DATA MINING PROJECT PROPOSAL

GROUP 5

Group Members: Ashwin Dixit, Deyu Kong, Haruna Salim

1. What problem did you select and why did you select it?

   The soccer industry is the largest sports industry in the world. For developing countries in europe, most decisions on soccer matches and team management are made based on sentiments rather than actual data. This model will help teams predict matches based on their team selection and also help with some basic team-player analysis.

2. What database/dataset will you use? Does it need to be cleaned?

   The dataset is a sqlite file obtained from kaggle at European Soccer Database . It contains matches, players and team information. The database has some missing values but they are very minimal. Therefore it will be cleaned in order to produce a more accurate result.
   Dataset : https://www.kaggle.com/hugomathien/soccer

3. What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?

   - Multiclass classification for game outcome - SVM

4. What packages will you use to implement the network? Why?
   - Sqlite3
   - Numpy
   - Panda
   - Matplotlib
   - Scipy

- PyQty
- SKlearn

5. What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?

   The reference materials will be from Kaggle and other scholarly journals,

   - https://www.kaggle.com/hugomathien/soccer/code
   - https://www.kaggle.com/yoyocm/how-predict-the-outcome-of-48-matches
   - https://link.springer.com/article/10.1007/s10994-018-5726-0
   - https://journals.sagepub.com/doi/full/10.1177/1471082X18810971

6. How will you judge the performance of your results? What metrics will you use? • Provide a rough schedule for completing the project.

   Performance & Metrics:

   - The percentage of successful predictions

   SCHEDULE:

   1. Week 1 : Data Preprocessing.
   2. Week 2 & 3: Building the GUI & Implementation of Mining Algorithms.
   3. Week 4: Model parameter tuning and prepare for the presentation.