SIT 720 - Machine Learning

Lecturer: Chandan Karmakar | karmakar@deakin.edu.au

School of Information Technology,
Deakin University, VIC 3125, Australia.

Assessment Task 5 (35 marks)

In this assignment, you will use a lot of concepts learnt in this course to come up with a good solution for a given chronic kidney disease prediction problem.

Submission Instruction

1. Student should insert Python code or text responses into the cell followed by the question.
2. For answers regarding discussion or explanation, **maximum ten sentences are suggested**.
3. Rename this notebook file appending your student ID. For example, for student ID 1234, the submitted file name should be A5_1234.ipynb.
4. Insert your student ID and name in the following cell.

```
# Student ID:

# Student name:
```

The dataset

Dataset file name: chronic_kidney_disease.csv

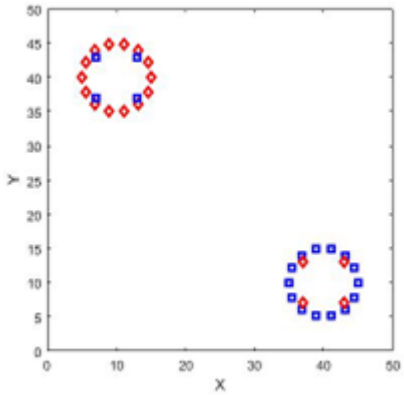
Attribute Information:

There are 24 features + class = 25 attributes

1. Age(numerical): age in years
2. Blood Pressure(numerical): bp in mm/Hg
3. Specific Gravity(nominal): sg - (1.005,1.010,1.015,1.020,1.025)
4. Albumin(nominal): al - (0,1,2,3,4,5)
5. Sugar(nominal): su - (0,1,2,3,4,5)
6. Red Blood Cells(nominal): rbc - (normal,abnormal)
7. Pus Cell (nominal): pc - (normal,abnormal)
8. Pus Cell clumps(nominal): pcc - (present,notpresent)
9. Bacteria(nominal): ba - (present,notpresent)
10. Blood Glucose Random(numerical): bgr in mgs/dl

11. Blood Urea(numerical): bu in mgs/dl
12. Serum Creatinine(numerical): sc in mgs/dl
13. Sodium(numerical): sod in mEq/L
14. Potassium(numerical): pot in mEq/L
15. Hemoglobin(numerical): hemo in gms
16. Packed Cell Volume(numerical)
17. White Blood Cell Count(numerical): wc in cells/cumm
18. Red Blood Cell Count(numerical): rc in millions/cmm
19. Hypertension(nominal): htn - (yes,no)
20. Diabetes Mellitus(nominal): dm - (yes,no)
21. Coronary Artery Disease(nominal): cad - (yes,no)
22. Appetite(nominal): appet - (good,poor)
23. Pedal Edema(nominal): pe - (yes,no)
24. Anemia(nominal): ane - (yes,no)
25. Class (nominal): class - (ckd, notckd)

Part 1: Short questions: **(6 marks)**



1. For the above figure, what value of k in KNN method will give the best accuracy for leave-one-out cross-validation. Report accuracy and k value. **(3 marks)**

```
# CODE and/or COMMENT
```

2. In classification, overfitting and underfitting is a big problem. Does it happen in Random Forest or not? Why? **(3 marks)**

```
# CODE and/or COMMENT
```

Part 2: **(24 marks = 4 methods x 6)**

Using the following four supervised machine learning methods, answer questions(A-D).

1. Support vector machine
2. K-Nearest Neighbour
3. Decision tree, and
4. Random forest

A. Build optimised classification model to predict the chronic kidney disease from the dataset. **(1 marks)**

B. For each optimised model, answer the followings - **(3 marks)**

- which hyperparameters were optimised? [Hint: For SVM, kernel can be considered as one of the hyperparameters.]
- what set or range of values were used for each hyperparameter?
- which metric was used to measure the performance?
- justify your design decisions.

C. Plot the prediction performance against hyperparameter values to visualise the optimisation process and mark the optimal value. **(1 marks)**

D. Evaluate the model (obtained from A) performance on the test set. Report the confusion matrix, F1-score and accuracy. **(1 marks)**

```
# CODE and/or COMMENT
```

Part 3: Discussion **(5 marks)**

Based on the results obtained in Part-2, which classification method showed the best performance and why? Do you have any suggestions to further improve the model performances? **(5 marks)**

```
# CODE and/or COMMENT
```