CSE521 DATA MINING TEST 1 SUMMER 2021
(70pts + 5 extra credit)

TEAM NUMBER: Leader **NAME** and **ID:**

Team members Leader **NAMES** and **IDs:**

1.

2.

FORMAT OF SUBMISSION

1. The TEAM LEADER submits **ONE PDF** file for the TEAM

2. Name your PDF file as $< TEAMLEADERID >$.pdf

Example: 11384578.pdf

All team members receive the same grade.

PROJECT DESCRIPTION

PROJECT GOAL

The main goal of the project is to use the Project Data and Internet based Classification Tools to build **descriptive** Decision Tree classifiers and **non-descriptive** Neural Network classifiers and to compare obtained results and to compare the functionality of the tools used.

PROJECT TOOLS

TOOL WEKA: https://www.cs.waikato.ac.nz/ml/weka/

LIST OF TOOLS https://www.softwaretestinghelp.com/data-mining-tools/

Use WEKA for the Project and compare WEKA functionality and WEKA's results with one other tool.

You can use a tool from the LIST OF TOOLS or any other tool of your choice.

You will get 5 EXTRA POINTS for using more than one secondary tool.

PROJECT DATA is provided on the course web page.

This is a real life classification data with TYPE DE ROCHE (Rock Type) as a CLASS attribute.

There are 98 records with 48 attributes and 6 classes.

Classes are:

C1 : R. Carbonatees AND R. Carbonatees impures

C2 : Pyrate

C3 : Charcopyrite

C4 : Galene

C5 : Spahlerite

C6 : Sediments terrigenes

Most important attributes (as determined by the expert) are: **S, Zn, Pb, Cu, CaO+MgO, CaO, MgO, Fe2O3**

This is a real life experimental data and it contains a lot of missing data (no value) and not ready to use without some preprocessing

PROJECT STEPS

Project follows the following steps of **DM Process** to build the classifiers.

S1: Data Preparation (2pts)

Use attributes selection, cleaning the data, filling the missing values, etc... operations to build your Project DATA - **PD**

Create Project EXPERT DATA - **PED** as a subset of Project DATA - **PD** using only the most important attributes **S, Zn, Pb, Cu, CaO+MgO, CaO, MgO, Fe2O3** as determined by the expert.

S2: Data preprocessing (3pts)

1. For the Decision Trees **Descriptive Classifier** you use your chosen method of the **discretization** of Project DATA - **PD** creating a set **PD1** of data with no more than 4 values (bins) for each attribute. Different attributes do not need to have the same number of values (bins) and you do not need to use the same discretization methods for all of them. Describe which discretization method (must use at least two) you used for each attributes.

Create the EXPERT DATA - **PED1** from the **PD1** by using only the most important attributes **S, Zn, Pb, Cu, CaO+MgO, CaO, MgO, Fe2O3**.

2. For the Neural Network **Non - Descriptive Classifiers** use your chosen method of normalization of the Project DATA - **PD** and Project EXPERT DATA - **PED**. Specify which.

S3: Building Classifiers (25pts)

1. For the Decision Tree **Descriptive Classifiers** use the sets of data **PD**, **PED** for **Experiments 1, 2** and **PD1**, **PED1** for **Experiment 3**.

2. For the Neural Network **Non- Descriptive Classifiers** use the normalized sets of data **PD**, **PED**.

Use at least TWO different Network TOPOLOGIES; one can be the Tool default one, and the other designed by you. Compare results.

PROJECT TOOLS

Use WEKA and a TOOL of your choice to build classifiers conducting **Experiments 1, 2** and only WEKA for **Experiment 3**.

PROJECT Classification STEPS for **EXPERIMENTS 1 and 2**

.1 (5pts)

For each TOOL and each set of DATA and each EXPERIMENT describe shortly the resulting Classifier and the parameters (testing method, network topology, number of epochs to converge... type of Decision Tree, write set of discriminant rules. etc..) used to build it.

2. (5pts)

Compare the resulting **Descriptive Classifiers** with each other testing methods, Rules Accuracy and Predictive Accuracy

3. (5pts)

Compare the resulting **Non Descriptive Classifiers** with each other on topology used testing method and predictive accuracy

4. (5pts)

Compare each **Descriptive Classifier** with the resulting **Non-Descriptive Classifier** on Predictive Accuracy only.

5. (5pts)

Describe shortly the TOOLS you use.

Describe their similarity and differences for the tasks you used them for.

Describe your experience with the TOOLS and your preferences for one or the other regarding different tasks.

EXPERIMENTS 1- 3

Experiment 1

Use the **PD** and **PED** data to perform the **full classification**, i.e. to build

Decision Tree classifier, and Neural Network Classifier for all classes **C1- C6** simultaneously.

Experiment 2

Use the **PD** and **PED** data to perform the **contrast classification** for the class **C1**

i.e. to build a Decision Tree classifier, and Neural Network Classifiers contrasting the class **C1** with a class **notC1** that contains other classes.

Experiment 3 (5 extra points)

Use **PD1** and **PED1** data and **only** WEKA Decision Tree to build two kind of **contrast classifiers** for the class **C1** for each set **PD1**, **PED1** of data.

1. Classifier **M1** with branches grown for each known value of each attribute.

2. Classifier **M2** with only binary branches branches

Write obtained classifiers as sets of discriminant rules in predicate form.

Compare the classifiers built out of **PD1** and **PED1** data on base of predictive accuracy and rules accuracy.