The computational study of the interactions between small molecules (ligands) and proteins (receptors) is very important, specially in Drug Discovery experiments. To study such interactions, many scoring functions were developed. In order to computationally obtain the binding affinity between proteins and ligands.

Many scoring functions have been developed and can be classified in four categories acoording to the method that is used to obtain the protein-ligand binding affinity: Physics-based that uses force fields to calculate protein-ligand binding; Empirical that calculates the protein-ligand fitness through the sum of individual terms which represent important energetic factor in protein-ligand binding; Knowledge-based, that obtains the score from summing pairwise statistical potentials between protein and ligand; and Machine Learning methods that trains scoring functions using features obtained from known protein-ligand binding experiments.

For this work, 723 molecular descriptors of the protein-ligand complex collected by different software were considered as input to estimate the "pKd" attribute. To carry out the training on predictive models, the PDBbind 2018 dataset was first selected and applied several tools for data extraction in order to create an ideal dataset for the elaboration of the work. Then, the dataset is submitted to three attribute selection methods, namely: Principal Component Analysis (PCA) that consists of a multivariate statistical technique which aims to transform the set of original attributes into another set containing the new attributes that must have the same dimension, called principal components. To choose how many resources should be kept in order to maintain 90% of the data variability, the cumulative PCA sum was applied, resulting in 77 attributes that should be kept; Anova F-Value is used to compare the values of the multiple means of the dataset and check whether there is a significant difference between the mean values of different groups. Anova F-Value is used mainly when data sets are based on experimental data; Random Forest Regression is built from a set of decision trees, each of which is composed of a set of nodes or inner leaves.

For each of the attribute selection methods, the models were trained with 80% of the data destined for the training set and 20% for the test set, using three different regression methods, namely: Elastic Net Regularization (ENR) that is a weighted combination of the penalties of Ridge Regression (L2 regularization) and Lasso (L1 regularization); Neural Networks that are computational models that seek to simulate the human central nervous system, in which they are capable of machine learning as well as pattern recognition. We employed a NN with two layers of 50 neurons. The threshold was set to 0.01; Support Vector Machines are classifiers that learn hyper-planes in a multidimensional space from data input that best separate the labelled classes. For the evaluation of the models, the RMSE and Pearson Correlation metrics were used. For both metrics, the model that returns the best result was the Neural Network using the Random Forest Regression, where it obtained a RMSE of 1.03 and Pearson's correlation of 0.86.