Diagnosis of Breast Cancer using Decision Tree Data Mining Technique

Course:	**INFO-3135**
Professor:	**Daniel Maclam**
Project:	**Project #2**
Due Date:	**December 10th @ 11:59PM**
Submitting:	**Please see the last page for instructions.**

Problem Description

Build a binary tree to help detect the type of Breast Cancer tumors.

Benign tumors have not yet spread to other parts of the body whereas Malignant tumors have already spread to other parts of the body. A study has been published that uses a Binary Decision tree to determine if a patient has benign or malignant breast cancer tumors to within a 94.5% accuracy.

We will build that binary decision tree as shown in Figure 13, that runs on a dataset with the information as shown in Table 3.

The data that we would like to run on will be in the csv file format in Figure 12, without the diagnosis (column K in the CSV).

You are to read in the data of patients from file, run the data through the binary decision tree and output the results as shown with the diagnosis Benign or Malignant (Figure 12).

Store the results in a new file named 'results.csv' following the same format as the input file, replacing the last column (0) with 2 or 4 based on the result for that row.

After processing all the rows in the file print a summary showing the following:

- Total Patients Processed
- Total Benign
- Total Malignant

Table 3. Wisconsin Breast Cancer Dataset Attribute

S.No	Attribute	Domain
1	Sample Code Number	Id number
2	Clump Thickness	1 – 10
3	Uniformity of Cell Size	1 – 10
4	Uniformity of Cell Shape	1 – 10
5	Marginal Adhesion	1 – 10
6	Single Epithelial Cell Size	1 – 10
7	Bare Nuclei	1 – 10
8	Bland Chromatin	1 – 10
9	Normal Nucleoli	1 – 10
10	Mitoses	1 – 10
11	Class	2(Benign) or 4(Malignant)

The dataset provided to you will be in the order of Table 3. The Class column has been set to 0 as this is what you will be calculating using your tree structure.

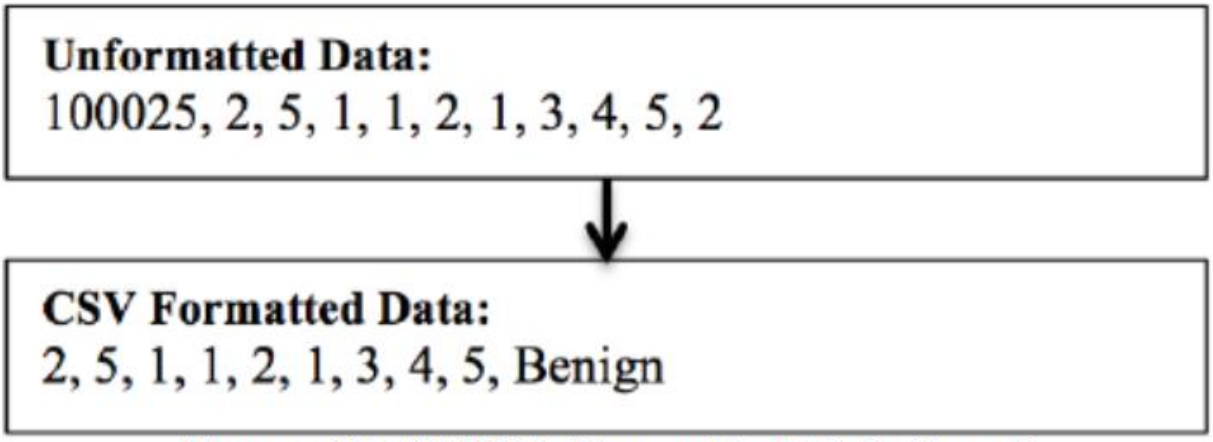

Figure 12. WEKA Formatted Data Input

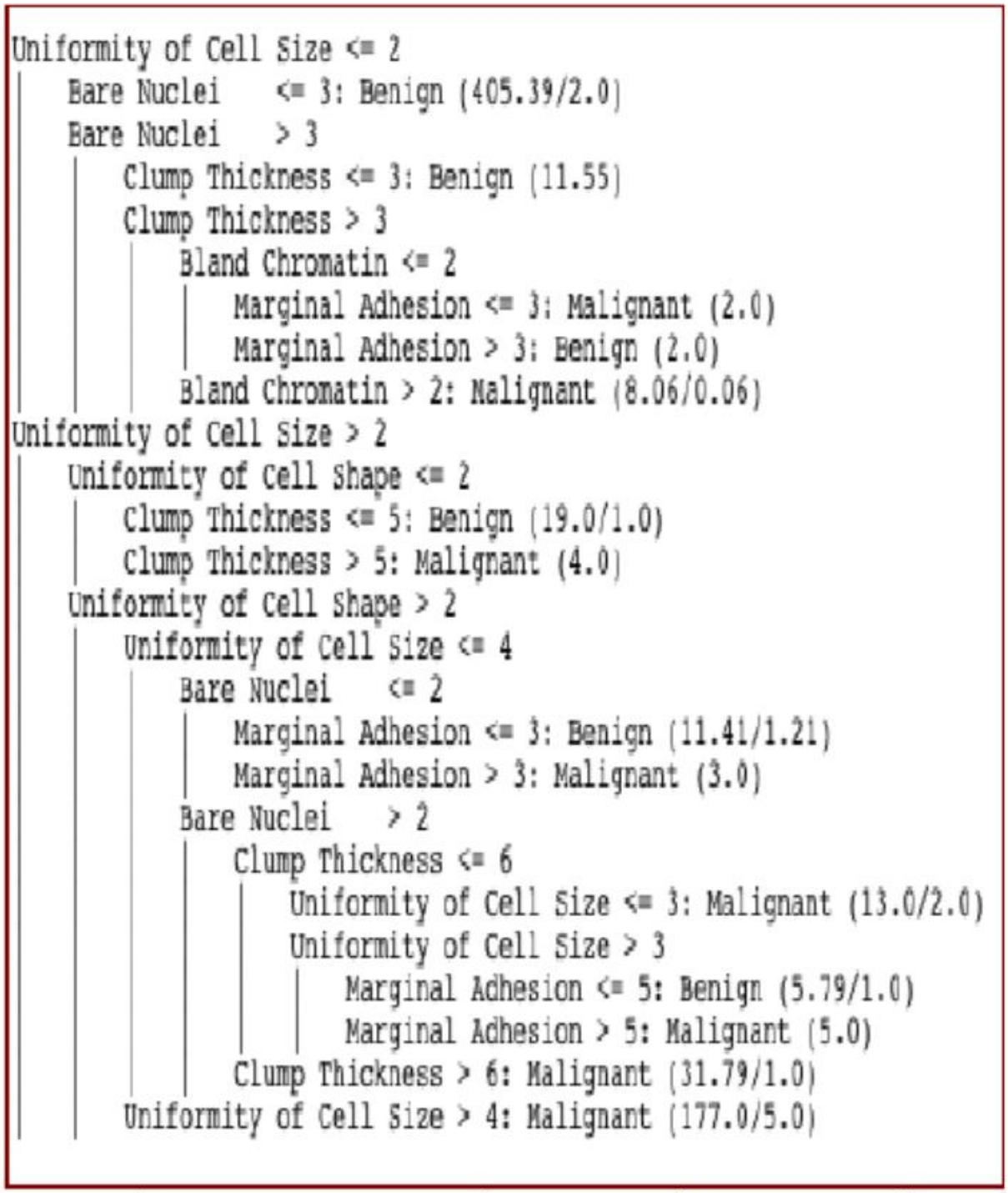

Figure 13. J48 Pruned Tree – Rules Generated

Reference:

To read the full research paper, go to:

https://www.researchgate.net/publication/272863357_Diagnosis_of_Breast_Cancer_using_Decision_Tree_Data_Mining_Technique

Submission:

The easiest way to submit is to provide your entire project in a zip file.

1. When creating a project ensure you choose 'empty project'
2. Make sure the target is set to x64
3. Code your project.
4. Make sure everything compiles and runs.
5. From the build menu select and execute "Clean Solution"
    a. This will ensure all compiled code is removed reducing the project size.
6. To export.
    a. Open the location of your solution file (.sln)
    b. Select all files and folders
    c. Right-click and send to compressed (zip) folder
    d. This will create a zip file in your folder.
    e. Rename the zip file to **LastNameFirstName_Project2.zip**
    f. Submit this zip file to the "Project 2 – Breast Cancer Diagnosis" drop box
7. To test your Project.
    a. Download your submission from FoL
    b. Extract the files to a new location on your pc
    c. Double click the .sln file to open
    d. Press F5 to run in Visual Studio

Submit your project on time!

Submissions must be made on time! Late projects will be subject to divisional policy on missed test and late projects.

10% per day for a maximum of 5 days after which the submission will receive a zero grade.

Submit your own work and *keep it to yourself*!

It is considered cheating to submit work done by another student or from another source. Helping another student cheat by sharing your work with them is also not tolerated. Students are encouraged to share ideas and to work together on practice exercises, but any code or documentation prepared for a project must be done by the individual student. Penalties for cheating or helping another student cheat may include being assigned zero on the project with even more severe penalties if you are caught cheating more than once. Just submit your own work and benefit from having made the effort on your own.

All work will be subject to “TurnItIn” scrutiny.

Grading Scheme

Marks Available	Any amount of plagiarism will be awarded a zero Non compiling code will receive a zero -50% Not using a Binary Decision Tree (Node Based Logic)	Marks Awarded
5	**Code is commented / Code is formatted**	
5	**Naming conventions / Following general coding guidelines**	
10	**Correctly diagnose the patient data by using the binary tree.**	
10	**Input from unformatted_data.csv file parsed correctly**	
10	**Output results.csv file with the diagnosis in the correct location of the file.** **Note: Output must match the format of the input file with the diagnosis (0) replaced with a (2 or 4). Do not overwrite the input file.**	
10	**Print a summary showing the following:** • Total Patients Processed • Total Benign • Total Malignant	
10	**Program runs, creates the appropriate files, outputs and displays to console the total number of patients, number of patients diagnosed with benign tumors and malignant tumors**	
60	**TOTAL MARKS**	