ML Project Weekly Report

Coure Name: Machine Learning (CSE 523)
Week Number: 6 (10rd March - 16th March 2024)
Group Name: White
Instructor's Name: Mehul S Raval
Project no.: 7

Project 7 | Athlete Statistics Visualization and Prediction

Introduction

- In continuation to the methods employed in the data pre-processing last week, we aim to further bring a balance in dataset between the minority and majority classes.

Weekly Activities

1. **Data Balance**
    - Conducted research to study about SMOTE
    - Trials of Python implementation conducted

2. **Dashboard Study**
    - Studied the Dashboard that represents the athlete data through visualization from research paper provided
    - Discussed Dashboard layout

Challenges Faced

- SMOTE uses the concept of oversampling, hence it brings forward new instances of the minority class. This introduces noise since the probability of class overlap increases.

Learnings

- With guidance from the paper, we learnt that the application of ENN (Edited Nearest Neighbor) technique on the oversampled dataset can try to fix the aforementioned problem.
- It essentially implements the removal of those majority class instances whose neighbors belong to a different class completely. The identification of these neighbors is performed by KNN (K-Nearest Neighbor) algorithm.

Conclusion

- We have tried the optimal methods for preprocessing under the guidance of the research papers. As a final touch, we shall modify some of the techniques employed with other similar approaches for preprocessing in order to make way for a clean useful dataset.