**E Health Methods and Application: Project Report Part 1 GROUP 04**

**DATABASE**

STEP 1: Download Google-Playstore.csv dataset from Kaggle

We decided to work and apply the filters elaborated in the next phase on the csv format of the database in order to comfortably take advantage of the features offered by the Pandas library

STEP 2: Selection of app classified as “games”

Using the google-play-scraper we have implemented a filter that filters with the [genreId] all those applications falling into the game category, eliminating the others. Furthermore, since the google-playstore.csv was not updated, we have identified and eliminated those few applications that were present in the google-play-scraper but not in the google-playstore.csv.

STEP 3: Filtering apps by relevant features

We have implemented a function that filters games by certain categories ("educational", "educational", "family", "learning", "4-year-olds" and "4-year-olds"), and minimum requirements on rating (4.0 or above), rating count (>100000) and, in order to exclude those apps which are not translated in English, we provide a function capable of determine the set of characters that compose the app name and decide if it is an app written in English or not.

We chose those parameters based on our assumptions of what a “real” serious game should be, and taking into account the feedback provided by the users of the app too, as a reliable source of “good quality”.

STEP 4: Defining a first function to enrich database with description and reviews

We have implemented a function that is able to enrich the starting database in csv format with the description, reviews, learning category (which points to the learning field in which the app is specialized) and age range of each app through the google-play-scraper.

STEP 5: Using NLP approach to identify learning category and age range

We first defined eight learning categories ("science", "counting", "language", "creativity", "shape", "food", "music" and "sport") and for each of them we wrote a list of keywords associated according to natural language. Then we wrote a function that counts the maximum number of keywords found for each category for each app. In this way a category is assigned to the application, taking into account the number of keywords associated with it. With the same approach we have implemented the function that identifies the age range starting from four age ranges ("babies", "children", "adolescents" and "adults").

STEP 6: Create a new dataset to train the model we got

We chose 120 applications from google store randomly, including 40 serious games, 40 “fake” serious game (defined as those game that seem like “serious game”, but actually are not = misleading serious game), 40 normal and random apps.

This dataset (entirely built by human) has been compared to the one found by the algorithm in order to measure three parameters:

* *Accuracy*: the ability to recognize both the serious games and the misleading ones;
* *Sensitivity*: the ability to recognize the serious games;
* *Specificity*: the ability to recognize the misleading serious games.

STEP 7: Create the final database