AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING

Computer Engineering and Software Systems Program
i-Credit Hours Engineering Programs (CHEP)

Data Mining and Business Intelligence
Course Code: CSE385
Semester: Spring 2020
End Term Assesment
Research Paper

Description and Instructions

Instructor(s)	**Dr. Mahmoud Mounir**
Research Title:	**Using Decision Trees for Classification in Data Mining**

General Instructions

- Research Paper replaces the final written exam
- The research is individual
- Each part/component of the research has a weight
- Research submission is online in electronic format
- Due date will be the date announced by faculty/program administration
- Plagiarism checking will be applied. Research is subject to rejection in such case

1. Research Description

1.1. Research short description

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a dataset and transform the information into a comprehensible structure for further use.

This research is designed to allow you to deal with these concepts in addition to the application of data mining concepts in real life activities.

A decision tree is a map of the possible outcomes of a series of related choices. It allows an individual or organization to weigh possible actions against one another based on their costs, probabilities, and benefits. A decision tree typically starts with a single node, which branches into possible outcomes. Each of those outcomes leads to additional nodes, which branch off into other possibilities. This gives it a treelike shape .

A decision tree can also be used to help build automated predictive models, which have applications in machine learning, data mining, and statistics. Known as decision tree learning, this method takes into account observations about an item to predict that item's value. In these decision trees, nodes represent data rather than decisions. This type of tree is also known as a classification tree. Each branch contains a set of attributes, or classification rules, that are associated with a particular class label, which is found at the end of the branch. Decision trees can be used in biomedical engineering, Financial analysis, etc. Decision trees can handle both categorical and numerical data.

1.2. Research requirements

In this research, you should introduce the Decision Trees concept for classification and how it is used in machine learning. After finishing this research, you **must** cover the following items:

- What Decision Tree is?
- Mathematical formulation of Decision Trees.
- What are the terms in the Decision Trees mean and the intuitions behind them?
- Explain the ID3 and C4.5 methods to construct the decision tress how they are related to probability?
- Working numerical example of how Decision Trees are built.

- Working numerical example of how Decision Trees are used in classification.
- Applications of Decision Trees in real life.
    - Collecting a real dataset to be used for classification using Decision Trees (**you can get it from the internet from Kaggle or Google datasets or any other source you prefer**).
    - Exploratory data analysis for the collected dataset (**you may write a code to help you in this part using Python or R language or any other language you prefer**).
        - Analyze raw data using appropriate graphical and numerical procedures.
        - Describe Shape, Outliers, Center, and Spread of datasets in the context of your research.
        - Include appropriate graphical displays and numeric summaries.
    - **Using Weka:**
        - Perform outlier analysis using DBSCAN.
        - Perform cluster analysis to your data using a suitable technique of clustering.
        - Use Decision Trees to classify the data based on some aspects in the context of your collected real dataset (**or you may write a code to help you in this part using Python or R language or any other language you prefer if you don't prefer to use Weka**).

1.3. Research deliverables

You are required to **submit**

- A research paper document as a Microsoft WORD document following the instructions in the guidance to submission section in this document. Your

research must contain a cloud link to the source file of your implementation code **(See item number 10 in the Guidance to submission section).**

Your Report should consist of a summary of your research and survey as well as your personal conclusions. The goal is to enlighten the reader with words, numeric summaries, and appropriate graphs. Use the following format:

- Title Page
- Abstract
- Introduction
    - The introduction has three goals:
        a) To introduce the topic of the research;
        b) To present your research (which is to say the particular approach or argument the research will make);
        c) To tell the reader how the paper will be structured.
- Body
    - In the body of the research you will present the evidence and analysis that will substantiate your research. It is essential that the body of the paper be developed in a logical and orderly fashion following the preview that you presented in the introduction. The overall goal of this section is to develop your analysis and defend your argument – it is the main part of the research paper.
    - This logically ordered body of the paper will consist of a series of paragraphs. Each paragraph should develop one central theme that helps you further your argument. Introduce this theme in a topic sentence; expand on the theme through the use of evidence or examples; and analyze the evidence to show how it contributes to the specific point you are making in the paragraph and to the research as a

whole. Paragraphs should consist of several sentences rather than one, long sentence.

- Conclusion
  - The conclusion is designed to bring together your paper main points and to reassert or emphasize the strength of the research. A conclusion is more than a summary, in that it is important to indicate why there is merit to your research – what has been shown as a result of your investigation or exploration of the topic.
- References

1.4. Research references

The following references may be useful for your research paper:

- Han, J., "Data mining: Concepts and techniques", third edition, 2012, Waltham, Mass.: Morgan Kaufmann Publishers.
- Gareth James, et al., "An Introduction to Statistical Learning with Applications in R", Springer.
- https://www.kaggle.com/
- https://datasetsearch.research.google.com/

1.5. Guidance to submission

1. This research should be prepared using Microsoft Word following the required structure of IEEE papers.
2. Research to be submitted in (.DOC, .DOCX) format ONLY.
3. It should be individual work.
4. Research words are recommended to be within the range of 4000 to 5000 words (A4 size, font size 12, 1.5 line spacing).
5. Research to be Grammarly checked by student.

6. Used in research references must be at least 5 references and, at least, one of them must be an updated reference (after 2010).
7. Researches might undergo similarity checks for allocating plagiarism cases.
8. Submitted file size limit will be 20mb.
9. Only one file is allowed to be uploaded.
10. In case of having some attachments such source codes, or other similar materials, you should add these attachments to a separate virtual cloud and insert their links within the research body itself in addition to an attachments list by the end of the research.

Best Wishes
Dr. Mahmoud Mounir