Data Mining Project: *Tennis Matches*

A **project** consists in data analysis based on the use of data mining tools. The project has to be performed by a team of 2/3 students. It has to be performed by using Python. The guidelines require to address specific tasks and results must be reported in a unique paper. The total length of this paper must be **max 20 pages** of text including figures. The students must deliver both: paper and well commented Python notebooks.

DATASET DESCRIPTION

The dataset to be analyzed is composed of information about some tennis matches. Data are organized in .csv file:

- tennis_matches.csv where each row represents a match
- male_players.csv
- female_players.csv

The dataset contains different features, with the following semantic meaning:

- **Tourney_id**: a unique identifier for each tournament, such as 2020-888. The exact formats are borrowed from several different sources, so while the first four characters are always the year, the rest of the ID doesn't follow a predictable structure.
- **Tourney_name**: the name of the tourney
- **Surface**: kind of surface for the match
- **Draw_size**: number of players in the draw, often rounded up to the nearest power of 2. (For instance, a tournament with 28 players may be shown as 32.)
- **Tourney_level**: they are split for men and women.
    - **For men**: 'G' = Grand Slams, 'M' = Masters 1000s, 'A' = other tour-level events, 'C' = Challengers, 'S' = Satellites/ITFs, 'F' = Tour finals and other season-ending events, and 'D' = Davis Cup. F
    - **For women**, there are several additional tourney_level codes, including 'P' = Premier, 'PM' = Premier Mandatory, and 'I' = International. The various levels of ITFs are given by the prize money (in thousands), such as '15' = ITF $15,000. Other codes, such as 'T1' for Tier I (and so on) are used for older WTA tournament designations. 'D' is used for the Federation/Fed/Billie Jean King Cup, and also for the Wightman Cup and Bonne Bell Cup.
    - There is also some competition which can be for both men and women: 'E' = exhibition (events not sanctioned by the tour, though the definitions can be ambiguous), 'J' = juniors, and 'T' = team tennis, which does yet appear anywhere in the dataset but will at some point.
- **Tourney_date**: eight digits, YYYYMMDD, usually the Monday of the tournament week.
- **Match_num**: a match-specific identifier. Often starting from 1, sometimes counting down from 300, and sometimes arbitrary.
- **Winner_id**: the player_id used in this repo for the winner of the match.

- **Winner_entry**: 'WC' = wild card, 'Q' = qualifier, 'LL' = lucky loser, 'PR' = protected ranking, 'ITF' = ITF entry, and there are a few others that are occasionally used.
- **Winner_hand**:R = right, L = left, U = unknown. For ambidextrous players, this is their serving hand.
- **Winner_ht**: height in centimetres, where available
- **Winner_ioc**: three-character country code
- **Winner_age**: the age of the player, in years, depending on the date of the tournament
- **Best_of**: '3' or '5', indicating the number of set for this match
- **Minutes**: match length, where available
- **W_ace**: winner's number of aces
- **W_df**: winner's number of doubles faults
- **W_svpt**: winner's number of serve points
- **W_1stIn**: winner's number of first serves made
- **W_1stWon**: winner's number of first-serve points won
- **W_2stwon**: winner's number of second-serve points won
- **W_SvGms**: winner's number of serve games
- **W_bdSaved**: winner's number of breakpoints saved
- **W_bdFaced**: winner's number of breakpoints faced
- **Winner_rank**: winner's ATP or WTA rank, as of the tourney_date, or the most recent ranking date before the tourney_date
- **Winner_rank_points**: number of ranking points, where available.

We did not report the meaning of losers attribute because they are the same as the winners, but the feature names start with 'loser'.

Task 1 Data Understanding and Preparation (30 points):

Task 1.1: Data Understanding: Explore the dataset with the analytical tools studied and write a concise "data understanding" report assessing data quality, the distribution of the variables and the pairwise correlations.

Task 1.2: Data Preparation: Improve the quality of your data and prepare it by extracting *new features* interesting for describing the player profile and his behavior derivable from matches. These indicators have to be extracted for each player. Examples of Indicators to be computed are:

- how many times did the player win during a given period
- how many matches the player played in a given period
- a ratio between the previous indicators
- percentage of aces related to the number of first serves made
- number of breakpoints numbers w.r.t. all games
- ….

Note that these examples are not mandatory. You can derive indicators that you prefer and that you consider interesting for describing the players.

It is MANDATORY that each team defines **indicators** and their description and when it is necessary also their mathematical formulation. The profile will be useful for the clustering analysis (i.e., the second project's task).
Once the set of indicators is computed, the team has to explore the new features for a statistical analysis (distributions, outliers, visualizations, correlations).

Subtasks of DU

- Data semantics for each feature that is not described above and the new one defined by the team
- Distribution of the variables and statistics
- Assessing data quality (missing values, outliers, duplicated records, errors)
- Variables transformations
- Pairwise correlations and eventual elimination of redundant variables

Task 2: Clustering analysis (30 POINTS - 32 with optional subtask)

Based on the player's profiles explore the dataset using various clustering techniques. Carefully describe your decisions for each algorithm and which are the advantages provided by the different approaches.

Subtasks

- Clustering Analysis by K-means:
    1. Identification of the best value of k
    2. Characterization of the obtained clusters by using both analysis of the k centroids and comparison of the distribution of variables within the clusters and that in the whole dataset
    3. Evaluation of the clustering results
- Analysis by density-based clustering**:**
    1. Study of the clustering parameters
    2. Characterization and interpretation of the obtained clusters
- Analysis by hierarchical clustering
    1. Compare different clustering results got by using different merging strategies
    2. Show and discuss different dendrograms using the different merging strategies
- Final evaluation of the best clustering approach and comparison of the clustering obtained
- **Optional (2 points):** Explore the opportunity to use alternative clustering techniques in the library: https://github.com/annoviko/pyclustering/

Delivery of the first draft of the report with Task 1.1, Task 1.2 and Task 2: 5 November

Task 3: Predictive Analysis (30 POINTS)

Consider the problem of predicting for each player a label that defines if s(he) is a *high ranked player* or a *low ranked player* (binary task) by exploiting the feature related to the rank of the players.

The student need to:

1. Define a player profile that enables the above player classification. For this task, you can exploit the profile created for the clustering task, by adding or removing features, depending on the results previously obtained.
2. Compute the label for any customer. The extraction of the label can take advantage of several features related to the rank, such as loser_rank, winner_rank, loser_rank_points, winner_rank_points, etc. An example of simple label can be derived by:
    - computing the average rank per player by exploiting loser_rank, winner_rank
    - select a threshold for discretizing in two categorical labels the class.

    Note that you can define in different ways the labels.
3. Perform the predictive analysis comparing the performance of different models, discussing the results and discussing the possible preprocessing that you applied to the data for managing possible problems identified that can make the prediction hard. Note that the evaluation should be performed on both training and test set.

Note: The final report delivered within the end of December can also improve the already delivered tasks.