ITADATAhack 2023

Team Monty Hall

Gianluca Coletta, Simone Maiorani, Tommaso Martinelli, Marco Venturi

Road to the challenge

Università degli Studi di Perugia

IDEAL APPROACH

Split dataset in Train and Validation set

Grid Search for model selection

Prediction on the Test dataset

REAL APPROACH

BRUTE FORCE APPROACH FOR MODEL SELECTION

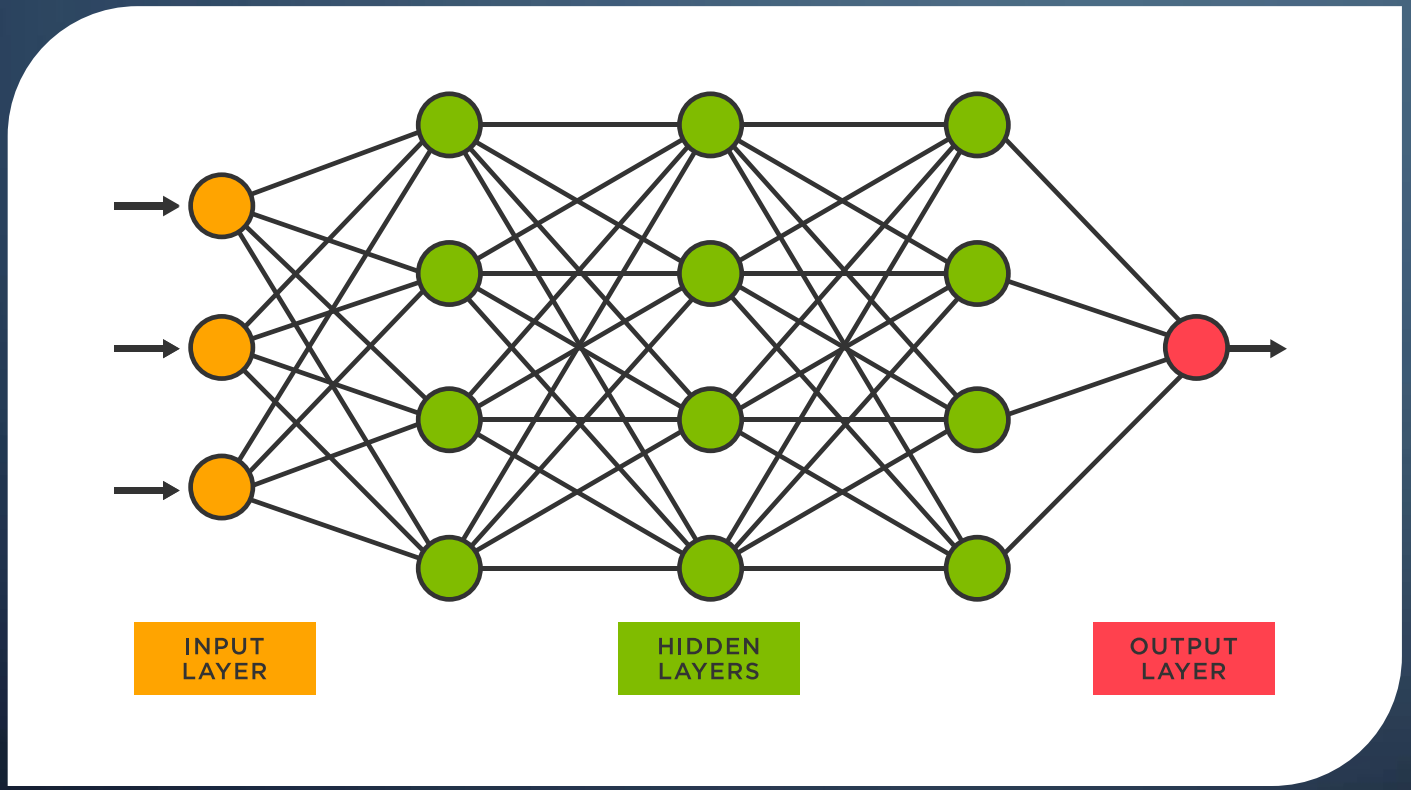
TRAINING ON THE TRAIN SET AND PREDICTION ON THE TEST SET

ANNOTATION OF RESULTS (DOES THE MODEL GENERALIZE WELL OR NOT?)

DAY 1

PREPROCESSING

- Removing numbers
- Removing punctuation
- Converting text to lower case
- Removing stopwords
- TF-IDF with N-grams



FINAL MODEL - MLP

- **Neural Network** from the scikit-learn library with 10 maximum iterations and two hidden layers of 100 and 50 neurons respectively.

DAY 2

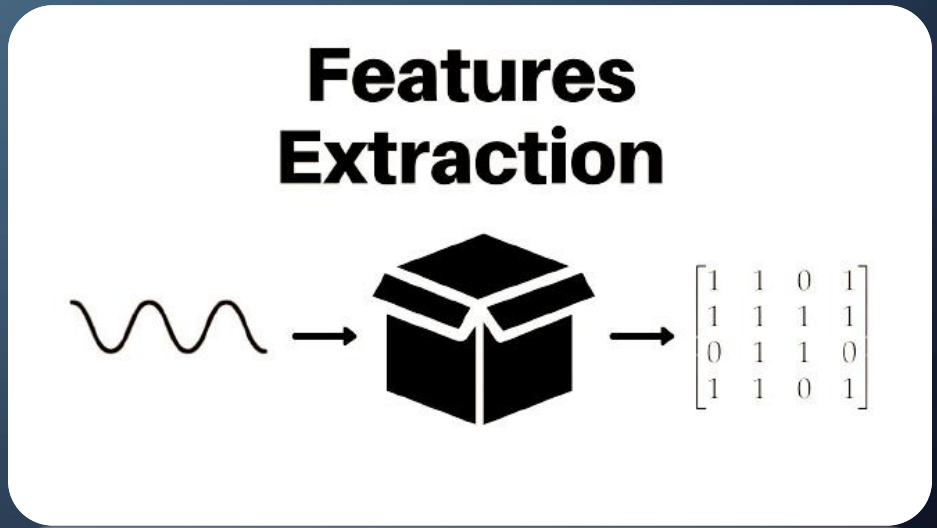
PREPROCESSING

Unlike day 1, we added:

- Removing stopwords with **punkt as tokenizer.**
- Reduction of the input text to the model by taking characters from 149 to 8000.
- TF-IDF (this time without N-grams).

FEATURE EXTRACTION

- On day two, citations were used as additional information.
- To overcome the problem of the necessary memory (about 110 GB are needed) to switch from the sparse matrix, generated by the TF-IDF, to the dense matrix of the Text feature, it was decided to use TF-IDF also to transform the citations.

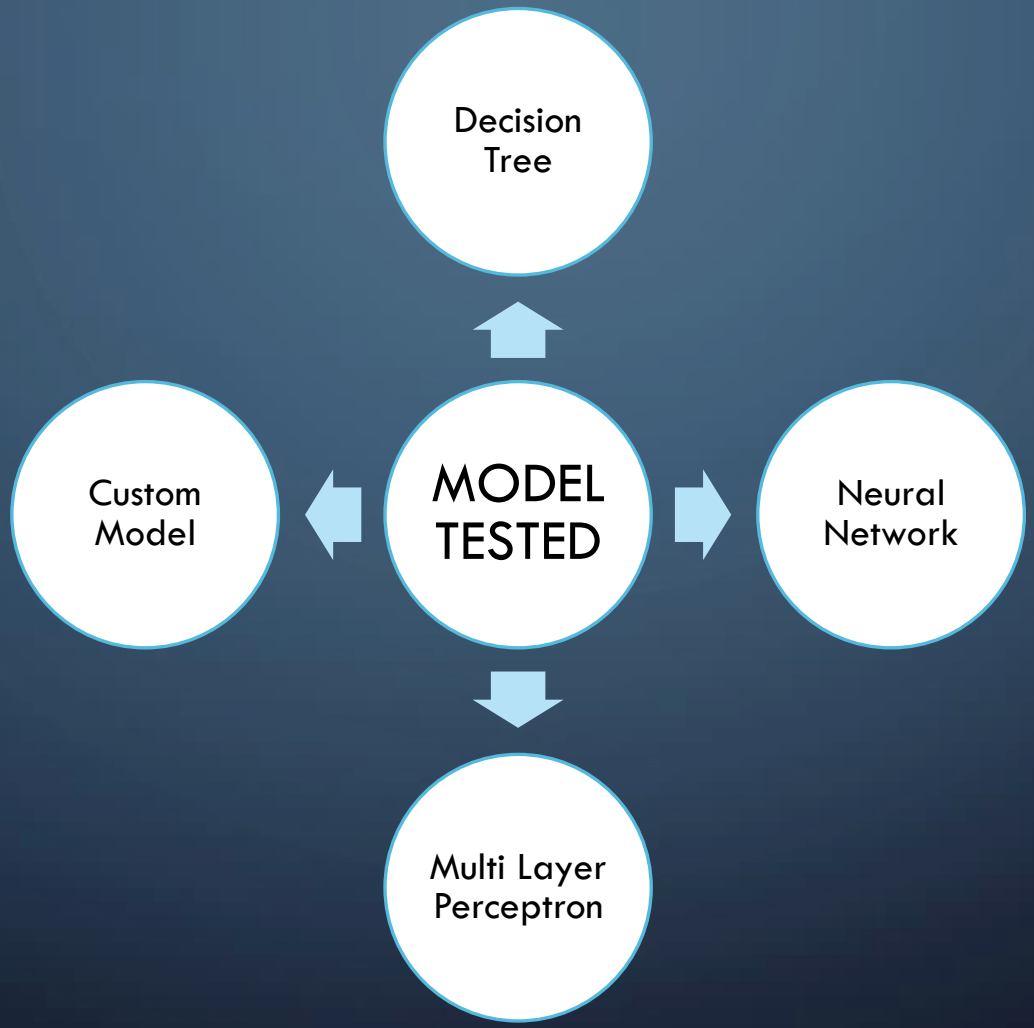


FINAL MODEL - MLP

Hidden Layers	Max Iterations	Activation Function	Optimization Algorithm	Learning Rate
(150, 150)	150	Identity	Adam	Adaptive

DAY 3

MODELS TESTED



FINAL MODEL - MLP

Multioutput Approach	Hidden Layers	Max Iterations	Activation Function	Optimization Algorithm	Learning Rate
MultiOutputClassifier	(150, 150)	150	Identity	Adam	Adaptive

THANKS FOR YOUR ATTENTION!