**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 (25% above the medium) rating count (>1000) and the minimum number of installations (>1000000).

But why this numbers and how we can optimize these requirements?

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

We have implemented a function that could enrich the starting database in csv format with the description and reviews 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. We then 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, includes 30 serious games, 30 educational

apps, 30 other games and 30 normal apps. Test to see if our model works properly can find out all the serious games.

STEP 7: Create the final database