DMG Assignment 3

Kaggle competition link:
https://www.kaggle.com/t/45933f6432c04e13af5448727daabd23

Dataset: (uploaded on Kaggle)

Dataset size -(11000, 11)
Target variable - 'Label'

Aim: Build a classification model with Association Pattern mining (Rules) only.
Evaluation Metric: Macro F1
Deadline: 23 November 2020

Instructions:

1. Use only the username you provided for submission on Kaggle.
2. Mention all assumptions if any in the report.
3. Report plus code in .py or standard format should be submitted in the classroom in a zip folder with the name 'A3_RollNumber1_RollNumber2'.
4. You can use spmf (python wrapper) or weka, You are also free to build the rules from scratch.
5. Machine learning (Scikit learn and likewise libraries) and Transfer learning techniques are **NOT** allowed. If found in use, it will lead to disqualification of the team with 0 marks.
6. Include one runner function in code which takes test_X.csv as input and produces result.csv. All preprocessing to be done on data before applying the model should be present in the runner function.
7. No restrictions on which libraries to use except standard ML libraries, However, you can use sklearn for train-test split or K-Fold validation (If required).
8. Some students will be randomly picked for a demo of assignment 3. So write the code on your own, make sure you don't cheat. If you can't answer the questions during your demo, 50% of your marks will be deducted.
9. A single team member will submit on the google classroom and will mention the contributions of each member in the report.

Kaggle Instructions:

1. Make a team of two on Kaggle. The team name should be the roll numbers of both members: RollNo1_RollNo2
2. In case of doubts, comment on the classroom and not on the Kaggle discussion forum.
3. The maximum daily submission limit is 10.
4. Do not share the competition link.

The following should be included in the Report:

1. Explain your methodology: approach and reason clearly in the report.
2. Visualize skewness of data before and after preprocessing (if done any).
3. Add all data analysis steps which you have performed on the dataset.
4. Compare your F1 score with the Random forest-based baseline given on the kaggle leaderboard.
5. Make a section “Learning”, which describes your learning in doing this assignment.

Evaluation:

- 50% marks for a report containing algorithm and classification Score (F1) using the association approach, Also compare your score (F1) with the random forest baseline on kaggle.
- 50% marks as per the final leaderboard rankings.

Ranking Bucket	**Points**
Top 10 %	50
Next 10 %	45
Next 10%	40
Next 20 %	35
Next 30%	30
Next 20% (At least a valid Kaggle submission)	20