Work Integrated Learning Programmes Division
M.Tech (Data Science and Engineering)
Machine Learning
DSECLZ G565
Second Semester, 2021 -22

Assignment 2 – PS 12

Kidney Disease Prediction- [Weightage 10%]

Instructions for Assignment Evaluation

1. Please follow the naming convention as <Group no>_<Dataset name>.ipynb.

   Eg – for group 1 with a weather dataset your notebooks should be named as - Group1_WeatherDataset.ipynb.

2. Inside each jupyter notebook, you are required to mention your name, Group details and the Assignment dataset you will be working on.
3. Organize your code in separate sections for each task. Add comments to make the code readable.
4. Deep Learning Models are strictly not allowed. You are encouraged to learn classical Machine learning techniques and experience their behavior.
5. Notebooks without output shall not be considered for evaluation.
6. Prepare a jupyter notebook (recommended - Google Colab) to build, train and evaluate a Machine Learning model on the given dataset. Please read the instructions carefully.
7. Each group consists of up to 3 members. All members of the group will work on the same problem statement.
8. Each group should upload in CANVAS in respective locations under ASSIGNMENT Tab. Assignment submitted via means other than through CANVAS will not be graded.

9. Only two files should be uploaded in canvas without zipping them. One is ipynb file and other one html output of the ipynb file. No other files should be uploaded.

Problem Statement

The data was taken over a 2-month period in India with 25 features (eg, red blood cell count, white blood cell count, etc). The target is the 'classification', which is either 'ckd' or 'notckd' - ckd=chronic kidney disease. There are 400 rows. Early diagnosis can lead to lifestyle changes and more effective treatment, making predictive models for chronic disease risk important tools for public and public health officials.

Dataset: Chronic Kidney Disease dataset

https://drive.google.com/file/d/1NykVFA1f5oGXZ5JlrBGXPBJrnfRncREh/view?usp=sharing

1. **Import Libraries/Dataset**

   1. Download the dataset
   2. Import the required libraries

2. **Data Visualization and Exploration [1.5 M]**

   1. Print 2 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
   2. Comment on class imbalance with appropriate visualization method.
   3. Provide appropriate data visualizations to get an insight about the dataset.
   4. Do the correlational analysis on the dataset. Provide a visualization for the same. Will this correlational analysis have effect on feature selection that you will perform in the next step? Justify your answer. **Answer without justification will not be awarded marks**

3. **Data Pre-processing and cleaning [1.5 M]**

   1. Do the appropriate pre-processing of the data like identifying NULL or Missing Values if any, handling of outliers if present in the dataset, skewed data etc. Mention the pre-processing steps performed in the markdown cell with justification.

   2. Apply appropriate feature engineering techniques for them. Apply the feature transformation techniques like Standardization, Normalization, etc. You are free to apply the appropriate transformations depending upon the structure and the complexity of your dataset. Provide proper justification. **Techniques used without justification will not be awarded marks**

4. Model Building [4M]

1. Split the dataset into training and test sets. Justify your choice of split. You may experiment with different split to get the final split. **Answer without justification will not be awarded marks. [1M]**
2. Build Model Development using Decision Tree and KNN. Identify the best parameter and justify your answer. **[3 M]**

5.   Performance Evaluation [3 M]

1. Do the prediction for the test data and display the results for the inference. Calculate all the evaluation metrics and choose best for your model. Justify your answer. **Answer without justification will not be awarded marks. [2M]**
2. Compare the model in tabulated form. **[1M]** Justify your comment. **Answer without justification will not be awarded marks.**