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

Assignment 1 – PS11 - [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.

Problem Statement

Part A [5M]

Dataset: Diabetes dataset to predict the risk of a person given the medical parameters

https://drive.google.com/file/d/16oFnnycwRG6PUBWWpssUb7x04v73hvZp/view?usp=sharing

1. Import Libraries/Dataset

1. Download the dataset
2. Import the required libraries

2. Data Visualization and Exploration [1M]

1. Print 2 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
2. Print the description and Basic statistical details.
3. Print each class label count (Activity) and create a pie chart for each class (% of data distribution). Write your observation on data balancing.
4. Plot Activities by Subject/Participants and Provide appropriate comments on visualized data.
5. Try exploring the data and see what insights can be drawn from the dataset.

3. Data Pre-processing and cleaning [2M]

1. Do the appropriate preprocessing steps
    1. Identify NULL or Missing Values based on column. Apply appropriate feature engineering techniques for them.
2. Use MinMax normalization for feature transformation.
3. Do the correlational analysis on the dataset. Provide a visualization for the same.

Part B

1. Model Building [5M]

1. Perform Model Development using Naïve Bayes with appropriate hyper parameters.

2. Train the model and print the training accuracy, Recall, F1 Score for case 1, case 2 separately.
3. Deep Learning Models are strictly not allowed.

2. Performance Evaluation [2M]

1. Do the prediction for the test data and display the results for the inference.
2. Print test Accuracy, Recall, F1 Score for case 1 and case 2 separately.
3. Print the confusion matrix for all cases. Provide insights on the most suitable matrix in this case.
4. Compare the accuracy of train data with test data. Provide appropriate analysis for the same for all cases.
5. Write your observation for result of each question and justify your answer.