BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Work Integrated Learning Programmes Division

Machine Learning
DSECCZ G565
Second Semester, 2021 -22

Assignment 1 – PS2

Mushroom Classification- [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

Mushroom Classification

Dataset: Mushroom Classification Dataset

https://drive.google.com/file/d/1Urc0bLTFCV65FPHK-MGw9rXoIJhwreiH/view?usp=sharing

Import Libraries/Dataset

1. Download the dataset
2. Import the required libraries

Data Visualization and Exploration [2 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 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 an effect on feature selection that you will perform in the next step? Justify your answer. **Answers without justification will not be awarded marks.**
5. Any other visualisation specific to the problem statement.

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

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. Explore a few latest data balancing tasks and its effect on model evaluation parameters.

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.** Explore a few techniques for identifying feature importance for your feature engineering task.

3. Model Building [4M]

1. Split the dataset into training and test sets. **Answers without justification will not be awarded marks. [0.5M]**

   Case 1 : Train = 80 % Test = 20% [x_train1,y_train1] = 80% ; [x_test1,y_test1] = 20% ;
   Case 2 : Train = 10 % Test = 90% [x_train2,y_train2] = 10% ; [x_test2,y_test2] = 90%

2. Explore k-fold cross validation. **[0.5M]**

3. Build Model/s using 1) Logistic Regression 2) MLE **[2M]**

4. Explore the need of regularization and incorporate few relevant techniques for the problem statement. **[0.5M]**

5. Compare models with and without regularization in a tabular format and justify the findings. **[0.5M]**

4. Performance Evaluation [2 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. **Answers without justification will not be awarded marks. [1M]**
2. Comment on underfitting/overfitting/just right model. Justify your comment. **Answers without justification will not be awarded marks. [1M]**