Федеральное государственное бюджетное образовательное учреждение высшего образования   
«Саратовский государственный технический университет имени Ю. А. Гагарина»

Кафедра прикладных информационных технологий

**Практическая работа по курсу**

**Интеллектуальные информационные системы и технологии на тему:**

**«Разработка набора данных для построения дерева решений по тематике выбора одежды»**

Выполнил студент б1-ИФСТ-31

Яновский Евгений Валерьевич

Проверил преподаватель

Кошкина Елена Васильевна

Саратов, 2021

Интеллектуальные информационные системы и технологии

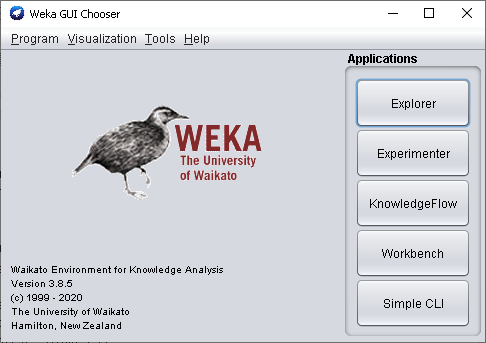
**Тема занятий**

**Практическая работа на тему:**

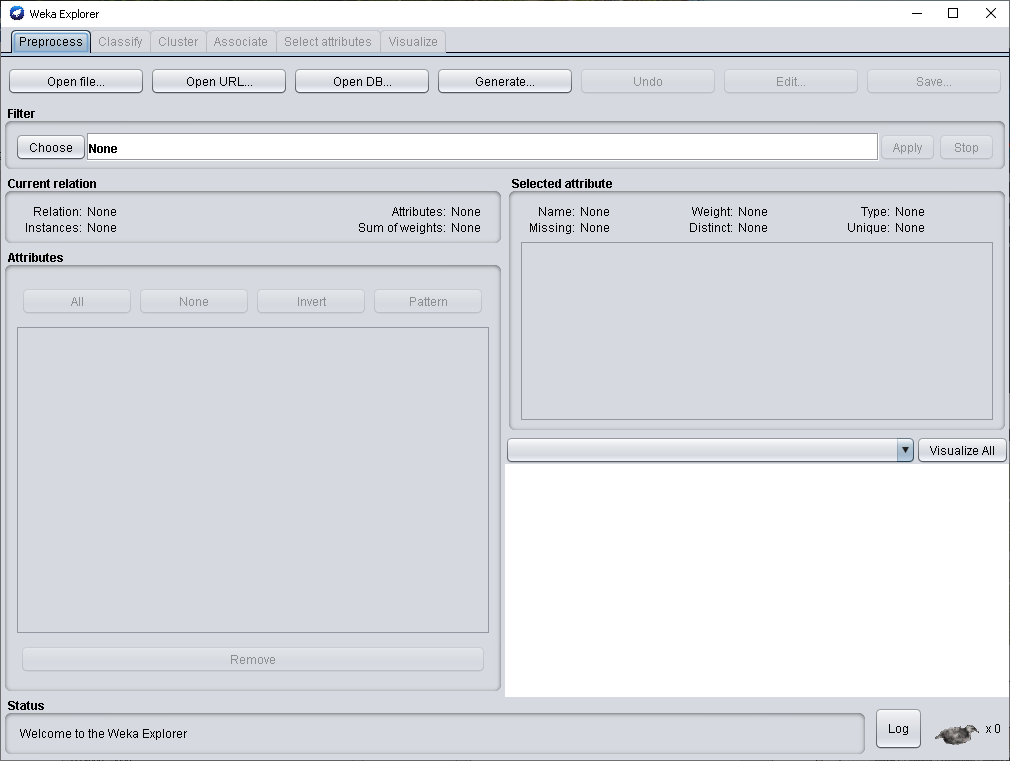
**«Разработка набора данных для построения дерева решений по тематике выбора одежды»**

**Первый этап выполнения работы. Пример построения дерева решений с использованием демонстрационного набора данных**

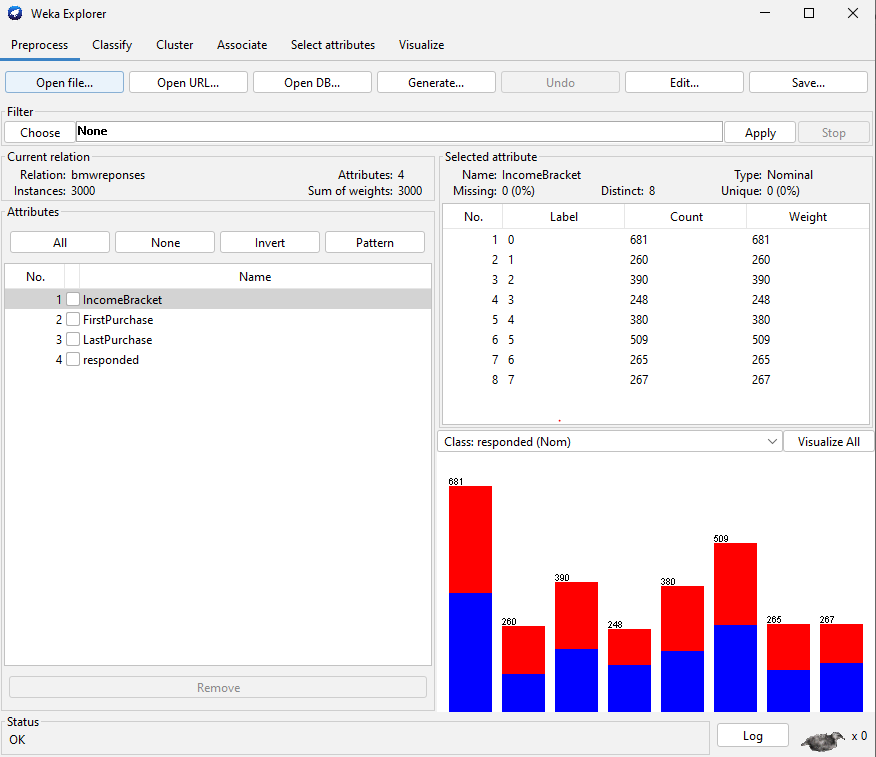
Откройте этот набор данных, выполнив последовательность шагов. Запустите WEKA



Щелкните по кнопке Explorer

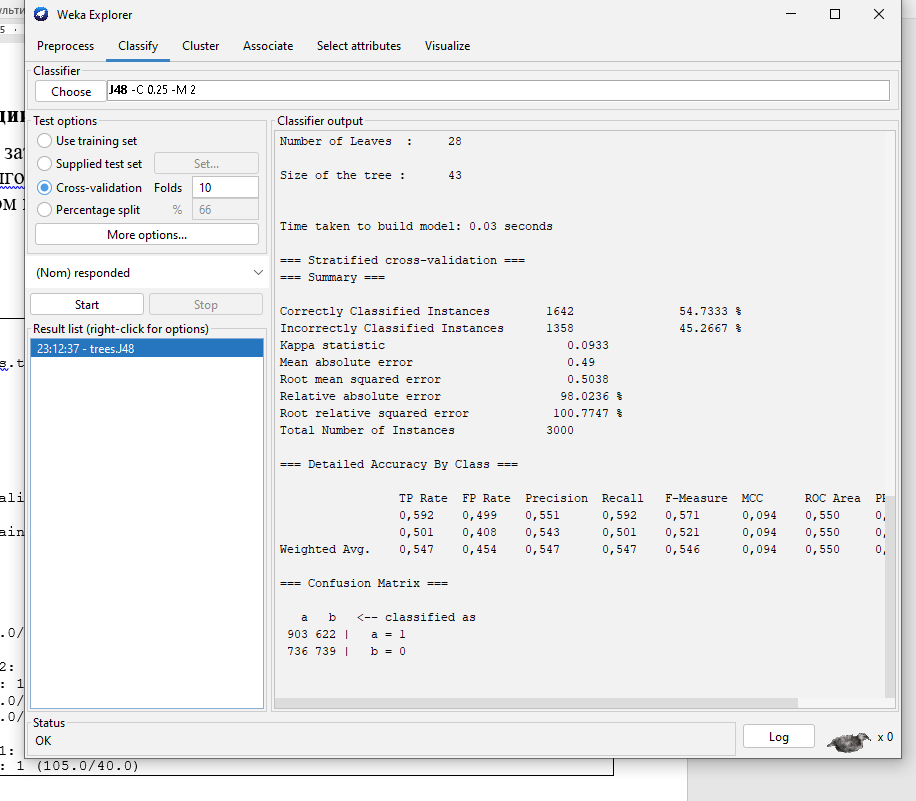


Откройте файл **bmw-training.arff.** “тот файл содержит 3000 кейсов. После загрузки данных экран должен выглядеть примерно так, как показано на рисунке



**Выполнение классификации**

Выберите вкладку Classify, затем выберите (Choose) trees, затем J48. Это один из реализованных в WEKA алгоритмов построения деревьев решений. Затем нажмите **Start** (кнопка подсвечена кружком на рисунке).



Результаты вычислений

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: bmwreponses

Instances: 3000

Attributes: 4

IncomeBracket

FirstPurchase

LastPurchase

responded

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

------------------

FirstPurchase <= 200011

| IncomeBracket = 0: 1 (271.0/114.0)

| IncomeBracket = 1

| | LastPurchase <= 200512: 0 (69.0/21.0)

| | LastPurchase > 200512: 1 (69.0/27.0)

| IncomeBracket = 2: 1 (194.0/84.0)

| IncomeBracket = 3: 1 (109.0/38.0)

| IncomeBracket = 4

| | LastPurchase <= 200511: 0 (54.0/22.0)

| | LastPurchase > 200511: 1 (105.0/40.0)

| IncomeBracket = 5

| | LastPurchase <= 200505

| | | LastPurchase <= 200504: 0 (8.0)

| | | LastPurchase > 200504

| | | | FirstPurchase <= 199712: 1 (2.0)

| | | | FirstPurchase > 199712: 0 (3.0)

| | LastPurchase > 200505: 1 (185.0/78.0)

| IncomeBracket = 6

| | LastPurchase <= 200507

| | | FirstPurchase <= 199812: 0 (8.0)

| | | FirstPurchase > 199812

| | | | FirstPurchase <= 200001: 1 (4.0/1.0)

| | | | FirstPurchase > 200001: 0 (3.0)

| | LastPurchase > 200507: 1 (107.0/43.0)

| IncomeBracket = 7: 1 (115.0/40.0)

FirstPurchase > 200011

| IncomeBracket = 0

| | FirstPurchase <= 200412: 1 (297.0/135.0)

| | FirstPurchase > 200412: 0 (113.0/41.0)

| IncomeBracket = 1: 0 (122.0/51.0)

| IncomeBracket = 2: 0 (196.0/79.0)

| IncomeBracket = 3: 1 (139.0/69.0)

| IncomeBracket = 4: 0 (221.0/98.0)

| IncomeBracket = 5

| | LastPurchase <= 200512: 0 (177.0/77.0)

| | LastPurchase > 200512

| | | FirstPurchase <= 200306: 0 (46.0/17.0)

| | | FirstPurchase > 200306: 1 (88.0/30.0)

| IncomeBracket = 6: 0 (143.0/59.0)

| IncomeBracket = 7

| | LastPurchase <= 200508: 1 (34.0/11.0)

| | LastPurchase > 200508: 0 (118.0/51.0)

Number of Leaves : 28

Size of the tree : 43

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 1642 54.7333 %

Incorrectly Classified Instances 1358 45.2667 %

Kappa statistic 0.0933

Mean absolute error 0.49

Root mean squared error 0.5038

Relative absolute error 98.0236 %

Root relative squared error 100.7747 %

Total Number of Instances 3000

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0,592 0,499 0,551 0,592 0,571 0,094 0,550 0,538 1

0,501 0,408 0,543 0,501 0,521 0,094 0,550 0,537 0

Weighted Avg. 0,547 0,454 0,547 0,547 0,546 0,094 0,550 0,538

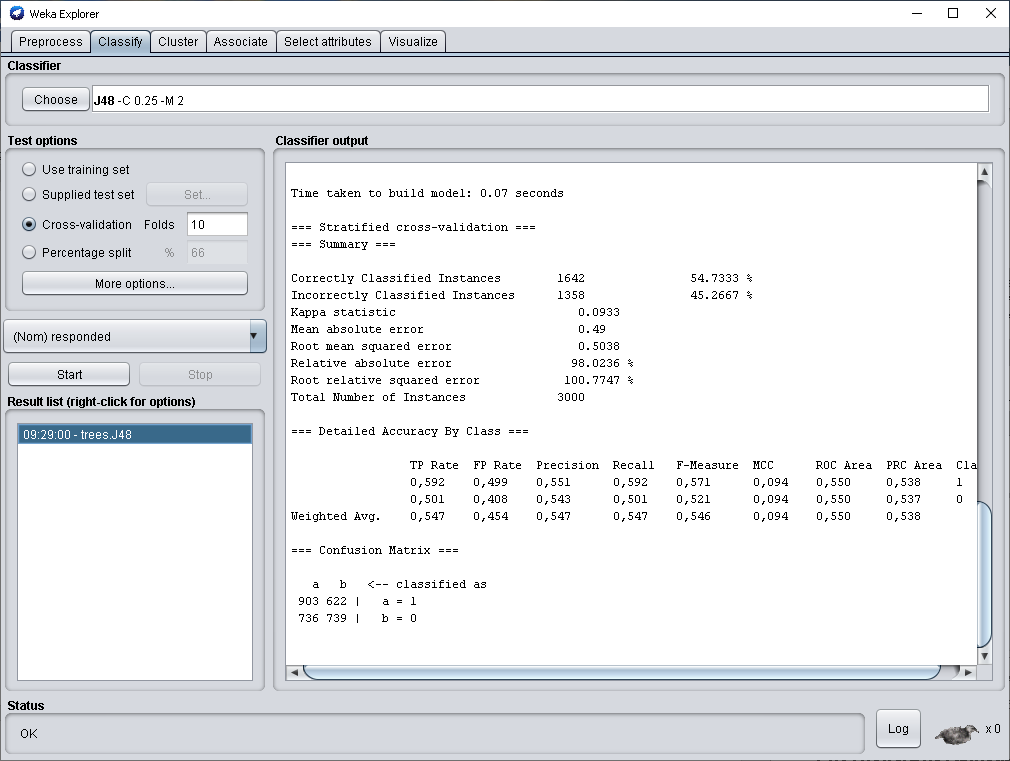
=== Confusion Matrix ===

a b <-- classified as

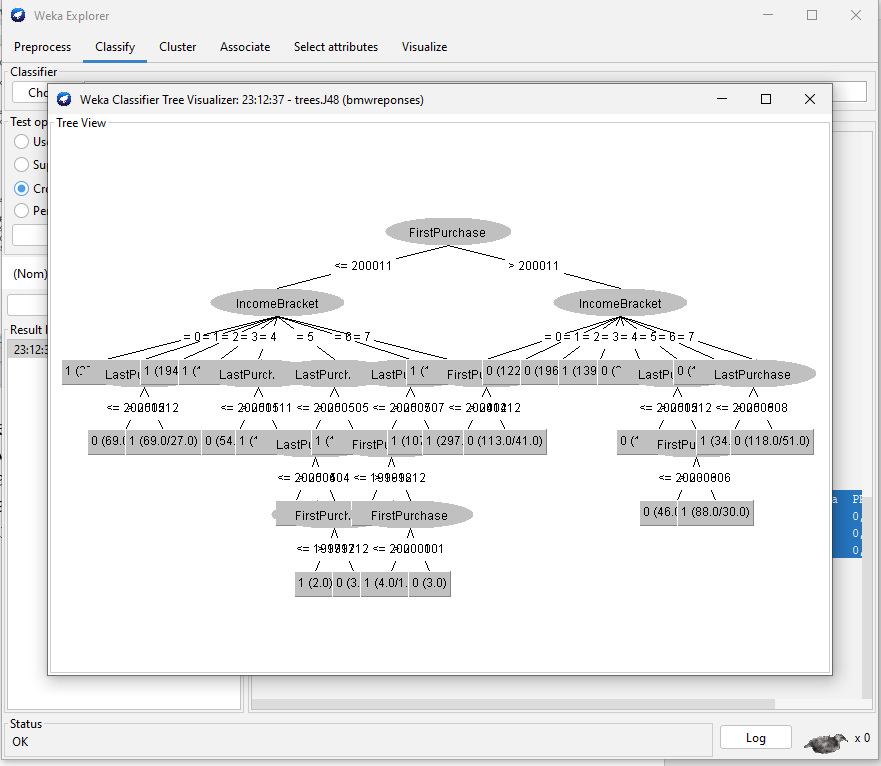
903 622 | a = 1

736 739 | b = 0

Как увидеть дерево решений? Чтобы увидеть это дерево, щелкните правой кнопкой мыши на модели, которую вы только что создали (модель отмечена на рисунке)



. Во всплывающем меню выберите **Visualize tree**. Вы увидите созданное программой дерево классификации. В этом примере визуальное дерево не предоставляется полезной для практического использования Другой способ увидеть это дерево - посмотреть в выходных данных классификатора, где текстовый вывод показывает все дерево с узлами и листьями. Дерево классификации в этом случае представлено на рисунке



Важные цифры, которые следует посмотреть здесь, это те, которые находятся рядом с «Правильно классифицированными экземплярами» (54,7%) и «Неверно классифицированными экземплярами» (45,3%).

**Correctly Classified Instances 1642 54.7333 %**

**Incorrectly Classified Instances 1358 45.2667 %**

Другим важным номером является номер в первой строке столбца «ROC Area» (0.616), мы должны разобрать это важное понятие отдельно

TP Rate FP Rate Precision

Recall F-Measure MCC **ROC Area**  PRC Area Class

0,592 0,499 0,551 0,592 0,571 0,094 0,550 0,538 1

0,501 0,408 0,543 0,501 0,521 0,094 0,550 0,537 0

Weighted Avg. 0,547 0,454 0,547 0,547 0,546 0,094 0,550 0,538

«Матрица» показывает количество ложных срабатываний и ложных отрицаний. В этой матрице ложное срабатывание составляет 622, а ложное отрицание – 736

=== Confusion Matrix ===

a b <-- classified as

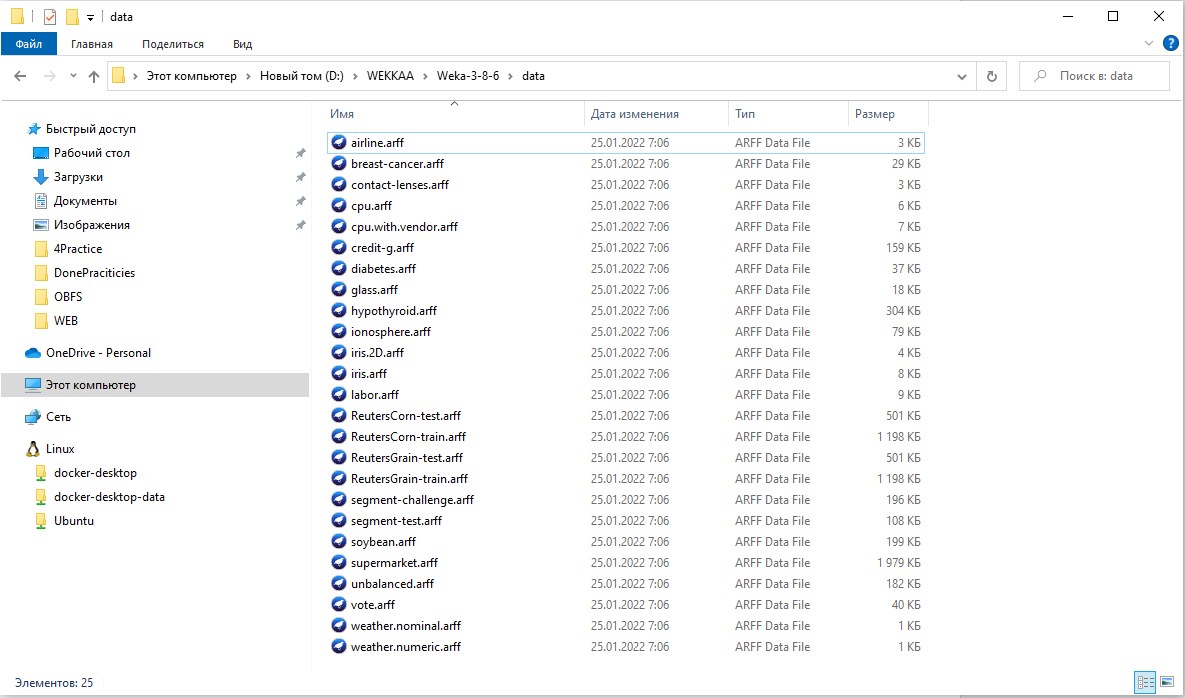
903 622 | a = 1

736 739 | b = 0

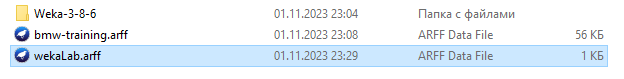
Как оценить полученную модель? Так как показатель точности составляет всего 54,7%, необходимо сделать вывод, что модель представляет небольшую ценность.

**Второй этап выполнения работы. Разработка собственного шаблона данных с небольшим количеством кейсов**

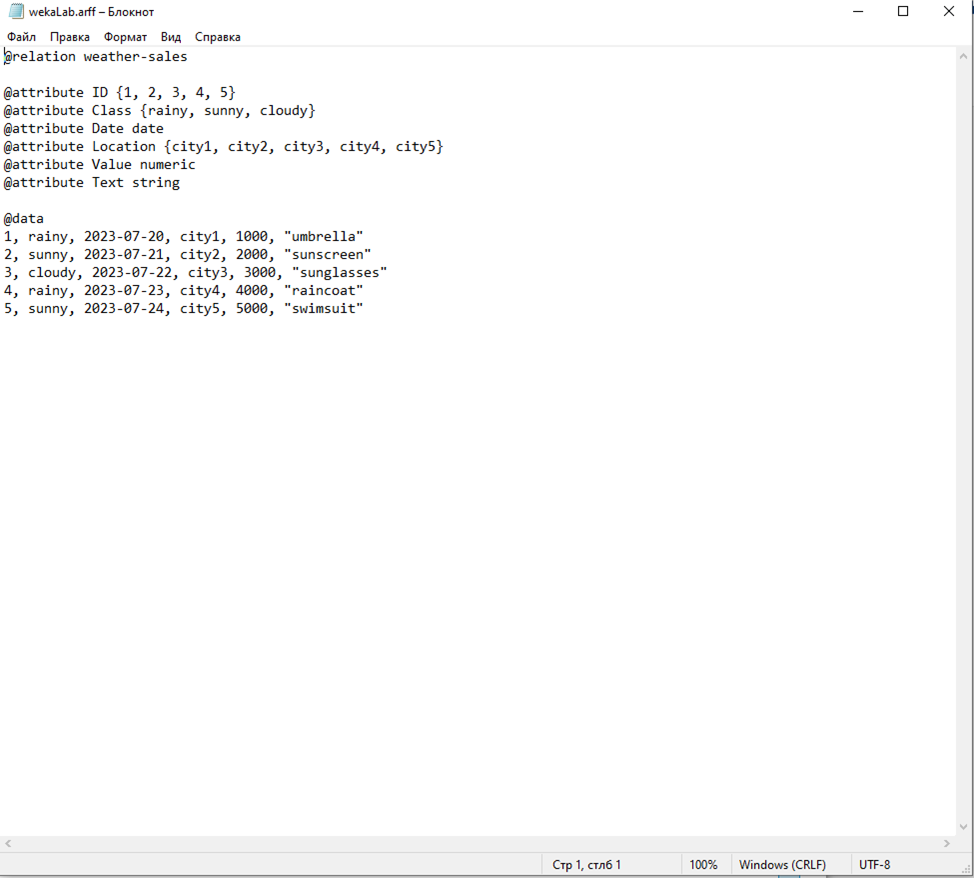
1. Просмотрите в блокноте файлы данных, которые предложены в пакете инсталляции ПО WEKA (Путь установки...:\WEKA DataMining\Weka-3-8-5\data).



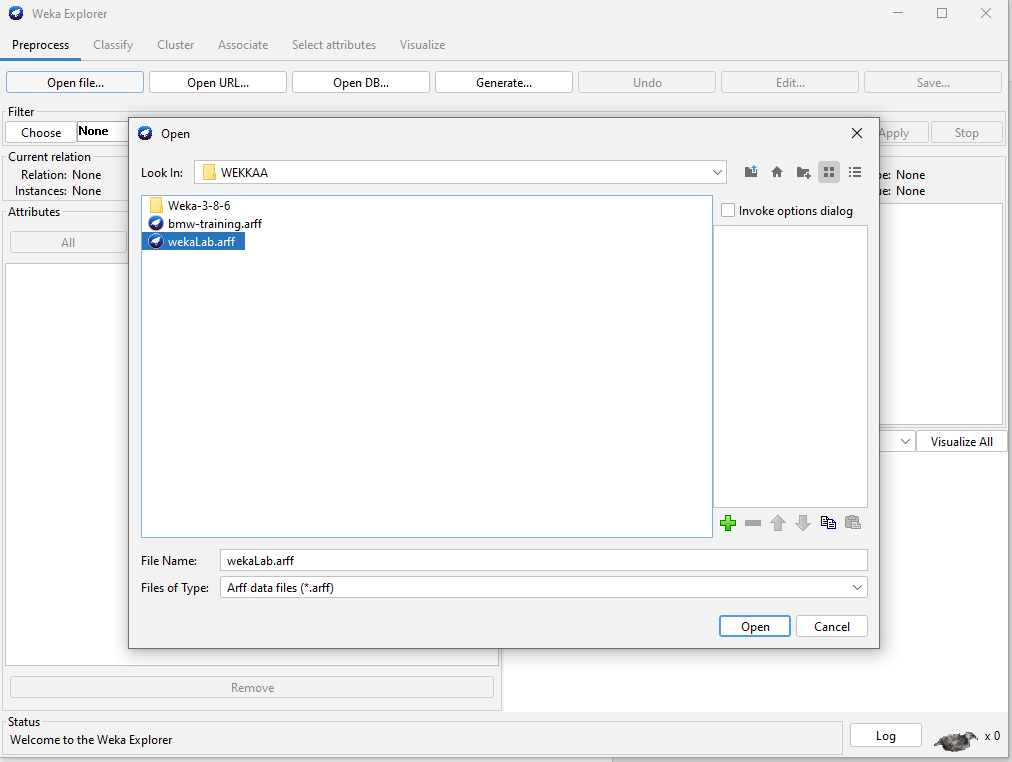
1. Проанализируйте способы представления данных, которые измеряются различными шкалами

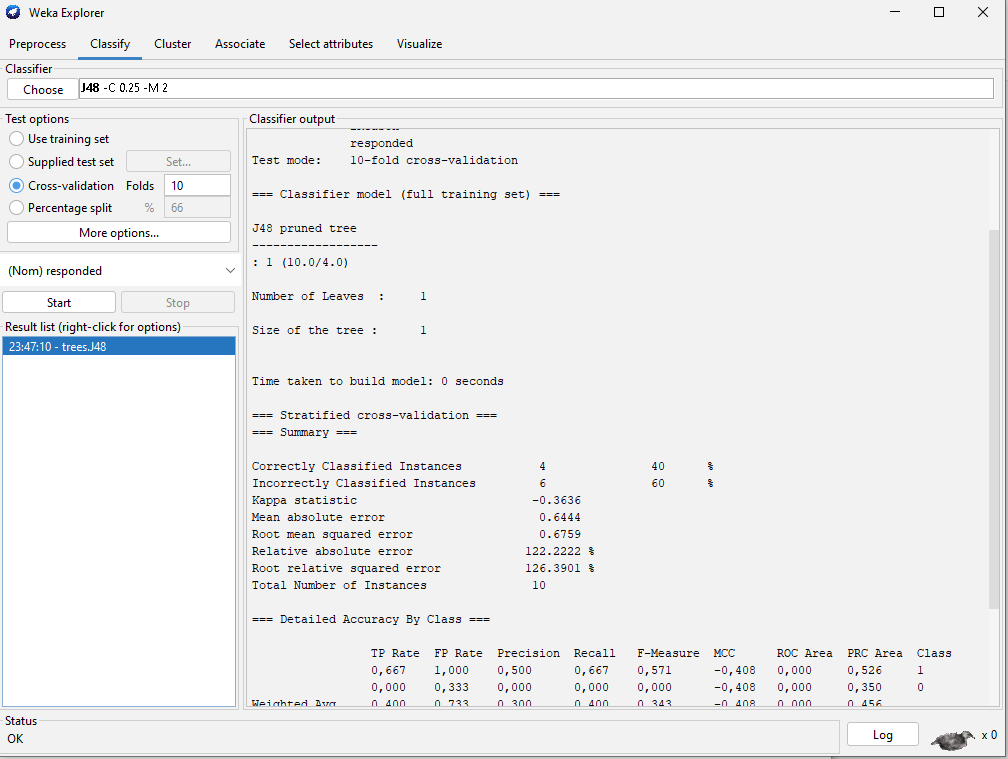


1. Предложите вариант шаблона данных, который использует поля различного типа (номинальные, ординальные) . Эти поля данных могут соответствовать используемым нами ранее условиям (Погода, Ветер, Цель выхода из дома…)
2. Введите пробный набор данных 5-6 строк.



1. Загрузите этот набор данных в программу WEKA





Вывод: в ходе выполнения практической работы я ознакомился с таким инструментальным средством, как Weka, которое используется для интеллектуального анализа данных.