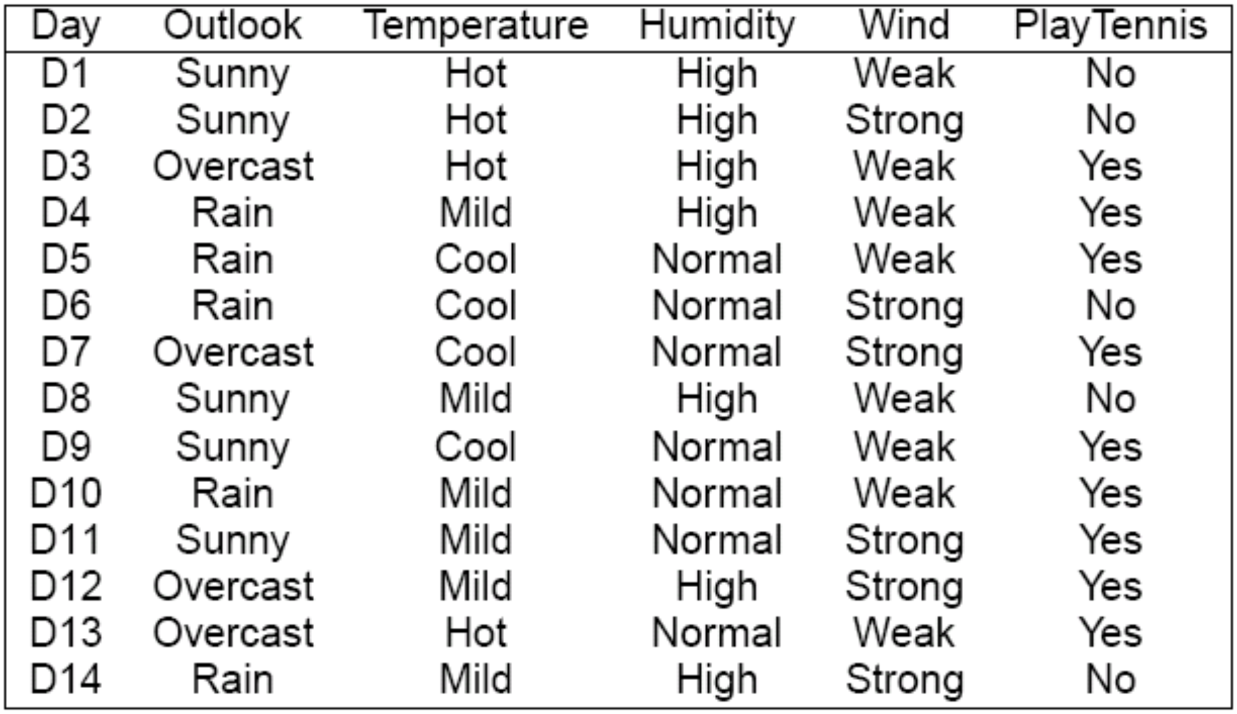
BME 6938 Multimodal Data Mining
Mini Project 2: Machine Learning

(Total: 50 points)

Due: March 12, 11:59 PM via Canvas

1. Classification: Playing Tennis (10 points)

Provide necessary steps in the problem solving (e.g. list the steps and intermediate results to calculate entropy, information gain, and GainRatio, prior probabilities). Draw the final decision trees (e.g. use Powerpoint to draw and save as a picture or any other way you prefer)



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

a) Build a decision tree using ID3 algorithm (Information Gain). (3 pts)

b) Build a decision tree using CART algorithm (Gini index). (3 pts)

c) Make predictions of (D15: Rain, Mild, Normal, Strong) using both trees. (2 pts)

d) Use Naïve Bayes classifier to predict the result in (c). (2 pts)

2. Python Programming: Classifying Parkinson's Disease (40 points)

Use the accompanied dataset for this mini project. Read the dataset description below carefully to make sure that you understand the dataset features and values.

> ***Dataset description:*** *This dataset belongs to 20 People with Parkinson (PWP) (6 female, 14 male) and 20 healthy individuals (10 female, 10 male). From all subjects, multiple types of sound recordings (26 voice samples including sustained vowels, numbers, words and short sentences) are taken. A group of 26 linear and time or frequency-based features are extracted from each voice sample. UPDRS ((Unified Parkinson's Disease Rating Scale) score of each patient which is determined by expert physician is also available in this dataset. Therefore, this dataset can also be used for regression.*
>
> *The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording; the subject ID is identified in the first column.*

a) Complete the template MP2.ipynb on Canvas. Follow the instructions to fill in all TODO fields. (30 points)

b) Bonus (Optional): Perform regression using UPDRS as the label. You can use support vector regression or XGBoost Regression. (bonus: 5 points)

c) Write a Mini-Project report to describe and record the optimal parameters and performance of each machine learning algorithms. Show and compare the learning and cross-validation curves for each algorithm. How would using additional training data affect this optimum? Are these models overfitting, underfitting, or balanced? (10 points)

Submission:

1. Submit the solution to Question 1 and Report to Question 2 in one PDF file via Canvas.

2. Upload your completed MP2.ipynb to your Github (make it public) and paste the link to Canvas submission.

- No late homework will be accepted (you will not be able to submit the repository link to Canvas after the due date). Do not change the repository after the due date; your homework will not be graded.