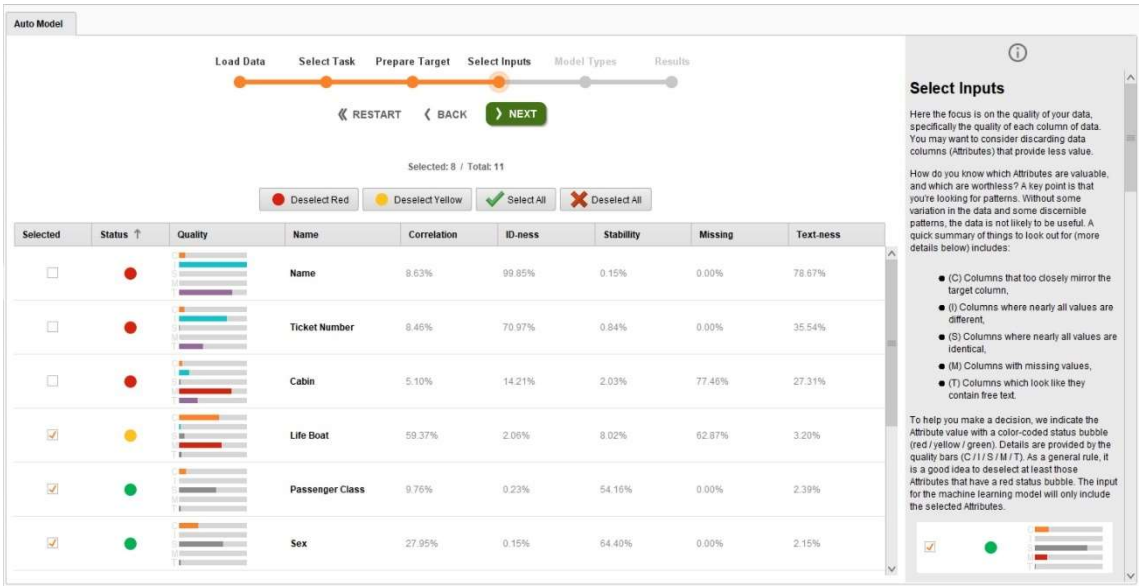
Auto Model と Turbo Prep のオートクリーニング機能を再現

—“Quality Measures”によるデータクリーニング—

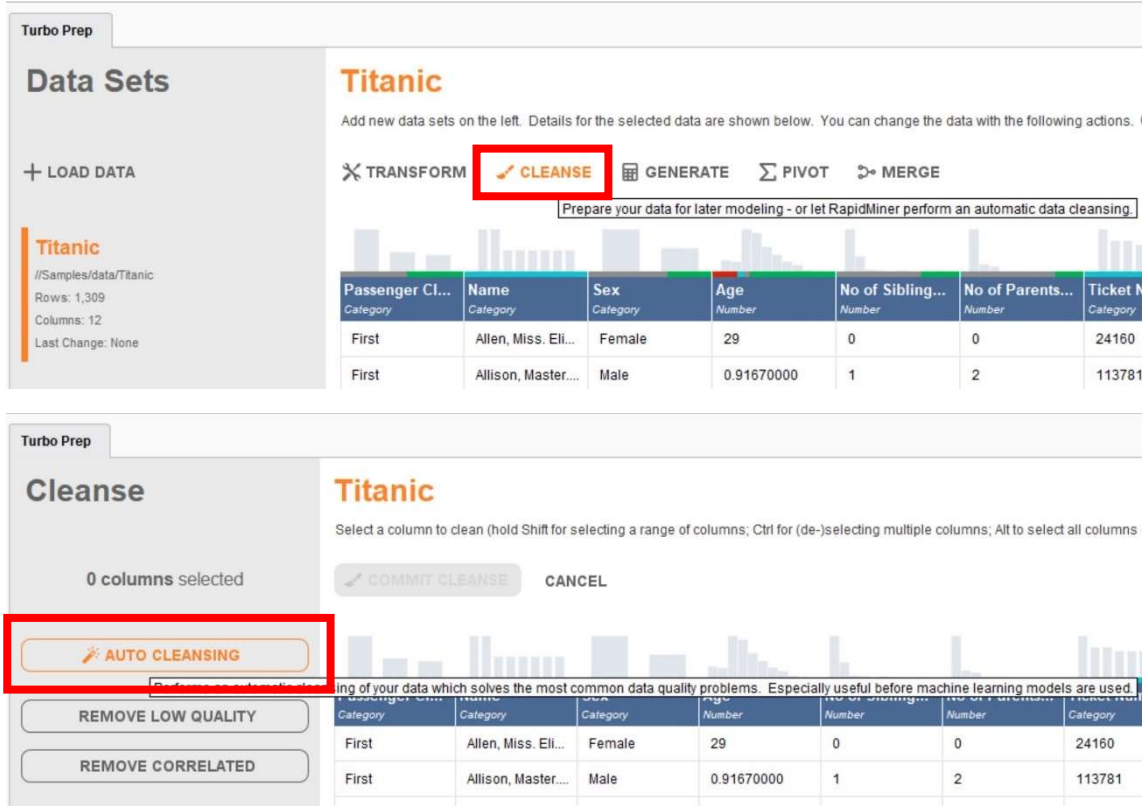
Auto Model と TurboPrep の両方に、データセットから不要な特徴量を削除できる「オートクリーニング」機能があります。 RapidMiner は特徴量を削除する必要があるかどうかを判断するために「品質測定」(Quality Measures)を実行します。この「品質測定」をご存じでない場合は、以下を参照してください:

- **Correlation:** 説明変数と目的変数の線形相関測定
- **ID-ness:** 特徴量の値が独自的（ユニーク）であるかの測定
- **Stability:** 特徴量の値がどの程度同一であるかの測定
- **Missing:** 特徴量の欠落値の％測定
- **Text-ness:** 特徴量に含まれるフリーテキストの量の測定

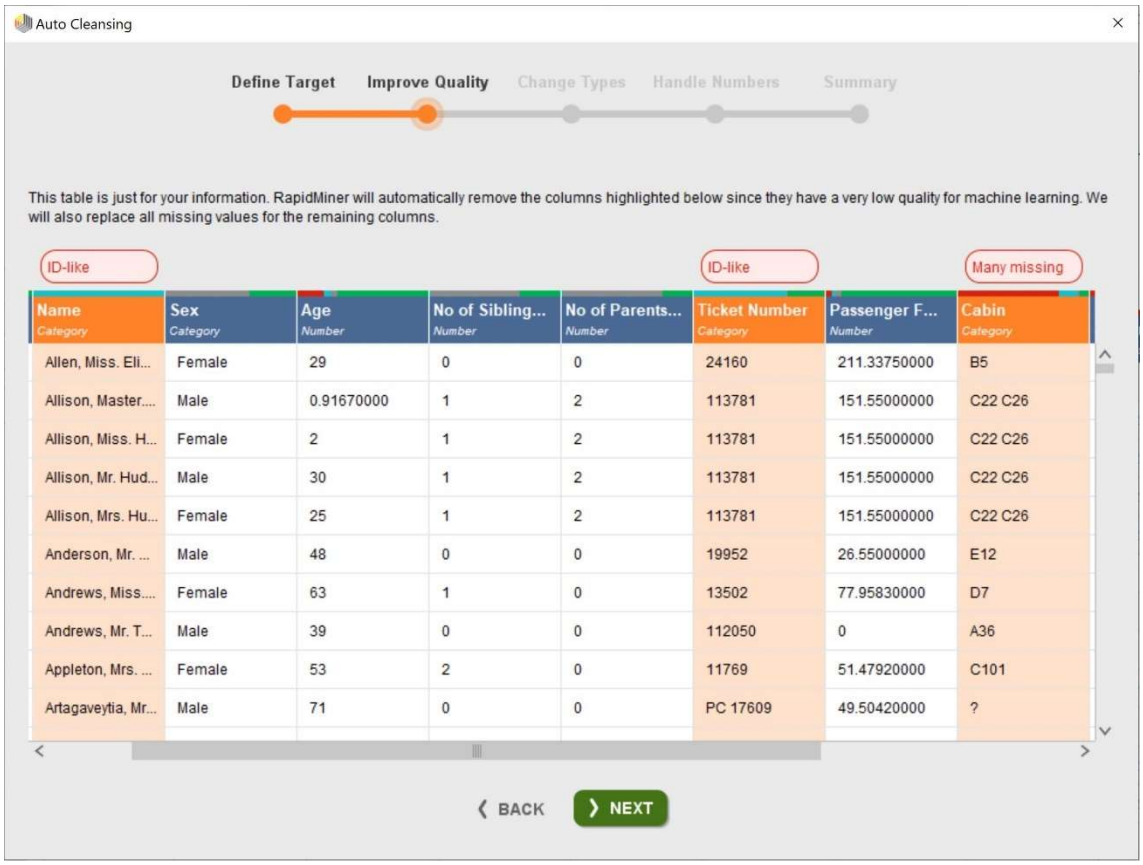
Auto Model では、「品質測定」は「入力の選択」ページで使用されます:



Turbo Prep では最初に「Clense」ボタンをクリックしてから、左側の「Auto Cleansing」オプションをクリックすることで利用できます:



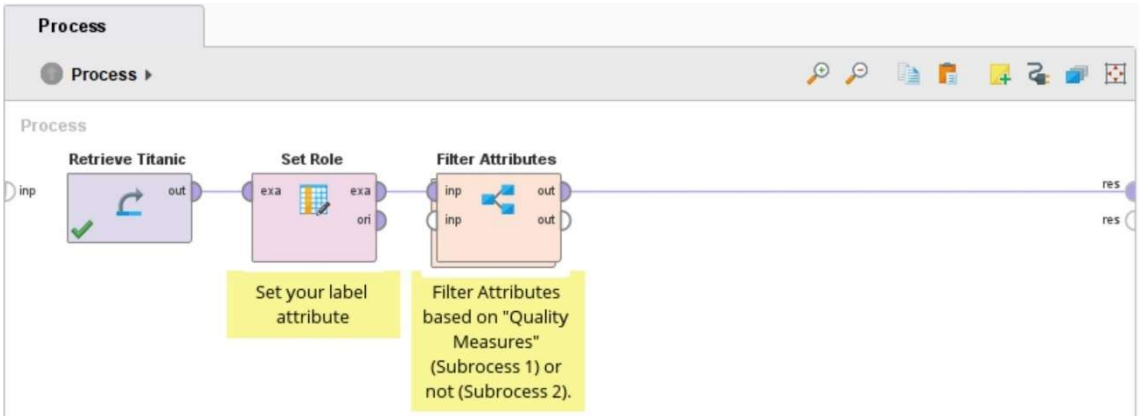
そこから、「品質測定」は「Improve Quality」ステップで使用されます:



これらの 2 つのオプションは便利ですが、クリーニングのために特徴量を選択する方法をあまり制御できません。

この便利な機能を自分のプロセスで使用するには、「Quality Measures」オペレーターを使用できます。 このオペレーターは、データの特徴量について、Auto Model および TurboPrep と同じ方法で計算します。 このオペレーターをフィルタリングと組み合わせると、お好みに合わせてクリーニングプロセスを制御できます。 見てみましょう。

実装の内容

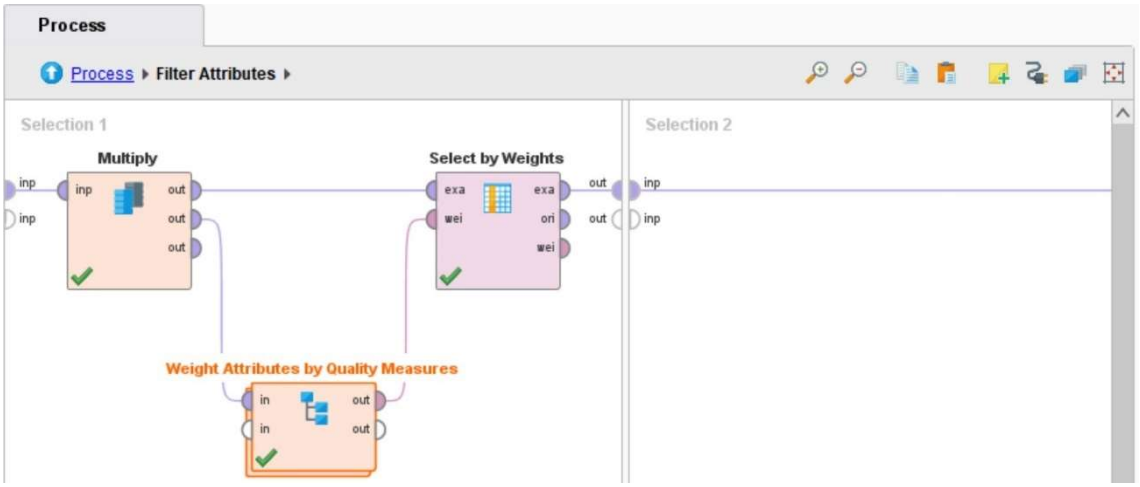


まず、属性の 1 つに「ラベル」の役割を与えるようにしてください。 そうしないと、「相関」(Correlation)の品質測定値を作成できません。

Row No.	Survived	Passenger ...	Name	Sex	Age	No of Sibling...	No of Parent...	Ticket Numb...	Passenger F...	Cabin	Port of Emb...	Life Boat
1	Yes	First	Allen, Miss. E...	Female	29	0	0	24160	211.33750000	B5	Southampton	2
2	Yes	First	Allison, Mast...	Male	0.91670000	1	2	113781	151.55000000	C22 C26	Southampton	11
3	No	First	Allison, Miss. ...	Female	2	1	2	113781	151.55000000	C22 C26	Southampton	?
4	No	First	Allison, Mr. H...	Male	30	1	2	113781	151.55000000	C22 C26	Southampton	?
5	No	First	Allison, Mrs. ...	Female	25	1	2	113781	151.55000000	C22 C26	Southampton	?
6	Yes	First	Anderson, Mr....	Male	48	0	0	19952	26.55000000	E12	Southampton	3
7	Yes	First	Andrews, Mis...	Female	63	1	0	13502	77.95830000	D7	Southampton	10
8	No	First	Andrews, Mr. ...	Male	39	0	0	112050	0	A36	Southampton	?
9	Yes	First	Appleton, Mrs...	Female	53	2	0	11769	51.47920000	C101	Southampton	D
10	No	First	Artagaveytia, ...	Male	71	0	0	PC 17609	49.50420000	?	Cherbourg	?
11	No	First	Astor, Col. Jo...	Male	47	1	0	PC 17757	227.52500000	C62 C64	Cherbourg	?
12	Yes	First	Astor, Mrs. Jo...	Female	18	1	0	PC 17757	227.52500000	C62 C64	Cherbourg	4
13	Yes	First	Aubart, Mme. ...	Female	24	0	0	PC 17477	69.30000000	B35	Cherbourg	9
14	Yes	First	Barber, Miss. ...	Female	26	0	0	19877	78.85000000	?	Southampton	6
15	Yes	First	Barkworth, Mr...	Male	80	0	0	27042	30	A23	Southampton	B
16	No	First	Baumann, Mr...	Male	?	0	0	PC 17318	25.92500000	?	Southampton	?
17	No	First	Baxter, Mr. Qu...	Male	24	0	1	PC 17558	247.52080000	B58 B60	Cherbourg	?
18	Yes	First	Baxter, Mrs. J...	Female	50	0	1	PC 17558	247.52080000	B58 B60	Cherbourg	6
19	Yes	First	Bazzani, Miss...	Female	32	0	0	11813	76.29170000	D15	Cherbourg	8

ExampleSet (1,309 examples, 1 special attribute, 11 regular attributes)

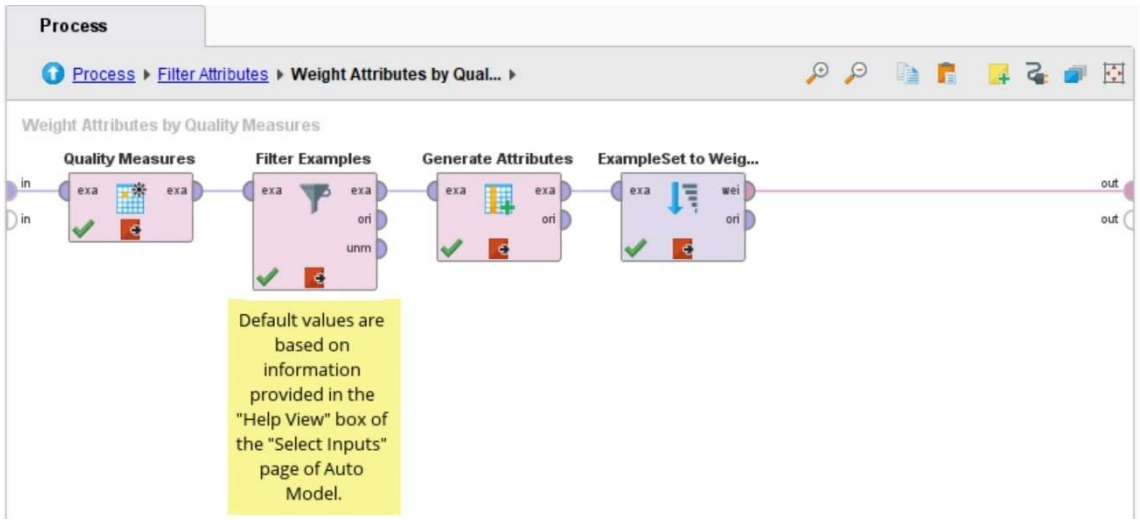
その後、データは特徴量フィルタリングのプロセスに入ります。 フィルタリングプロセスをこの場合このように作成しましたが、お好みで変えても構いません。



「Select Subprocess」オペレーター（ここでは Filter Attributes と名前を変えています）内にこのように 3 つの全体的なステップに分けてオペレーターを配置します:

1. データを複製します（Multiply オペレーター）
2. 「Quality Measures」オペレーターで、フィルタリングする特徴量を選択します
3. 「Select by Weights」オペレーターを使用して特徴量をフィルタリングします

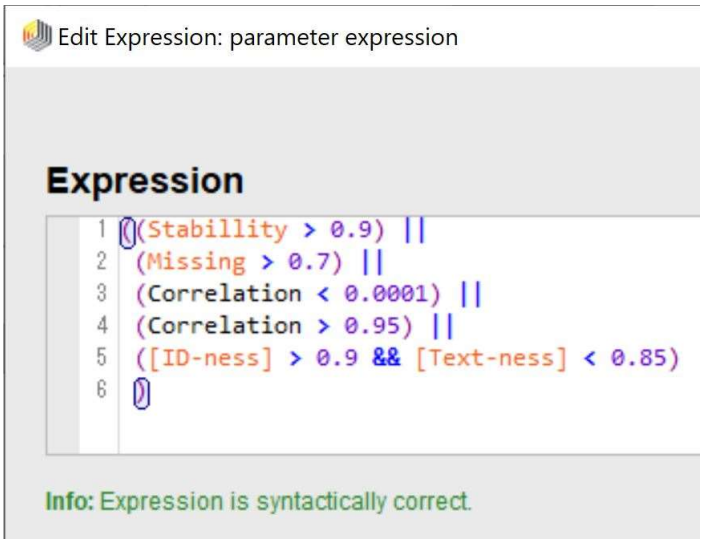
「Weight Attributes by Quality Measures」と呼ばれた 2 番目のステップで、ほとんどの作業を行っています。



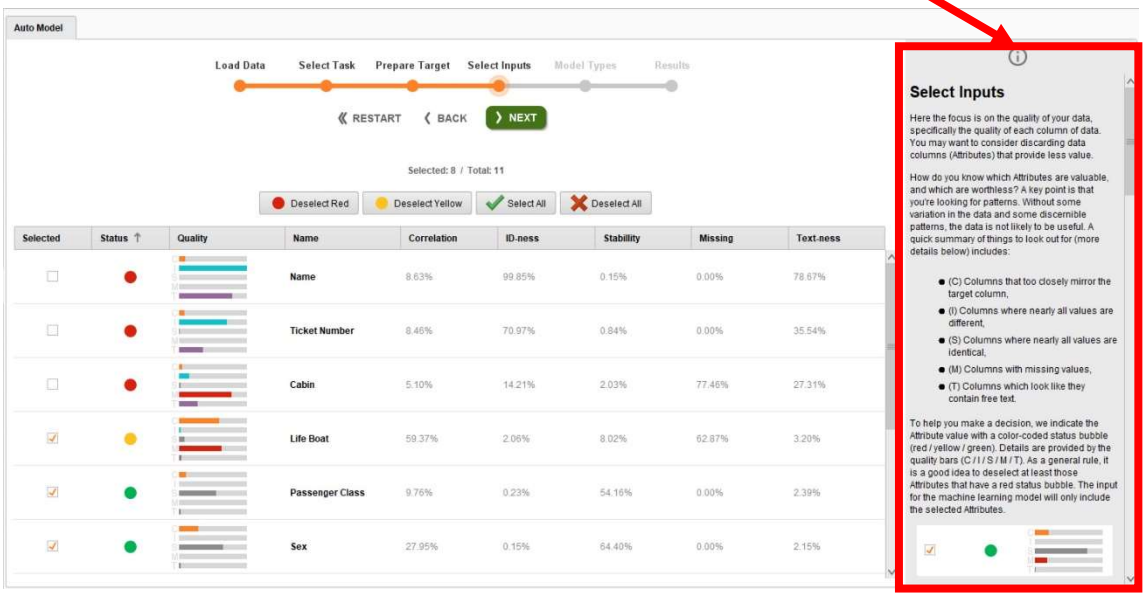
最初に、品質測定値が計算されて、結果（以下に表示）には、元のデータの各特徴量の行と、その特徴量の品質測定値に関する情報が含まれます。

Row No.	Attribute	Correlation	ID-ness	Stability	Missing	Text-ness
1	Passenger Class	0.09763710	0.00229183	0.54163484	0	0.02392666
2	Name	0.08632738	0.99847212	0.00152788	0	0.78673797
3	Sex	0.27951638	0.00152788	0.64400306	0	0.02145149
4	Age	0.00308164	0.05500382	0.04493308	0.20091673	0
5	No of Siblings or Spouses on Board	0.00077424	0.00534759	0.68067227	0	0
6	No of Parents or Children on Board	0.00683260	0.00611154	0.76546982	0	0
7	Ticket Number	0.08463433	0.70970206	0.00840336	0	0.35536542
8	Passenger Fare	0.05966562	0.02444614	0.04587156	0.00076394	0
9	Cabin	0.05095917	0.14209320	0.02033898	0.77463713	0.27311488
10	Port of Embarkation	0.00905755	0.00229183	0.69931140	0.00152788	0.04739947
11	Life Boat	0.59374771	0.02062643	0.08024691	0.62872422	0.03197988

その後、「Filter Examples」オペレーターを使用してフィルター処理する必要があります。 そうすれば、保持したい特徴量だけが残ります。 次のように式を書いてあります:



使用される値は、AutoModel の[Select Inputs]ページの[Help View]ボックスに表示される情報に基づいています。 ただし、必要に応じて変更できます。



その後、「Generate Attributes」オペレーターを使い、新しく weight 列を作成します。

Row No.	Attribute	Correlation	ID-ness	Stabillity	Missing	Text-ness	weight
1	Passenger Class	0.09763710	0.00229183	0.54163484	0	0.02392666	1
2	Sex	0.27951638	0.00152788	0.64400306	0	0.02145149	1
3	Age	0.00308164	0.05500382	0.04493308	0.20091673	0	1
4	No of Siblings or Spouses on Board	0.00077424	0.00534759	0.68067227	0	0	1
5	No of Parents or Children on Board	0.00683260	0.00611154	0.76546982	0	0	1
6	Passenger Fare	0.05966562	0.02444614	0.04587156	0.00076394	0	1
7	Port of Embarkation	0.00905755	0.00229183	0.69931140	0.00152788	0.04739947	1
8	Life Boat	0.59374771	0.02062643	0.08024691	0.62872422	0.03197988	1

次に、「ExampleSet to Weights」オペレーターを使用して、サンプルセットを「weights vector」に変換しました。

attribute	weight
Passenger Class	1
Sex	1
Age	1
No of Siblings or Spouses on Board	1
No of Parents or Children on Board	1
Passenger Fare	1
Port of Embarkation	1
Life Boat	1

最後に、「weights vector」は、「Select by Weights」オペレーターが関連する特徴量を選択するために使用します。

Row No.	Survived	Passenger Class	Sex	Age	No of Siblings or Spouses on Board	No of Parents or Children on Board	Passenger Fare	Port of Embarkation	Life Boat
1	Yes	First	Female	29	0	0	211.33750000	Southampton	2
2	Yes	First	Male	0.91670000	1	2	151.55000000	Southampton	11
3	No	First	Female	2	1	2	151.55000000	Southampton	?
4	No	First	Male	30	1	2	151.55000000	Southampton	?
5	No	First	Female	25	1	2	151.55000000	Southampton	?
6	Yes	First	Male	48	0	0	26.55000000	Southampton	3
7	Yes	First	Female	63	1	0	77.95830000	Southampton	10
8	No	First	Male	39	0	0	0	Southampton	?
9	Yes	First	Female	53	2	0	51.47920000	Southampton	D
10	No	First	Male	71	0	0	49.50420000	Cherbourg	?
11	No	First	Male	47	1	0	227.52500000	Cherbourg	?
12	Yes	First	Female	18	1	0	227.52500000	Cherbourg	4
13	Yes	First	Female	24	0	0	69.30000000	Cherbourg	9
14	Yes	First	Female	26	0	0	78.85000000	Southampton	6
15	Yes	First	Male	80	0	0	30	Southampton	B
16	No	First	Male	?	0	0	25.92500000	Southampton	?
17	No	First	Male	24	0	1	247.52080000	Cherbourg	?
18	Yes	First	Female	50	0	1	247.52080000	Cherbourg	6
19	Yes	First	Female	32	0	0	76.29170000	Cherbourg	8

ExampleSet (1,309 examples, 1 special attribute, 8 regular attributes)

この出力は、必要に応じてさらに処理したり、モデリングで使用したりできます。

以上のやり方で通常のプロセスでも「品質測定」を実行出来るようになりますので、是非、試してみてください。