|  |  |
| --- | --- |
| **Practicum Case** |  |
| COMP6140 | COMP6140001 | COMP6140049  Data Mining |
| **Computer Science** | **O191-COMP6140-NP01-06** |
| ***Valid on*** *Even Semester Year 2019/2020* | **Revision 00** |

**Learning Outcome**

* LO1 – explain concept of data and data preprocessing
* LO 2 – apply various data mining techniques

**Topic**

* Session 06 – Introduction to RapidMiner

## Sub Topics

* Data Preprocessing (Missing Values, Discretization, Normalization)
* Data Reduction (Sampling, PCA)
* Data Description

## Soal

*Case*

**BluejackFC**

**BluejackFC** is a game title of a multiplayer football game which is planned to be released in a year. To accurately illustrate the abilities of each football players in the world, they need to analyze the records of various physical tests and observations from these characters which will then be used to conclude the degree of **offensive**, **defensive**, and **goalkeeping** ability of the players. As a data scientist, you are asked to analyze the given data from the document named **FullData.csv** and analyze the potentials of the soon-to-be released characters.

To achieve the goal, you are asked to perform various processes as follows:

1. **Retrieve** the **data** from **FullData.csv**.
2. **Remove** player data with **empty** **National\_Position** information.



***Figure 1. Data with Empty National Position***

1. **Parse** the attribute **Club\_Joining** as **date**.
2. **Remove** data with **Club\_Joining** attribute **older than** **31** **December 2010**.
3. **Discretize** the attribute **Rating** into ranges of values by **binning**. Divide the data into **5** **bins**.
4. **Set the role** of the attribute **Rating** to **Label**.
5. **Sample** the data with **Stratified Sampling** with sample **ratio 0.4**, using **local random seed 1993**.
6. **Select these attributes** to use in the next processes:
   * Agression
   * Attacking\_Position
   * Ball\_Control
   * Composure
   * Dribbling
   * Finishing
   * GK\_Diving
   * GK\_Handling
   * GK\_Kicking
   * GK\_Position
   * GK\_Reflexes
   * Interceptions
   * Name
   * Rating
   * Reactions
   * Shot\_Power
   * Sliding\_Tackle
   * Standing\_Tackle
   * Vision
7. **Normalize all attributes** except **Name** and **Rating**.
8. **Set** the **role** of the attribute **Name** to **identifier**.
9. **Multiply** the data into **three** so that it can be used in **multiple processes**.
10. For the **first** **data** (**Offensive Ability Analysis**), perform the following processes.
    * 1. **Select Attributes** to use in the next processes.
         + Aggresion
         + Attacking\_Position
         + Finishing
         + Name
         + Rating
         + Shot\_Power
         + Sliding\_Tackle
         + Standing\_Tackle
      2. Perform **Principal Component Analysis** (**PCA**) to the data while **reducing its dimensionality** to **1**.
      3. **Rename** the **PCA** result **column** **name** to **Offensive Component**.
11. For the **second data** (**Defensive Ability Analysis**), perform the following processes.
    * 1. **Select Attributes** to use in the next processes.
         + Ball\_Control
         + Composure
         + Dribbling
         + Interceptions
         + Name
         + Rating
         + Reactions
         + Vision
      2. Perform **Principal Component Analysis** (**PCA**) to the data while **reducing its dimensionality** to **1**.
      3. **Rename** the **PCA** result **column** **name** to **Defensive Component**.
12. For the **third data** (**Goalkeeping Ability Analysis**), perform the following processes.
    * 1. **Select Attributes** to use in the next processes.
         + GK\_Diving
         + GK\_Handling
         + GK\_Kicking
         + GK\_Positioning
         + GK\_Reflexes
         + Name
         + Rating
      2. Perform **Principal Component Analysis** (**PCA**) to the data while **reducing its dimensionality** to **1**.
      3. **Rename** the **PCA** result **column** **name** to **Goalkeeping Component**.
13. **Join** the results of these **three processed data** into one and **show the result**.



***Figure 2. Result***