**Project: MTBI – Myers Briggs Personality Type Dataset**

Website: <https://www.kaggle.com/datasnaek/mbti-type>

**Introduction**

This project is based on a dataset from Kaggle.com® [1] related to information of sets of posts that individuals had posted, and for each one of them have a label, that is a combination of the following personality types based on the Myers Briggs Personality Type research[2],

* Introversion (I) – Extroversion (E)
* Intuition (N) – Sensing (S)
* Thinking (T) – Feeling (F)
* Judging (J) – Perceiving (P)

From these labels each type is constructed by combining in sets of four letter in parenthesis the kind of personality an individual has. For example, INTP or INTJ, in the case of INTJ, it’d be Introversion, Intuition, Thinking and Judging.

In this project the dataset had been used for identifying the individuals personality according to their posts writings. It is based on one of the most popular personality test in the world. It is used in businesses, online, for fun, for research and lots more[2]. In particular, the dataset from this project is based on the work done on cognitive functions by Carl Jung i.e. Jungian Typology. This was a model of 8 distinct functions, thought processes or ways of thinking that were suggested to be present in the mind [1].

**Dataset description**

This dataset contains over 8600 rows of data, on each row can be found every individual personality type corresponding to 4 letters as mentioned above, and a set of posts, around 50 (Each entry separated by "|||" (3 pipe characters).

The posts have a wide variety of words, websites, etc., that had been written by individuals that published their posts, that for everyone is around 50 posts, which is a lot. It is interesting that in each individual post their writings reflect in some way his or her personality. Whit this quantity of information it would be possible to make some research analyzing the writings in order to identify the kind of personality these people have.

**Exploratory Data Analysis**

The dataset have just two columns, one with the type of personality and the other with the set of posts an individual had posted. In the case of the type of personality, the distribution of the variable showed that it is not balanced at all. Then it had been chosen three of the values, INTP, INTJ and INFJ that were the most, kind of balanced.

Every post values have around 50 posts individuals have done, everyone separated by three pipes “|||”, in these case these characters had been deleted in order to have a set of ideas posted without differentiating among the posts.

With the help of these notebooks [3], [4] it had been identified some websites in. the posts that had been removed, some general characters, and of course the standard elimination of stop words and punctuation, after that, Lematization had been used. It had been used bag of words with word frequency count and also Tfidf transformation for the processing of the posts.

**Modeling and performance**

The technique that had been used was xgboost with the metric f1 score, it worked ok with 1000 max features in countvectorizer, the result was around 70%, this result goes down when there is an increment in max\_features. In the case of using Tf-idf, the results were around 68%. There is something to take into account, the quantity of words was very high, and the posts inside the posts, I think are, in general, from different matters, then, the variety of aspects considered in the posts are high; in this situation, is much more difficult to get higher results. The story would had been other if the individuals writing were related to some specific topic, maybe not, but I think it is something to tale into account.

**Bibliography**

[1] <https://www.kaggle.com/datasnaek/mbti-type>

[2] <https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/home.htm>

[3] <https://www.kaggle.com/lbronchal/what-s-the-personality-of-kaggle-users>

[4] <https://www.kaggle.com/depture/multiclass-and-multi-output-classification>