

1. Импортируем большой датасет Data в RapidMiner Studio (File -> Import Data -> My Computer).
2. Формат данных оставляем по умолчанию.

Specify your data format

<input checked="" type="checkbox"/> Header Row 1	File Encoding UTF-8	<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 #

3. На следующем шаге, не забываем поставить галочку:

Format your columns.

Date format:

☒ Replace errors with missing values

	dur <small>real</small>	proto <small>polynomial</small>	service <small>polynomial</small>	state <small>polynomial</small>	spkts <small>integer</small>	dpkts <small>integer</small>	sbytes <small>integer</small>	dbytes <small>integer</small>
1	0.000	udp	-	INT	2	0	496	0
2	0.000	udp	-	INT	2	0	1762	0
3	0.000	udp	-	INT	2	0	1068	0
4	0.000	udp	-	INT	2	0	900	0

4. Перетаскиваем загруженный датасет в рабочую область, перейдя в режим Design

Views: Design Results Turbo Prep Auto Model

5. Добавляем оператор **Filter Examples**

Operators

Filter

- ▼ Blending (5)
 - ▼ Attributes (3)
 - ▼ Selection (3)
 - Select Attributes
 - Remove Useless Attributes
 - Remove Correlated Attribute
 - ▼ Examples (2)
 - ▼ Filter (2)
 - Filter Examples**
 - Filter Example Range
 - ▼ Cleansing (1)

6. Добавляем фильтр полю *is_anomally* – “is not missing” (рисунок 1). Необходимо для того, чтобы на всякий случай исключить отсутствующие значения.

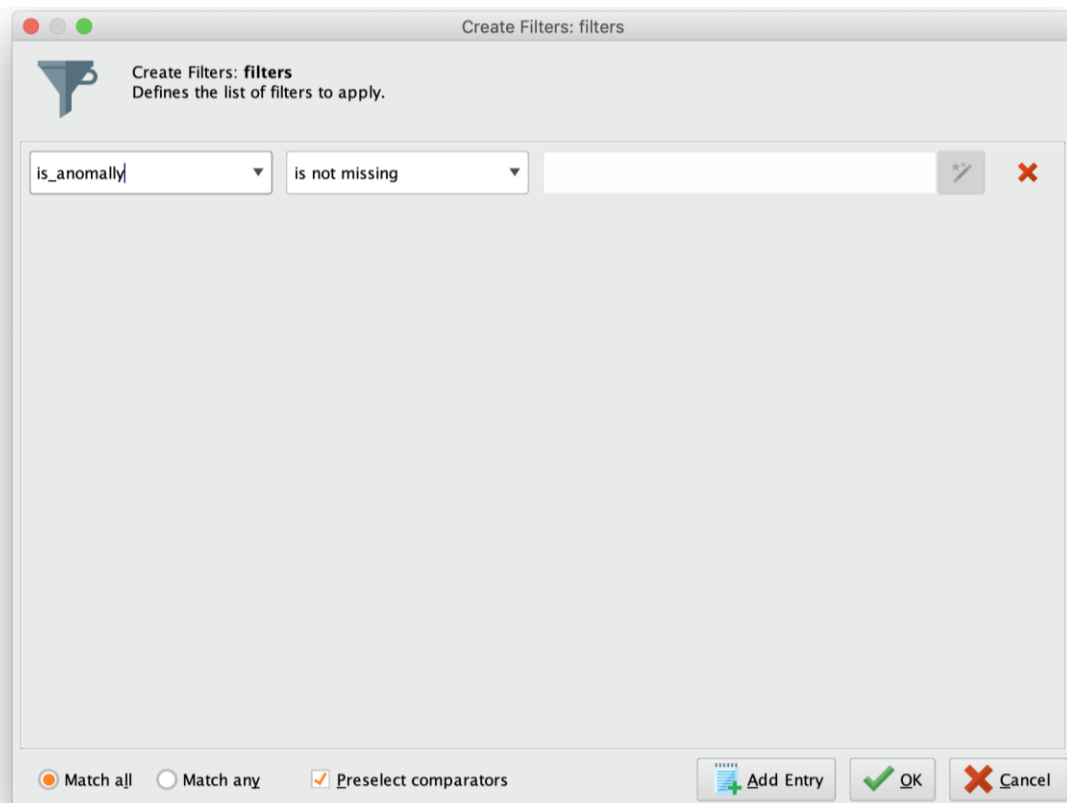
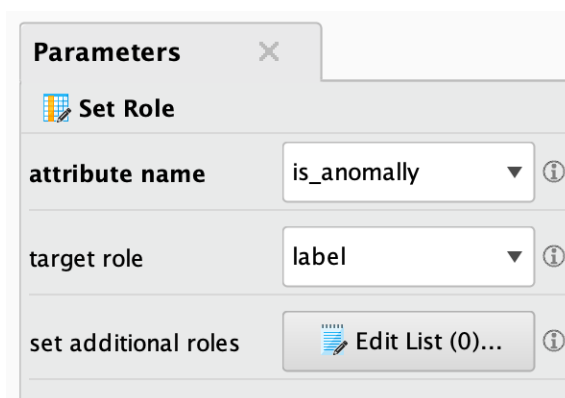


Рисунок 1

7. Далее с помощью оператора **Set Role** выбираем нужное поля для предсказания. В нашем случае – *is_anomaly*, target role помечаем как label.



8. Добавляем элемент **Nominal to Numerical** (нужен для избежания ошибок на этапе предсказания). Соединяем линии так, как указано на рисунке 2.

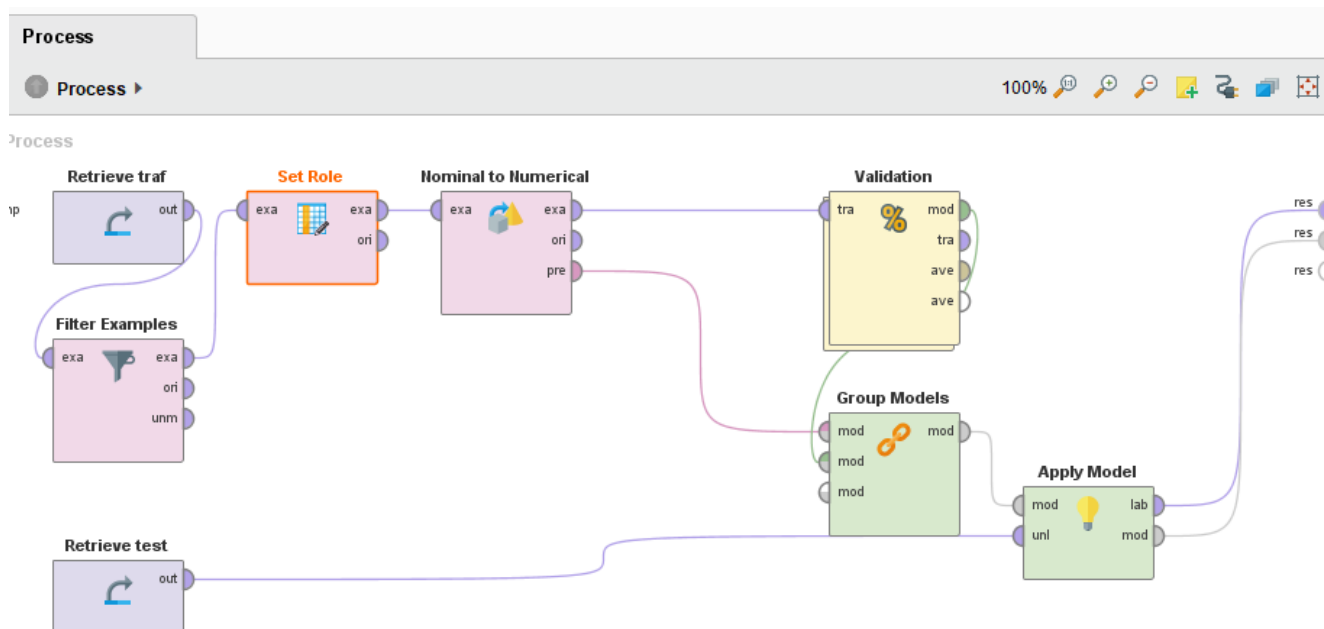
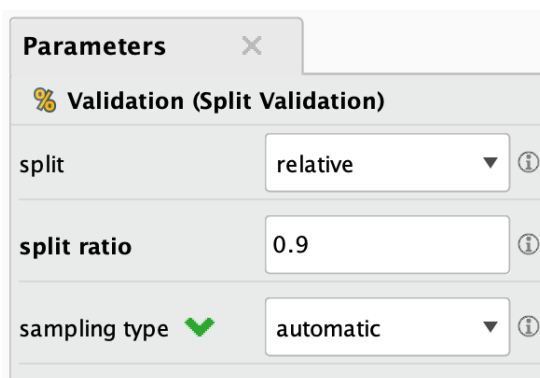


Рисунок 2

9. Добавляем оператор **Split Validation**, внутри него организовываем разбивку датасета на Training и Testing в пропорциях 90% и 10% соответственно.



10. Открываем оператор **Validation** двойным кликом и вставляем операторы **Decision Tree**, **Apply Model**, **Performance**, соединяем как указано на рисунке 3.

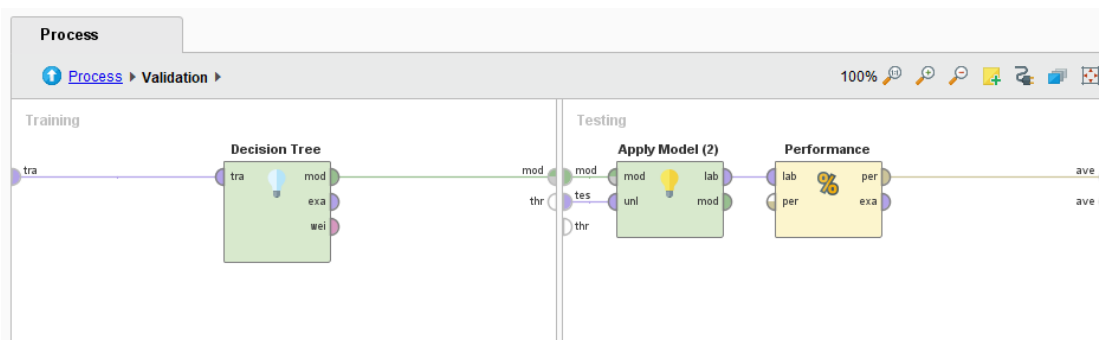


Рисунок 3

11. Выходим из **Split Validation**, далее в общую схему добавляем оператор **Group Models** (необходим для классификации новых данных), также добавляем ещё один оператор **Apply Model**. Импортируем новые данные (датасет без указания, аномальный ли трафик) также с помощью **Import Data**, добавляем его в схему и соединяем линиями так, как это указано на рисунке 4, должна получиться подобная схема.

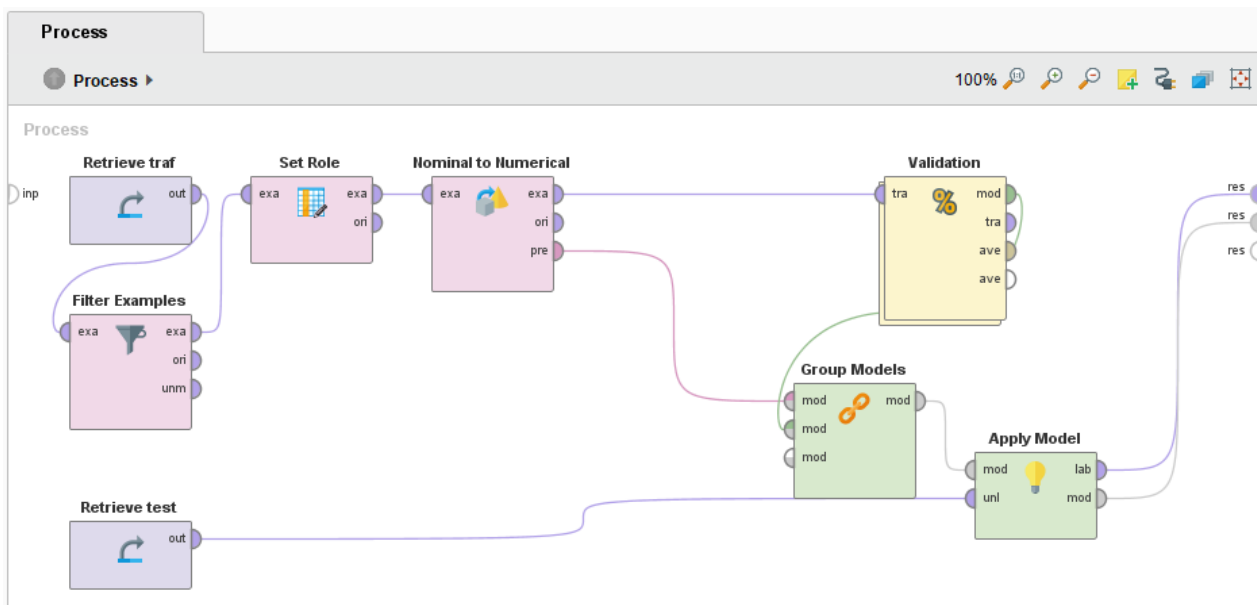


Рисунок 4

12. Далее нажимаем на Play (начать классификацию), открываем результаты и смотрим предсказание по трафику добавленным новым данным (рисунок 5).

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Turbo Prep Auto Model

Result History GroupedModel (Group Models) ExampleSet (Apply Model)

Open in Turbo Prep Auto Model

Row No.	prediction(is...	confidence(No)	confidence(Yes) ↑	proto = udp	proto = arp	proto = tcp	proto = igmp	proto = ospf	proto = sctp
1	Yes	0.008	0.992	0	0	1	0	0	0

Data Statistics

Рисунок 5