MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
**NATIONAL TECHNICAL UNIVERSITY**
**“KHARKIV POLYTECHNICAL INSTITUTE”**

# GUIDELINES FOR LABORATORY WORK

on the topic

«Classification Using Weka (Decision Tree, Naive Bayes)

for students of specialties

121 "Software engineering", 122 "Computer science»

Approved by the University
Editorial and Publishing Council
Protocol No. 2 dated 27.06.2024.

Kharkiv
NTU “KhPI”
2024

Guidelines for laboratory work on the topic «Classification Using Weka (Decision Tree, Naive Bayes) for students of specialties 121 "Software engineering", 122 "Computer science»/ compilers: Yangolenko Olga, Ivaschenko Oksana, Melnyk Karina, Yershova Svetlana, Lutenko Irina – Kharkiv : NTU “KhPI. – 2024. – 11 p.

Compilers: Yangolenko Olga, Ivaschenko Oksana, Melnyk Karina, Yershova Svetlana, Lutenko Irina

Reviewer prof. Moskalenko Valentina

Department of Software Engineering and Management Intelligent Technologies

# CONTENTS

# INTRODUCTION

**Aim:** To fulfil classification (Decision Tree, Naive BayeS) using Weka mining tool.

**Tools/ Apparatus**: Weka mining tool..

# 1 THEORETIC WORK MATERIAL

## 1.1 Theory

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time.

In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case.

Classifications are discrete and do not imply order. Continuous, floating point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.

The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating.

In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown.

Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is

typically divided into two data sets: one for building the model; the other for testing the model.

Scoring a classification model results in class assignments and probabilities for each case. For example, a model that classifies customers as low, medium, or high value would also predict the probability of each classification for each customer.

Classification has many applications in customer segmentation, business modeling, marketing, credit analysis, and biomedical and drug response modeling.

Different Classification Algorithms Oracle Data Mining provides the following algorithms for classification:

1 Decision Tree. Decision trees automatically generate rules, which are conditional statements that reveal the logic used to build the tree.

2 Naive Bayes. Naive Bayes uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data.

### 1.2 Formatting Documents

The report is prepared in the text editor MS Word [2]. You need to open the template at the link [3] and save the document in .docx format with the appropriate name, for example, Report_1_DMT_Ivanov_KN_N424.docx. This template is a set of styles that should be used when creating reports on laboratory work, this template also explains the cases of using one or another style, formatting and design of various elements of the report. Before making a report, you should study the requirements for making reports, which are given in the template and below the text.

It is considered that the student has basic skills in MS Word (or similar editors, such as OpenOffice.org Writer).

Document formatting refers to the way a document is laid out on the page—the way it looks and is visually organized—and it addresses things like font selection, font size and presentation (like bold or italics), spacing, margins, alignment, columns, indentation, and lists. Basically, the mechanics of how the

words appear on the page. A well formatting document is consistent, correct (in terms of meeting any stated requirements), and easy to read [4].

Basic formatting standards include:

1 The report is made on sheets of A4 printing paper (297 mm x 210 mm). The margins must be: left, bottom and top - not less than 20 mm, right - not less than 10 mm.

2 14 pt. font in a consistent style throughout, including section headings and subsection headings, headers, footers, and visual labels. Font of notes, text elements in table can be 12 pt.

3 A standard, professional font is Times New Roman.

4 1.5 line spacing, with 1.25 indentation on the first line of the paragraph

5 Body text is aligned with both margins

6 Page numbers at upper right corner (Arabic numerals). The number is not placed on the title page, which is the first page of the document, but it is included in the general numbering (a sample of the title page for the report on laboratory work is given in Appendix A).

Documents usually have some form of “logical structure”: division into chapters, sections, sub-sections etc. to organize its content. Each element of structure has corresponding heading, and heading of each part of document has own formatting standards:

1 Heading of Section. New section begins with a new page (page break at the end of the previous part of the text). Formatting of heading of section: Times New Roman 14 pt., bold, Capital, centre text, 1.5 line spacing, after heading 21 points.

2 Heading of Subsection is separated from the text body by blank line. Formatting of heading of subsection: Times New Roman 14 pt., bold, justified text, 1.25 indentation.

3 Heading of Item is NOT separated by line from the text body. Formatting of heading of item:Times New Roman 14 pt., bold, justified text, 1.25 indentation.

To automatically generate table of contents, you must first configure the styles of elements of structure.

# 2 TASKS OF LABORATORY WORK

**Procedure:**

1 Open Weka GUI Chooser.

2 Select EXPLORER present in Applications. Select Preprocess Tab.

3 Go to OPEN file and choose file according to Assignment of datasets.xlsx.

4 Fulfil preprocessing of data whether, if it is necessary and save file.

5 Go to Classify tab.

6 Select tree j48 (the implemented C4.5 algorithm). Set appropriate parameters if it is necessary.

7 Select the attribute to use as a class.

8 Experiment with Test options to obtain the best model.

9 Click Start .

10 You can see the output details in the Classifier output.

11 Right click on the result list and select" visualize tree " option .

12 Present an argument for selecting attribute as a class and the best techniques that you can use to evaluate the performance of an algorithm (Test options). Describe results of classification.

13 Select NaiveBayes. Set appropriate parameters if it is necessary. Do steps 7-10, 12.

14 Make conclusions.

# REFERENCES

1 Papakyriakou, Dimitrios & Barbounakis, Ioannis. (2022). Data Mining Methods: A Review. International Journal of Computer Applications. 183. 5-19. 10.5120/ijca2022921884

2 Word help & learning // https://support.microsoft.com/en-us/word?ui=en-us&rs=en-us&ad=us, 07.06.2024

3 Templates for reports on laboratory work // https://iiiii.sharepoint.com/:f:/s/Profs.PIITU/ErRwourAhj1AjM3szMwHEsgBcqyrl0_Ik8_xHIJp2A-lLQ?e=8nfxUH, 01.06.2024

4 Chapter 8. Formatting Documents // https://ohiostate.pressbooks.pub/feptechcomm/chapter/8-formatting/, 07.06.2024

# APPENDIX A

## Sample of design of the title page

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

**NATIONAL TECHNICAL UNIVERSITY**
**"KHARKIV POLYTECHNICAL INSTITUTE"**

Institute (faculty) of Computer Sciences and Software Engineering
Department of Software Engineering and Management Information Technology
Program Subject Area 122 Computer science
Educational Computer science and intelligent systems
Specialization

**LABORATORY WORK №3 on course**

**« Data Mining Tools »**

Laboratory work subject Classification Using Weka (Decision Tree, Naive Bayes)

Executed by student 5 year, group KN-424

Pavel ZHERZHERUNOV
(signature, surname and name)

Checked by Oksana IVASHCHENKO
(signature, surname and name)

Kharkiv 2024

Educational edition

Guidelines for laboratory work

on the topic

«Classification Using Weka (Decision Tree, Naive Bayes)

for students of specialties

121 "Software engineering", 122 "Computer science»

compilers:

YANGOLENKO Olga

IVASCHENKO Oksana

MELNYK Karina

YERSHOVA Svetlana

LUTENKO Irina

Responsible for the publication of prof. GAMAYUN Igor

The work was recommended for publication by prof. BEZMENOV Mykola

In the author's edition

Plan 2024 p., pos. 575

Submitted for publication 2025. Font Times New Roman

---

Publishing Centre of NTU "KhPI", 2 Kyrpichova St., Kharkiv, 61002
Certificate of the subject of publishing business DK № 5478 dated 21.08.2017

---

Electronic version