Data Science

Домашнее задание

Что нужно сделать

У некого сотового оператора появилась новая услуга, и её предложили небольшому количеству абонентов. Увидели, что часть абонентов к услуге подключилась, а часть нет. Также увидели, что это зависит от некоторых параметров этих абонентов.

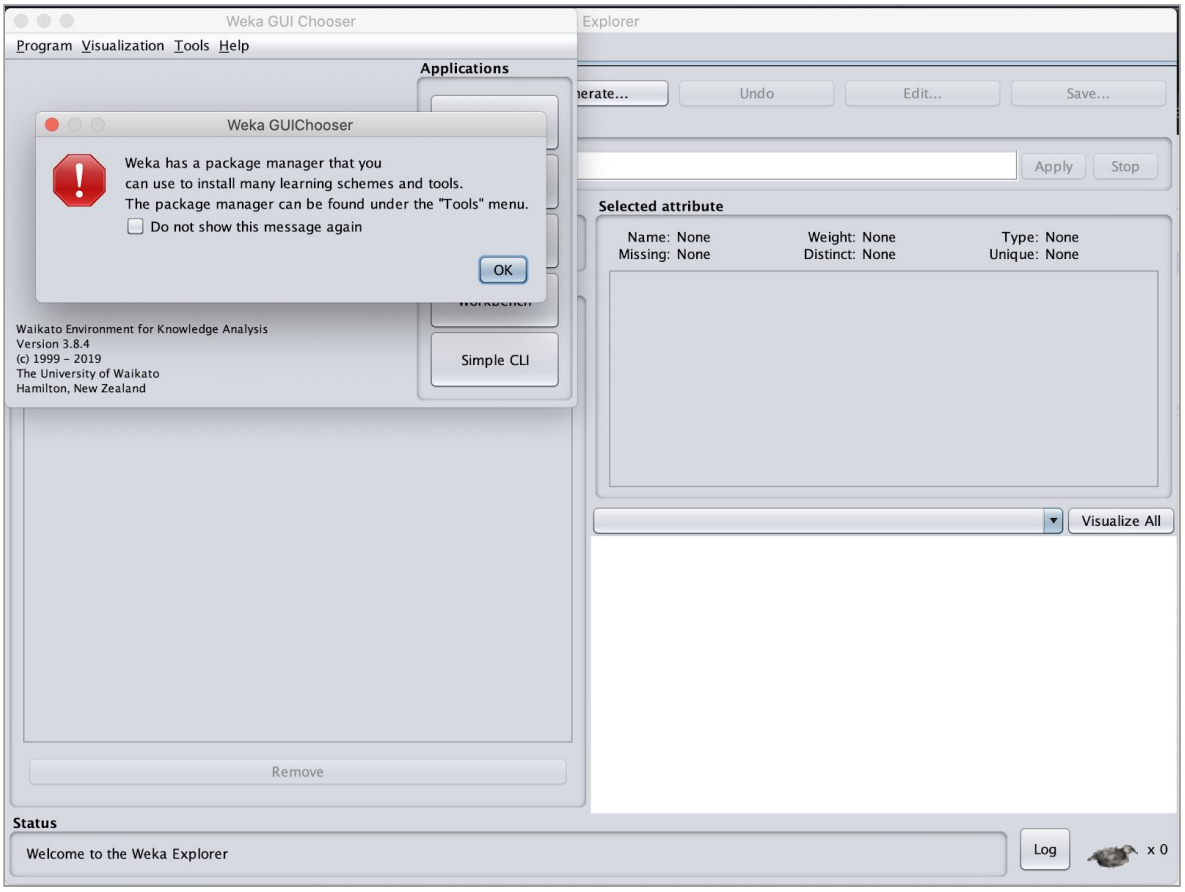
Вам необходимо найти наиболее подходящий и точный алгоритм классификации таких абонентов, чтобы наилучшим образом предсказывать, купят или не купят новую услугу другие абоненты. Это задание аналогично тому, которое демонстрировалось на видео.

- К заданию прилагается файл **data-homework.csv** с данными об абонентах сотовой сети и информацией о покупке ими новой услуги. В файле заданы следующие поля:

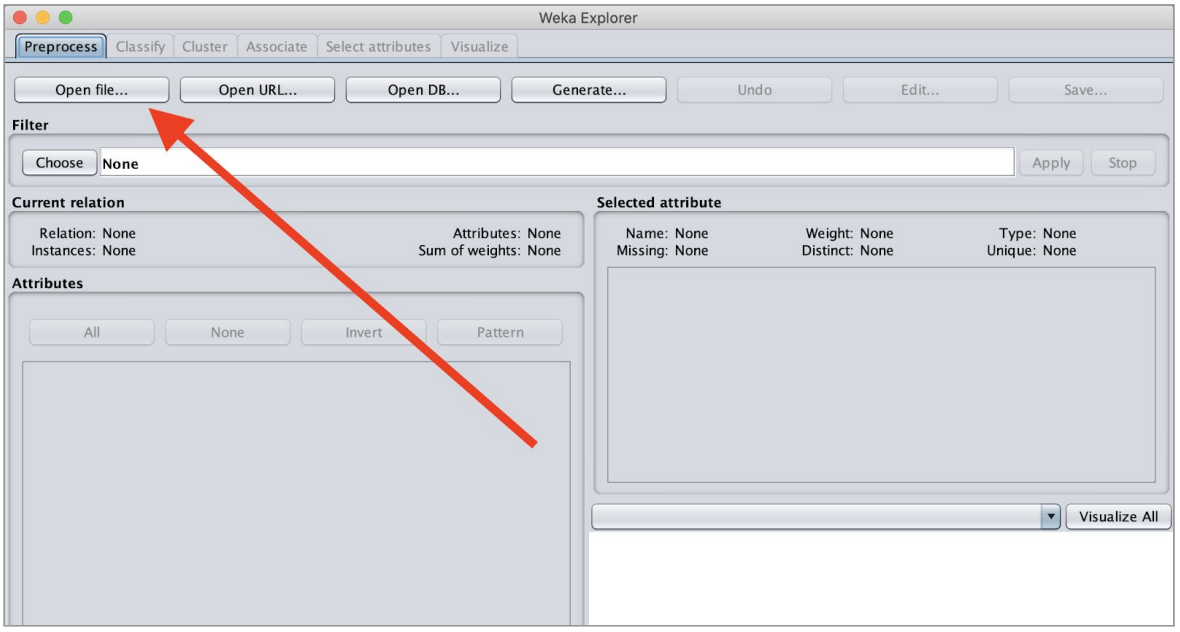
    - **gender** — пол абонента (M — мужской, F — женский);
    - **tenure** — срок владения номером, недель от момента оформления;
    - **service_1** — подключена ли у абонента некая услуга (условно названа «Услугой 1»), имеющая отношение к новой, к которой ему хотят предложить подключиться;
    - **service_2** — подключена ли у абонента некая вторая услуга (условно названа «Услугой 2»), имеющая отношение к новой, к которой ему хотят предложить подключиться;
    - **tariff** — тип тарифа, которым пользуется абонент — предоплатный (PRE) или постоплатый (POST);
    - **expencies** — средний расход в день за последнюю неделю, рублей;
    - **RESULT** — купил ли пользователь новую услугу (YES или NO).

- Скачайте программу Weka и установите её.

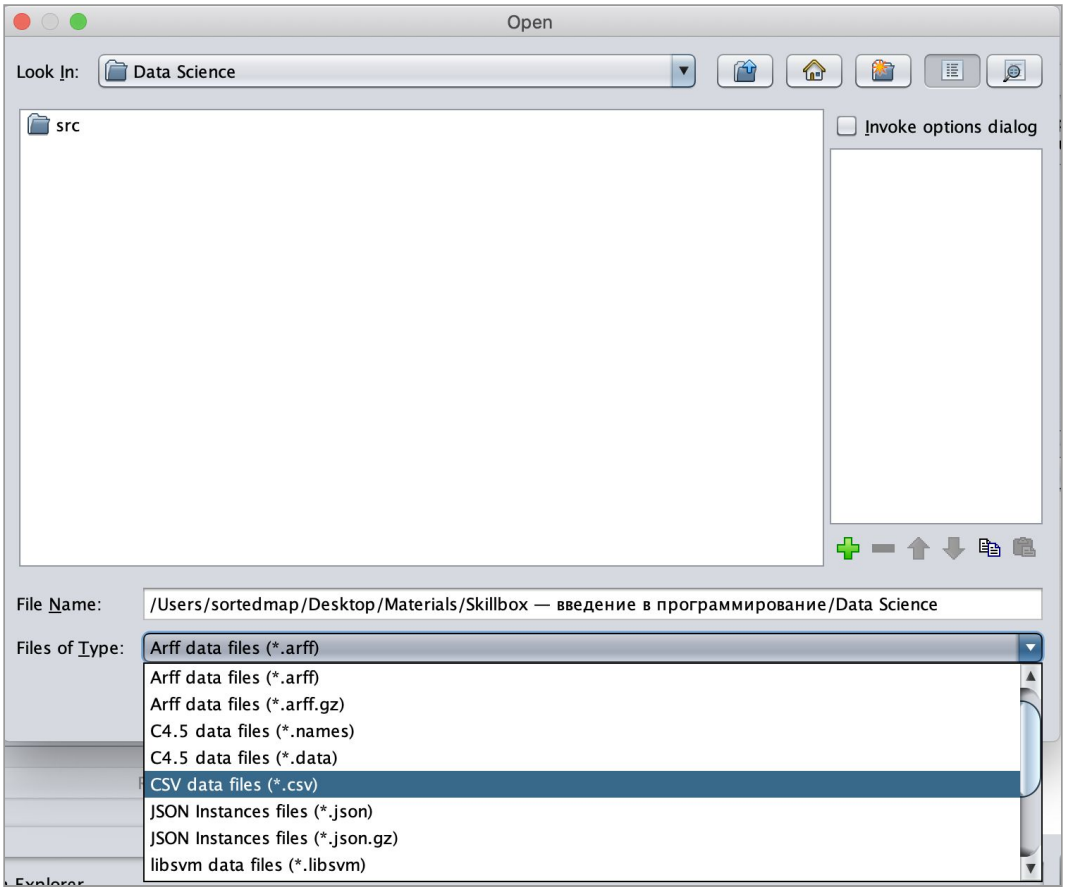
- Запустите программу. Откроются три окна:



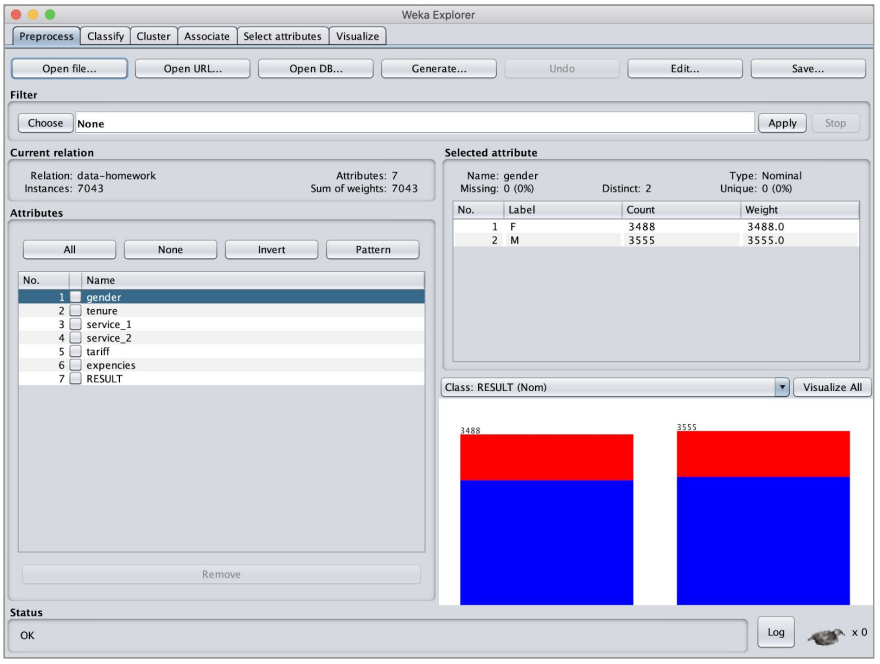
- Закройте два самых верхних маленьких окна, чтобы осталось основное большое. В нём нажмите на кнопку открытия файла **Open file**:



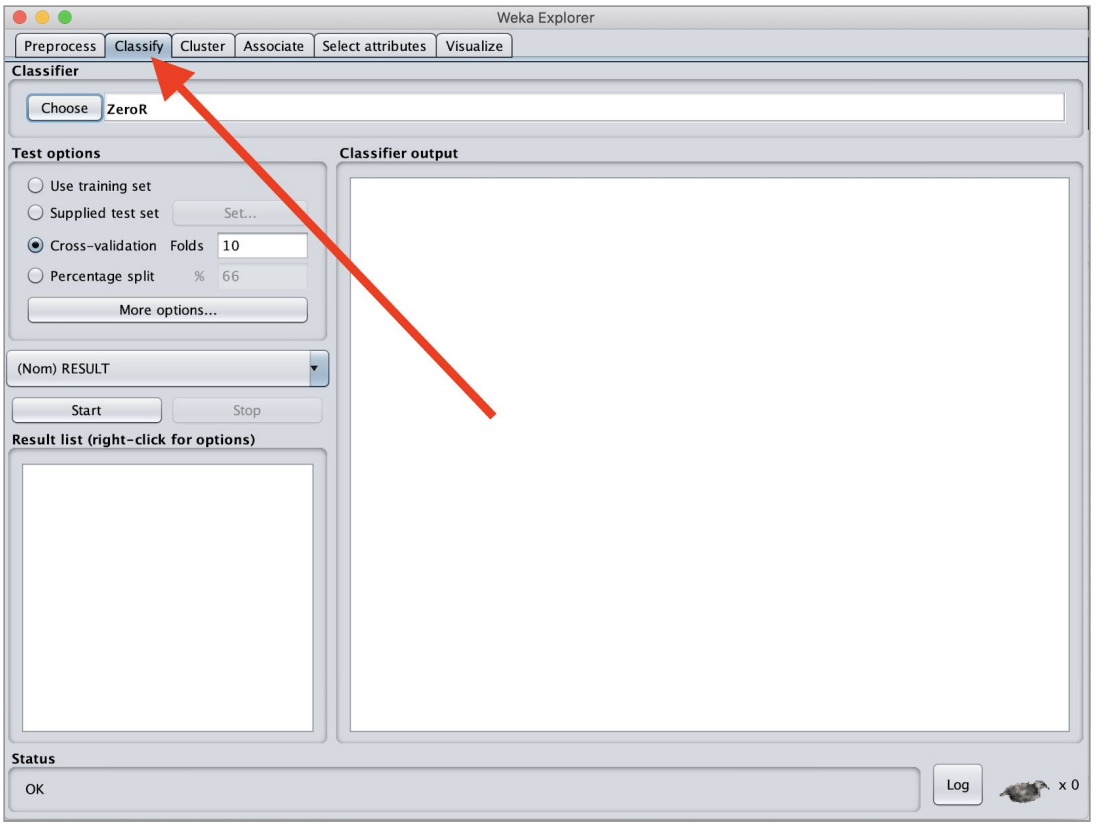
- В открывшемся окне выберите тип файла “CSV data files (*.csv)”:



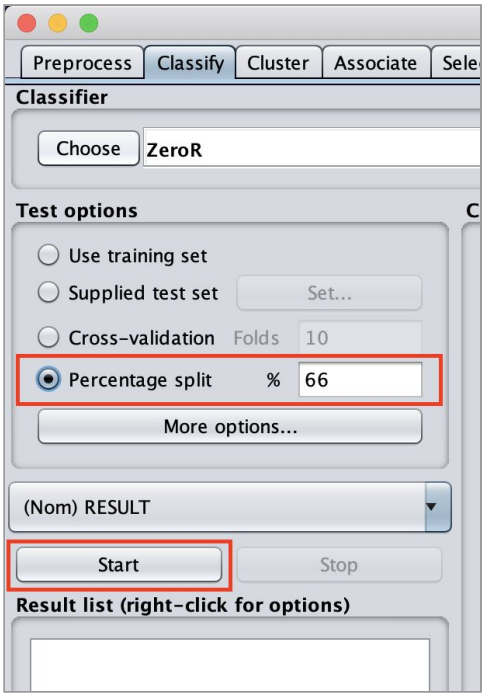
- Выберите файл **data-homework.csv**. Когда откроете его, в окне приложения появится информация о данных в нём:



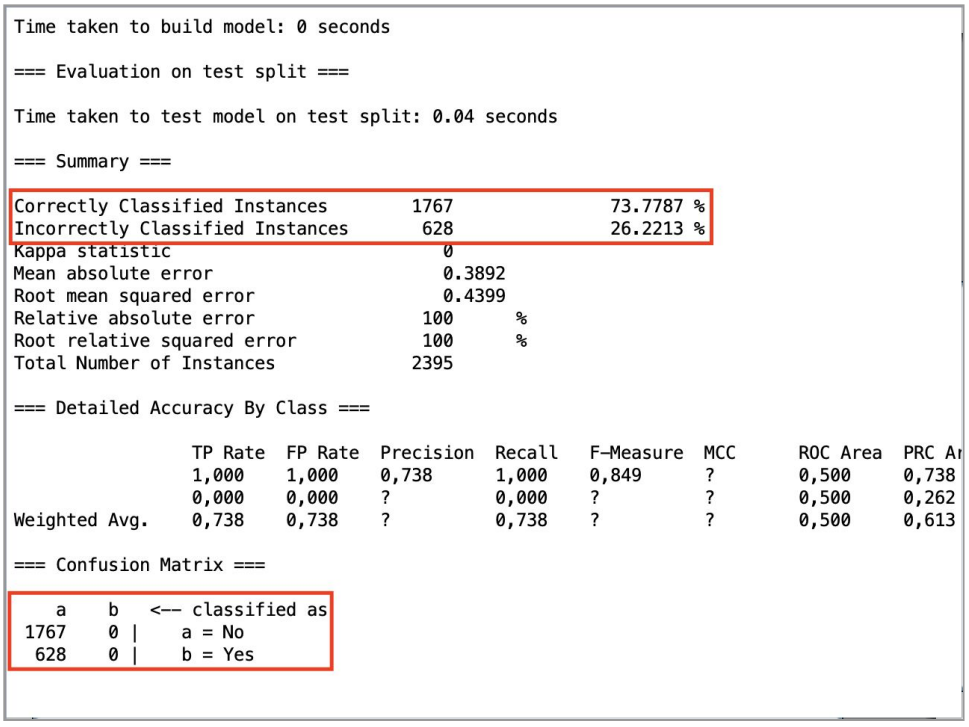
- Откройте вторую вкладку **Classify**:



- В левом поле **Test options** выберите пункт **Percentage split** и нажмите кнопку **Start**:



- По умолчанию выбран алгоритм классификации **ZeroR** (показано выше на рисунках). После недолгих вычислений справа появится информация о точности работы выбранного алгоритма:

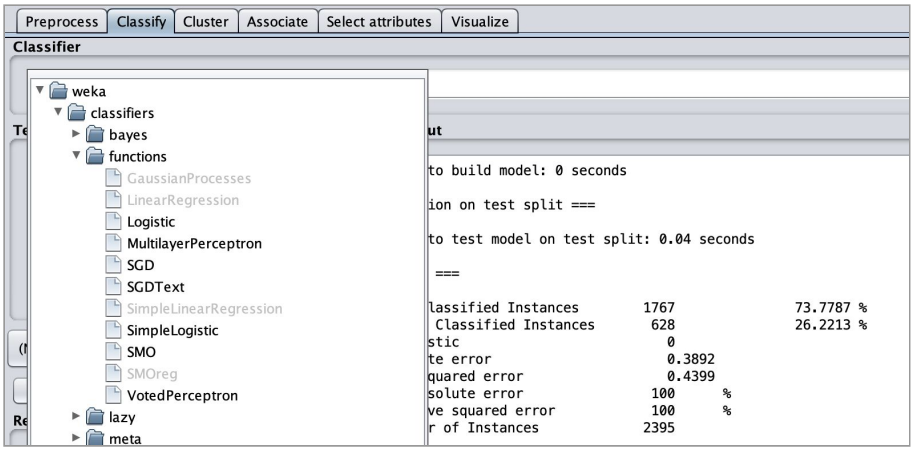


```
Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.04 seconds

=== Summary ===

Correctly Classified Instances        1767               73.7787 %
Incorrectly Classified Instances       628               26.2213 %
Kappa statistic                          0
Mean absolute error                      0.3892
Root mean squared error                  0.4399
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             2395

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Ar
                 1,000    1,000    0,738      1,000    0,849      ?        0,500     0,738
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,262
Weighted Avg.    0,738    0,738    ?          0,738    ?          ?        0,500     0,613

=== Confusion Matrix ===

   a    b   <-- classified as
1767    0 |    a = No
 628    0 |    b = Yes
```

По изображению видно, что точность данного алгоритма минимальна и составляет 73%.

Ниже показано, для каких абонентов результат изначально был **Yes**, а для каких — **No**. И там же видно, что все абоненты отнесены алгоритмом к классу **No**, даже те, которые в исходной таблице находились в классе **Yes**.

- Выбирайте разные алгоритмы с помощью кнопки **Choose** и запускайте их:



- По каждому алгоритму запишите точность, чтобы в найти самый точный. Рекомендуем составить таблицу из наиболее точно работающих алгоритмов с названием каждого алгоритма и его точностью в процентах.

Как проверить результат

Вы можете проверить результат самостоятельно. Важно, чтобы найденный вами алгоритм выдавал точность выше 77%.