Министерство науки и высшего образования Российской Федерации
Калужский филиал
федерального государственного бюджетного
образовательного учреждения высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(КФ МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ***ИУК «Информатика и управление»***

КАФЕДРА ***ИУК4 «Программное обеспечение ЭВМ, информационные технологии»***

ЛАБОРАТОРНАЯ РАБОТА №9

«Программа интеллектуального анализа данных WEKA»

ДИСЦИПЛИНА: «Технологии анализа данных»

Выполнил: студент гр. ИУК4-82Б _____ (Карельский М.К.)
(Подпись)

Проверил: _____ (Ерохин И.И.)
(Подпись)

Дата сдачи (защиты):

Результаты сдачи (защиты):

- Балльная оценка:
- Оценка:

Калуга, 2024

Цель: формирование и закрепление навыков по работе с системой анализа данных WEKA.

Задачи:

1. Получить общие теоретические сведения о платформе WEKA.
2. Получить навыки по установке WEKA.
3. Ознакомиться с форматами входных данных.
4. Получить навыки решения задач регрессионного анализа, задач классификации и кластеризации.

Задание:

1. Создайте регрессионную модель расчета расхода бензина (MPG - количества миль на галлон), исходя из нескольких параметров автомобиля. Модель учитывает несколько параметров машины – количество цилиндров, рабочий объем двигателя, его мощность, вес автомобиля, время разгона, год выпуска, производителя и марку автомобиля. БД можно найти по следующему адресу: https://cs.nyu.edu/courses/fall00/G22.3033-001/weka/weka-3-0-2/data/auto-mpg.arff
2. Создайте регрессионную модель расчета стоимости машины модели M5. Модель в качестве независимых параметров будет учитывать данные проданных автомобилей и параметры модели M5, а в качестве зависимого параметра – стоимость автомобилей, проданных дилерским центром.
3. Решите задачу Фишера о классификации цветков ириса. БД можно найти по следующему адресу: http://archive.ics.uci.edu/ml/datasets/Iris
4. Решите задачу классификации дней в зависимости от погоды. БД можно найти в папке _адрес_установки\Weka-3-8\data\weather.nominal.arff
5. Решите задачу классификации стекла в зависимости от типа. БД можно найти в папке _адрес_установки\Weka-3-8\data\glass.arff
6. Решите задачу кластеризации цветков ириса. БД можно найти по следующему адресу: http://archive.ics.uci.edu/ml/datasets/Iris
   Решите задачу кластеризации дней в зависимости от погоды. БД можно найти в папке _адрес_установки\Weka-3-8\data\weather.nominal.arff

Результат:

```
Linear Regression Model

mpg =

      1.1903 * cylinders=8,4 +
     -0.0129 * horsepower +
     -0.0006 * weight +
      1.6477 * model_year=71,70,72,74,81,82 +
      1.5989 * model_year=81,82 +
      1.6589 * origin=2,1 +
     30.6249

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correlation coefficient                  0.1642
Mean absolute error                     10.0164
Root mean squared error                 11.5414
Relative absolute error                 98.2735 %
Root relative squared error             98.6428 %
Total Number of Instances             1000
```

Рис. 1. Регрессионная модель

```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 1.7892369364018326

Initial starting points (random):

Cluster 0: 6.4,3.2,4.5,1.5,Iris-versicolor
Cluster 1: 7,3.2,4.7,1.4,Iris-versicolor
Cluster 2: 7.6,3,6.6,2.1,Iris-virginica

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0             1             2
                           (24.0)        (8.0)         (8.0)         (8.0)
================================================================================
sepal_length               5.9167        4.9125        6.3625         6.475
sepal_width                3.0833        3.3875          2.95        2.9125
petal_length               3.9125          1.45        4.5625         5.725
petal_width                 1.225        0.2375         1.425        2.0125
class                 Iris-setosa    Iris-setosa Iris-versicolor  Iris-virginica




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       8 ( 33%)
1       8 ( 33%)
2       8 ( 33%)
```

Рис. 2. Задача Фишера о классификации цветков ириса

```
=== Summary ===

Correctly Classified Instances          10               71.4286 %
Incorrectly Classified Instances         4               28.5714 %
Kappa statistic                          0.5692
Mean absolute error                      0.173
Root mean squared error                  0.2948
Relative absolute error                 39.0795 %
Root relative squared error             62.6814 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,400    0,111    0,667      0,400    0,500      0,337    0,922     0,853     sunny
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     overcast
               0,800    0,333    0,571      0,800    0,667      0,447    0,922     0,885     rainy
Weighted Avg.  0,714    0,159    0,728      0,714    0,702      0,566    0,944     0,906

=== Confusion Matrix ===

 a b c   <-- classified as
 2 0 3 | a = sunny
 0 4 0 | b = overcast
 1 0 4 | c = rainy
```

Рис. 3. Классификация дней

```
=== Summary ===

Correctly Classified Instances         158               73.8318 %
Incorrectly Classified Instances        56               26.1682 %
Kappa statistic                          0.6386
Mean absolute error                      0.0979
Root mean squared error                  0.2229
Relative absolute error                 46.2896 %
Root relative squared error             68.6958 %
Total Number of Instances              214

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,686    0,174    0,658      0,686    0,671      0,507    0,873     0,754     build wind float
               0,750    0,188    0,687      0,750    0,717      0,552    0,865     0,778     build wind non-float
               0,235    0,020    0,500      0,235    0,320      0,306    0,924     0,424     vehic wind float
               ?        0,000    ?          ?        ?          ?        ?         ?         vehic wind non-float
               0,846    0,005    0,917      0,846    0,880      0,873    0,997     0,958     containers
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     tableware
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     headlamps
Weighted Avg.  0,738    0,126    0,732      0,738    0,731      0,617    0,904     0,792

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 48 18  4  0  0  0  0 |  a = build wind float
 18 57  0  0  1  0  0 |  b = build wind non-float
  7  6  4  0  0  0  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  2  0  0 11  0  0 |  e = containers
  0  0  0  0  0  9  0 |  f = tableware
  0  0  0  0  0  0 29 |  g = headlamps
```

Рис. 4. Классификация стекла

Вывод: в ходе выполнения лабораторной работы были получены практические навыки по работе с системой анализа данных WEKA.