

Master's degree in computer science

PPI Report

PPI

PPI or Protein-Protein Interaction refers to the physical and functional interactions between proteins in a biological system. These interactions are fundamental to almost every biological process, from cellular signaling to the formation of complex molecular machines.

PPIs often form the basis for identifying biomarkers:

* **Direct Biomarkers:** Disease-specific protein complexes or interaction patterns;
* **Indirect Biomarkers:** Secondary changes in interacting proteins;

PPIs are, thus, highly relevant for disease prevention and detection because they play a central role in virtually all biological processes. When these interactions are disrupted or altered, it can lead to diseases (such as Cancer, Neurodegenerative diseases and Infectious diseases). Understanding PPIs provides critical insights into disease mechanisms, enabling the development of diagnostic tools and preventive strategies.

Dataset Description

The dataset is from Kaggle: <https://www.kaggle.com/datasets/spandansureja/ppi-dataset/data>.

The dataset is comprised of 2 smaller datasets (positive\_protein\_sequences and negative\_protein\_sequences). Positive protein sequences represent **interacting** protein pair sequences and negative protein sequences represent **non-interacting** protein pair sequences.

The datasets are very identical in size (positive\_protein\_sequences contains 36,652 interacting protein pair sequences and negative\_protein\_sequnces contains 36,480 non-interacting protein pair sequences).

Description of the Process Applied

Firstly, both datasets were loaded, labeled and merged into a single dataset (with both the positive and negative protein pair interactions). After this initial aggregation, the length of each protein sequence is calculated for both proteins in each pair. Length is relevant for various reasons:

* Structural Compatibility
  + Since proteins interact through specific regions, longer proteins may offer larger interaction surfaces and/or certain functional motifs or domains may only be present in sufficiently long sequences.
* Functional Implications
  + Many proteins consist of multiple functional domains. Longer proteins are more likely to contain multiple domains, increasing the likelihood of specific interactions. On the other hand, intrinsically disordered regions, which are often longer, can play a critical role in transient interactions.

After this, amino acid composition was calculated (that is, the percentage of each amino acid in a given sequence). This is relevant because:

* Polarity and Hydrophobicity
  + Hydrophobic amino acids are often found at PPI interfaces.
  + Polar and charged amino acids often mediate interactions.

Using Biochemical properties extraction, specific properties for each sequence were calculated and appended to the Data Frame, them being:

* **Molecular Weight:** Total weight of the sequence in Daltons.
* **Aromaticity:** Fraction of aromatic amino acids.
* **Isoelectric Point:** pH at which the sequence has no net charge.
* **Secondary Structure Fraction:** Proportion of sequence forming helices, sheets, or turns.

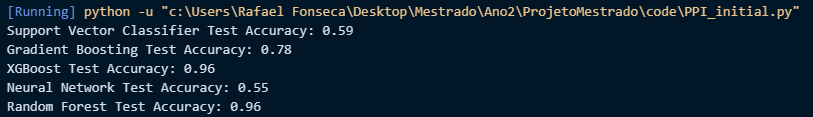
All these properties are extremely relevant providing insights into the protein’s physical, chemical and structural characteristics thus influencing greatly how they interact with other proteins.

All this information was plotted to further visualize their distributions based on the target variable (PPI).

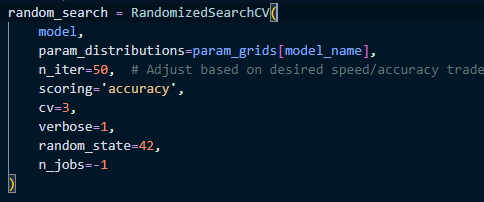
The dataset was split into 80-20 meaning that 80% of the dataset was used for training and then the remaining 20% for testing and evaluation of the models. Initially, the following models were used:

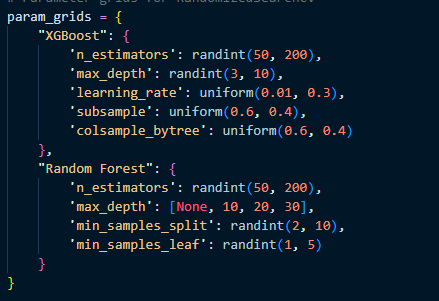
* SVC
* Gradient Boosting
* XGBoost
* Random Forest
* MLP

The models achieved the following accuracy.

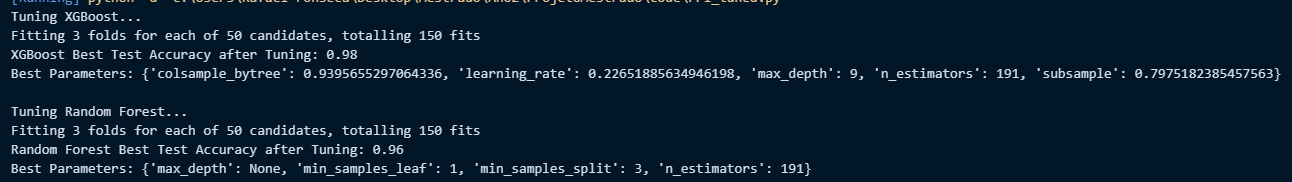


Then, XGBoost and Random Forest were tuned (using RandomizedSearchCV) to tune the parameters shown below, with the intent to further improve their already high accuracy.





The output was the following:



This shows a significant (96% to 98%) from XGBoost.

Furthermore, I also used CNN and MLP models (with the exact parameters I used for the first dataset, the one with the OCU’s here, in this dataset) having obtained 81% for the CNN and 77% for the MLP.

Result Analysis

XGBoost outperformed every model, also proven by this experiment on the same dataset on Kaggle, done by another person: <https://www.kaggle.com/code/ashwinsharma03/ppi-prediction>

Just like them, I reached the conclusion that XGBoost outperforms all the models used.

order to have medical relevance.