10351 ניתוח נתוני עתק

תרגיל #2

The purpose of this Lab is to practice SPARK Feature Manipulations

Problem definition:

1. Read data from the file “dataFIFA.scv”.

2. The purpose of this LAB is to find players that belong to the group of players with top 10 salaries. Decide about the type of feature defined in each column (categorical, numerical, binary). Choose 5 more promising feature for clustering to define if 10 players with biggest salary belong to the same cluster

Column name	Feature type	Way to manage a feature	Example before and after the feature transformation

3. Use Spark ML library to prepare the feature to find clusters.

4. Discover the number of clusters with the largest Silhouette. For this number of clusters create a result table (the number of rows as an amount of the clusters):

Clusters ID	The number of players from the 10 players with biggest salary

5. Use PCA to reduce the dimensions of feature vector. Is it more helpful for clustering? Fill the following table for applying PCA with 5, 4, and 3 features.

Clusters ID	The number of players from the 10 players with biggest salary

6. Perform task 5 with reduction to two features only. Draw 2D plot of clusters and top 10 players with biggest salary.

Requirements:

Fill all tables, supply the Jupiter Notebook as well.

Grading Policy:

- **10 points** for code quality:
  a. The code must be divided into small functions
  b. Use meaningful names for variables, functions, files, constants, etc.
  c. Place enough comments to understand the code
  d. No unused lines of code. Don't repeat the code – use functions!
  e. Write README.TXT file if special instructions are needed to run the solution. The file must be in the root folder of the solution.

- **90 points (**in total**)** – for proper implementation of the requirements.
  a. **45 points** for feature engineering
  b. **45 points** for clustering

Important:

- The Solution of this Homework must be suppled as Jupiter Notebook and tested on VLAB.
- The Homework must be uploaded to the Moodle and delivered in time. No delay will be accepted.

- The homework may be performed in pairs. Only one member of pair submits the solution through the Moodle. The whole project must be zipped and named as

111111111_222222222.zip

Where **111111111** is ID of the one student and **222222222** is ID of another student

בהצלחה!