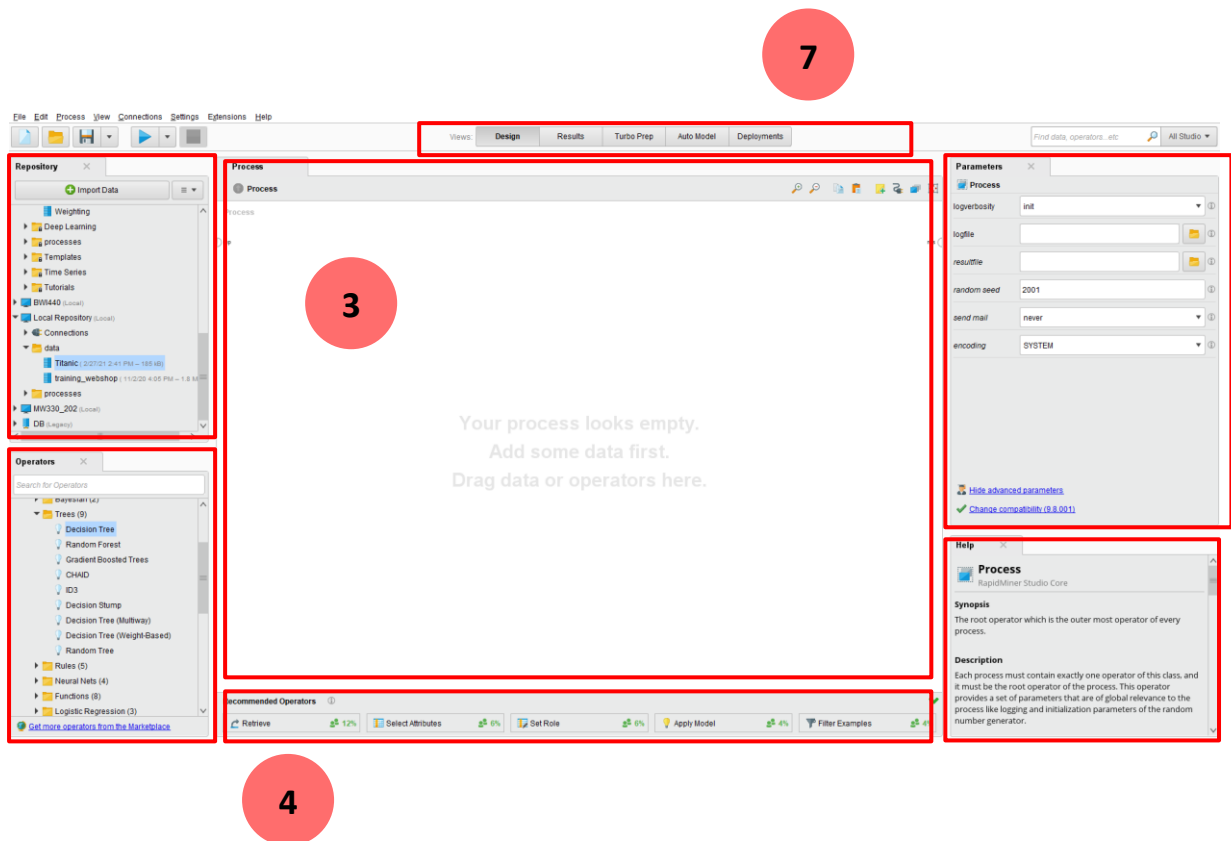


Use Case 1 – Einführung Rapid Miner



GUI

Wenn Sie Rapid Miner starten sehen Sie eine solche Benutzeroberfläche. Diese enthält nützliche Features zum Design Ihres Prozesses oder Ihrer Pipeline.



1. Repository: Hier werden Daten, Prozesse und Verbindungen gespeichert.
2. Operatoren: Operatoren werden zum modellieren/designen von Prozessen verwendet. Mit der Suchfunktion können alle verfügbaren Operatoren wie beispielsweise zum Laden, Verändern und Filtern von Daten gefunden werden. Ebenso Machine-Learning Modelle (Decision Tree etc.).
3. Dieser Bereich stellt den Prozessbereich dar, indem man durch Verknüpfen von Operatoren einen Prozess/Pipeline definieren kann. Alle Operatoren können per Drag & Drop in den Prozessbereich gezogen werden.
4. Diese Leiste steht für empfohlene Operatoren. Basierend auf den ausgewählten Operatoren wird hier eine Empfehlung für einen nächsten passenden Operator ausgesprochen. Diese richtet sich nach der Häufigkeit der Nutzung durch andere User.

Beispiel:

Empfehlung →  Retrieve  12%

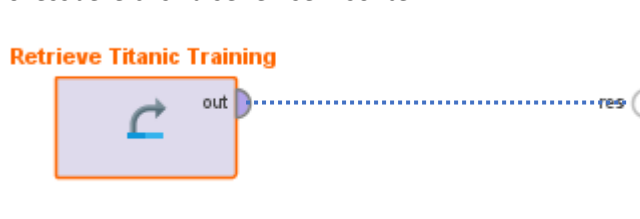
Aussage der Empfehlung: Basierend auf den ausgewählten Operatoren, haben 12% der User als nächsten Schritt einen Retrieve-Operator in ihren Prozessbereich eingebunden.

5. Rechts unten finden Sie den Help-Bereich. In diesem werden Operatoren und deren Parameter genauer beschrieben und ggf. auf andere Dokumentationen oder Hilfestellungen verwiesen.
6. Die Parameter sind ähnlich zu sehen wie die Eingabeparameter für eine Funktion/Methode. Hier sehen Sie welche Einstellungen für einen bestimmten Parameter modifizierbar sind und können diese dementsprechend anpassen. Besonders für das Machine Learning sind diese Parameter essenziell.
7. Views: Diese Leiste gewährt Ihnen Zugriff auf verschiedene Ansichten Ihres Prozesses. Beispielsweise kann hier zwischen „Design“ und „Result“ umgeschaltet werden. Dies ist nützlich, wenn man beispielsweise nach Begutachtung der Ergebnisse zum Prozessbereich zurückkehren will, um Änderungen vorzunehmen.

Erste Pipeline

Im folgenden werden Sie nun Ihre erste Pipeline definieren. Folgen Sie dazu den beschriebenen Schritten:

- I. Navigieren Sie in Ihrem Repository zu Samples → Data → Titanic Training Dataset.
- II. Ziehen Sie dieses mit Drag & Drop in den Prozessbereich.
- III. Verbinden Sie den Punkt „out“ mit dem „res“, welches am rechten Rand Ihres Prozessbereichs zu sehen sein sollte.



- IV. Anschließend führen Sie den Prozess aus, indem Sie auf das „Play“ Symbol in der oberen linken Ecke des Fensters drücken.

Sie sehen nun, dass Sie sich nun im „Result“-Tab Ihres Prozesses befinden.

Auf der linken Seite des Prozessbereichs sind nun mehrere zusätzliche Features wie Statistics, Visualizations und Annotations aufgetaucht. Mit diesen werden wir uns in den folgenden Use Cases näher auseinandersetzen.

- V. Wechseln Sie auf das Tab Statistics.

Sie sehen für jede Dimension eine detaillierte Beschreibung je nach Datentyp.

Aufgabe:

1. Können Sie nun sagen wieviel männliche und weibliche Fahrgäste sich an Bord befanden?

Erste Visualisierung

Unter dem Tab „Visualisierung“ können Sie sich das ganze Dataset grafisch aufbereitet anschauen. Hierfür können mehrere „Plots“ generiert werden.

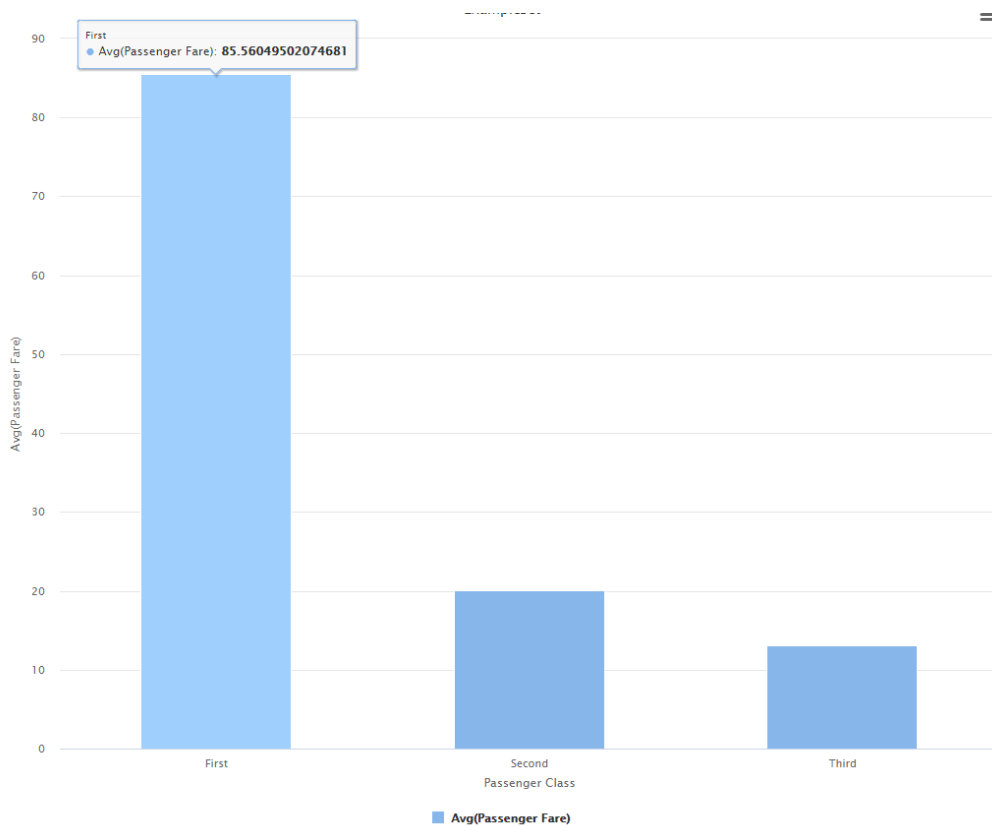
The image shows a configuration panel for 'Plot 1' with various settings and callouts explaining their functions:

- Plot type:** Set to 'Scatter / Bubble'. Callout: 'Hier kann der Chart Typ festgelegt werden, wie beispielsweise ein Bar/Column Chart etc.'
- X-Axis column:** Set to 'Sex'. Callout: 'Dimension, welche durch die X-Achse repräsentiert werden soll. In der Regel sind diese kategorisch, können jedoch auch numerisch sein.'
- Value column:** Set to 'Age'. Callout: 'Kennzahl, welche durch die Y-Achse repräsentiert wird. In der Regel numerisch oder kategorisch mit Aggregatsfunktion'
- Color:** Set to 'Survived'. Callout: 'Weitere Dimensionen, welche der Visualisierung hinzugefügt werden können.'
- Size:** Set to '-'. Callout: 'Chart-spezifische Parameter. Jitter fügt beispielsweise noise zu den Daten hinzu um Datenpunkte besser voneinander separieren zu können.'
- Jitter:** A slider control.
- Regression interpolation:** Set to 'None'.
- Plot style >>>** Callout: 'Formatoptionen'

Fertigen wir nun eine beispielhafte Visualisierung an (**Hinweis: Die folgenden Visualisierungen sind nur beispielhaft. Sie werden im Laufe der Veranstaltung grundlegende Vorgaben für Visualisierungen kennenlernen**):

- I. Wählen Sie den Chart-Typ aus, in diesem Falle ein „Bar Chart“.
- II. Ändern Sie die Dimension der X-Achse auf „Passenger Class“ und aktivieren den Haken „Aggregate Data“
Nun sehen Sie, dass ein Feld mit „Group By“ und „Aggregate Function“ aufgegangen ist. Stellen Sie als Aggregatsfunktion „Average“ ein.

- III. Ändern Sie die Value Column in „Passenger Fare“
Ihre Auswertung sollte nun wie folgt aussehen:



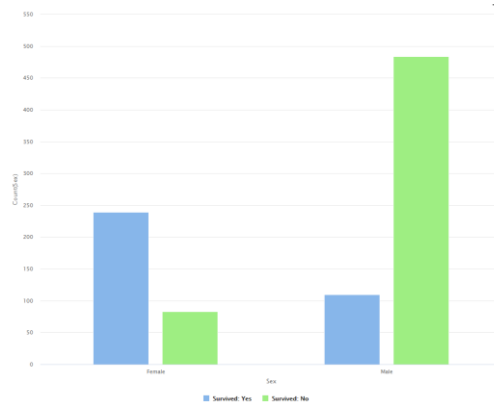
Sie können nun in der Color Group ebenfalls eine weitere Dimension hinzufügen. Fügen Sie beispielsweise „Sex“ als Gruppe hinzu.

Was sagt Ihnen diese Auswertung?

Aufgabe:

1. Fertigen Sie eine Visualisierung an, welche die absolute Anzahl an männlichen und weiblichen Passagieren gruppiert nach der Dimension „survived“ darstellt.

Die Visualisierung sollte in etwa so aussehen:



2. Was sagt Ihnen diese Auswertung?

Eigene Datasets im Repository speichern

Um auf eigene Datasets zugreifen zu können, müssen diese vorher im Repository verfügbar sein. Dies geht wie folgt:





- I. Laden Sie sich das „mpg“- Dataset aus dem Materialordner herunter.
- II. Klicken Sie im Repository-Fenster auf den Button „Import Data“.
- III. Wählen Sie „My Computer“.
- IV. Navigieren Sie zum Speicherort des Datasets und wählen Sie dieses aus.
- V. Die folgende Ansicht sollte etwas so aussehen:

<input checked="" type="checkbox"/> Header Row	1	File Encoding	windows-1252	<input checked="" type="checkbox"/> Use Quotes	"
Start Row	1	Escape Character	\	<input type="checkbox"/> Trim Lines	
Column Separator	Comma ","	Decimal Character	.	<input checked="" type="checkbox"/> Skip Comments	#

1	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	name
2	18	8	307	130	3504	12	70	1	chevrolet chev...
3	15	8	350	165	3693	11.5	70	1	buick skylark ...
4	18	8	318	150	3436	11	70	1	plymouth sate...
5	16	8	304	150	3433	12	70	1	amc rebel sst
6	17	8	302	140	3449	10.5	70	1	ford torino
7	15	8	429	198	4341	10	70	1	ford galaxie 5...
8	14	8	454	220	4354	9	70	1	chevrolet imp...
9	14	8	440	215	4312	8.5	70	1	plymouth fury iii
10	14	8	455	225	4425	10	70	1	pontiac catalina
11	15	8	390	190	3850	8.5	70	1	amc ambass...
12	15	8	383	170	3563	10	70	1	dodge challe...
13	14	8	340	160	3609	8	70	1	plymouth 'cud...
14	15	8	400	150	3761	9.5	70	1	chevrolet mon...
15	14	8	455	225	3086	10	70	1	buick estate ...
16	24	4	113	95	2372	15	70	3	toyota corona ...
17	22	6	198	95	2833	15.5	70	1	plymouth dust...

no problems.

- VI. Da „rapidminer“ das Format des Files schon richtig erkannt hat, ist hier keine Änderung vorzunehmen. → „Next“
- VII. Bei den Datentypen muss eine Anpassung gemacht werden. Das Feature horsepower wird als „polynomiales“ Feature erkannt, welches offensichtlich falsch ist. Ändern Sie dies indem Sie auf das Zahnrad neben dem Feature klicken → Change Type → integer.
- VIII. Ebenso bei dem Feature origin muss etwas geändert werden. Da es unser „label“ darstellt muss dies als „polynomiales“ Feature gekennzeichnet werden.

cylinders  ▼ <i>integer</i>	displacement  ▼ <i>real</i>	horsepower  ▼ <i>polynomial</i>	weight  ▼ <i>integer</i>
8	307.000	130	3504
8	350.000	165	3693

- IX. Da das Dataset Fehlwerte enthält, wird nun ein Error angezeigt. Klicken Sie unten rechts „Ignore Errors“ und stellen Sie sicher dass am oberen Teil des Fensters das Häkchen bei „Replace errors with missing values“ gesetzt ist.
- X. Wählen Sie einen Speicherort aus.