iNeuron

Drug Activity Prediction

HIGH-LEVEL DESIGN DOCUMENT (HLD)

Project By: -

Ashfaque Jamal

Beauty Atara

Date: - 22/09/2022

Table Of Content:

1. Introduction:

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

1.2 Scope:

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

1.3 Overview:

The HLD will:

- Present all of the design aspects and define them in detail.
- Describe the user interface being implemented.
- Describe the hardware and software interfaces.
- Describe the performance requirements.
- Include design features and the architecture of the project.
- List and describe the non-functional attributes like:
    - security
    - reliability
    - maintainability
    - portability
    - reusability
    - application compatibility
    - resource utilization
    - serviceability

2 . General Description:

2.1 Problem Statement:

Drugs are typically small organic molecules that achieve their desired activity by binding to a target site on a receptor. The first step in the discovery of a new drug is usually to identify and isolate the receptor to which it should bind, followed by testing many small molecules for their ability to bind to the target site. This leaves researchers with the task of determining what separates the active (binding) compounds from the inactive (non-binding) ones. Such a determination can then be used in the design of new compounds that not only bind, but also have all the other properties required for a drug (solubility, oral absorption, lack of side effects, appropriate duration of action, toxicity, etc.). The goal of this project is to allow us to develop predictive models that can
determine given a particular compound whether it is active (1) or not (0). As such, the goal would be developing the best binary classification model. A molecule can be represented by 100000 binary features which represent their topological shapes and other characteristics important for binding.

2.2 Project Perspective:

In this project, various experiments were conducted using various *Classification methods*, *Ensemble methods* with various combinations of *dimensionality reduction* algorithms along with *Resampling* since it is class imbalanced data. There was training data which was fed to the model which contained 800 drugs along with their activity and features affecting the activity. After trying all the experiments with different combinations of above parameters, the best model was arrived at. The resulting model was then used to compute the activities of the drugs in the test dataset into active or inactive classes.

2.3 Tools used:

Python Programming language, Jupyter notebook, Pandas, Numpy, Matplotlib, Seaborn, Sci-kit learn, Streamlit.

2.4 Constraints:

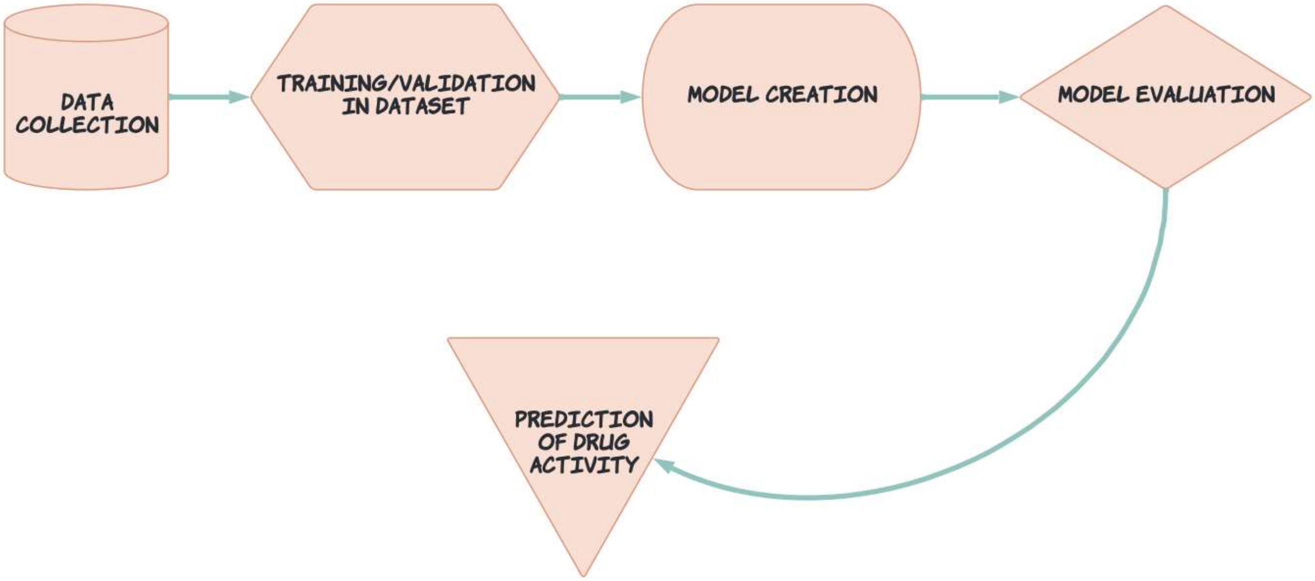
The Drug Activity Prediction Model should be able to correctly predict the classes of the drug and the results must be reliable. The user should not be required to know anything about the working of the model. When the model is deployed, the user should provide the details of the drug and the model will predict whether the given compound is active or not.

2.5 Assumptions:

The main objective of this project is to develop predictive models that can determine given a particular compound whether it is active (1) or not (0). Various Machine Learning Models were used for this task. It is also assumed that all the aspects of this project have the ability to work together in the way the user is expecting.

3. Design Details:

3.1 Proposed Methodology:



Data Collection:
The collected data was in two parts, training data and testing data. The data was in .dat format, so first of all we converted it into .csv format. The training data contained 800 rows and the testing data contained 350 rows. The data was then expanded to a dimension of 800*1000000 and 350*1000000 for training and testing data respectively.

Training/Validation on dataset:
After doing necessary data cleaning and data pre-processing, the dataset was fed to the models for training.

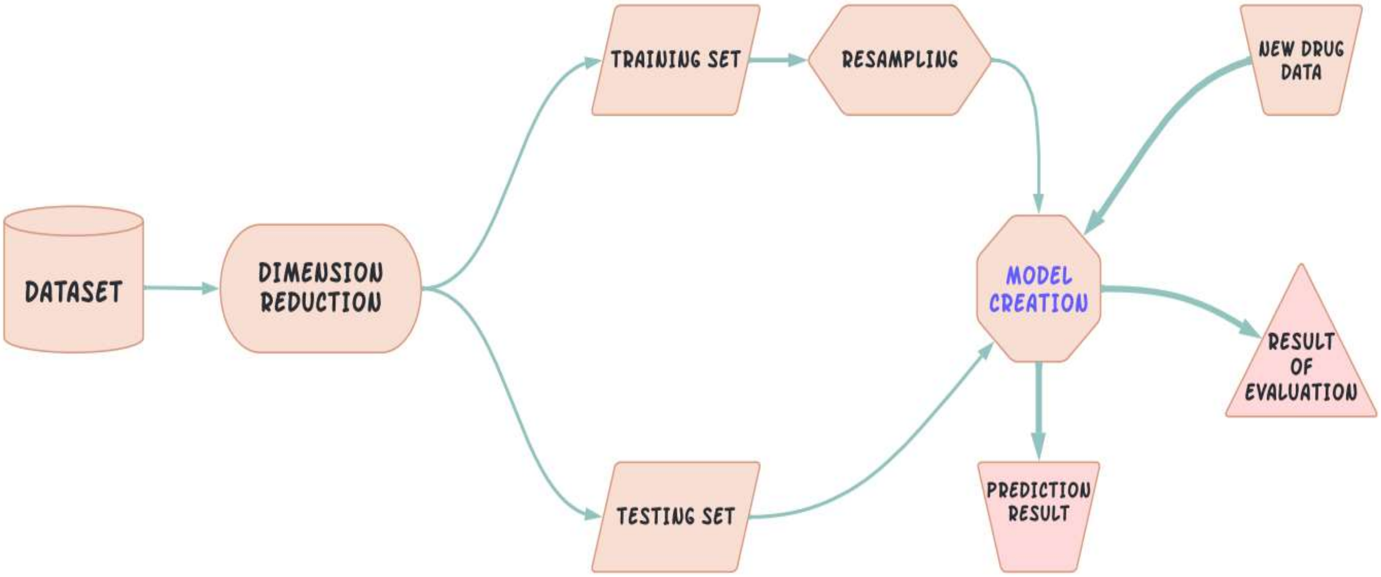
Model Creation:
Different Machine Learning Classification models were created using the training dataset and Hyperparameter tuning was done for all the models.

Model Evaluation:
After Hyperparameter tuning, best models were selected for evaluation using the testing dataset.

Prediction of drug activity:

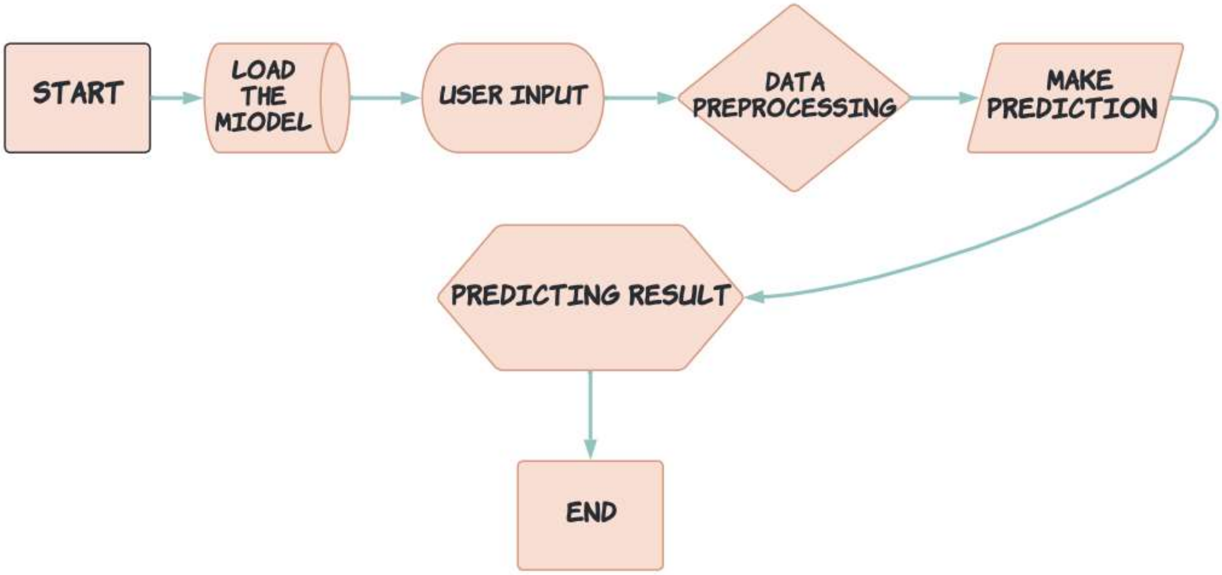
The model giving the best result was used to make the prediction of the drug activity.

3.2 Model Training and Evaluation:



The dataset collected first passes through data cleaning and data pre-processing. Then the Dimension of the dataset is reduced using PCA. The dataset is split into training and testing data. Since, the data was imbalanced, resampling of training data is done. The resampled dataset is then used for training the model. The testing dataset is used for model evaluation. The best model is selected, and new data is passed to the model and the model will make the predictions.

3.3 Deployment Process:



3.4 Error Handling:

Should errors be encountered, an explanation will be displayed as to what went wrong. An error will be defined as anything that falls outside the normal and intended usage.

3.5 Performance:

The Drug Activity Prediction Model will help the users to predict whether the particular drug molecule is active (binding) or inactive (non-binding). This information will help the researchers in the design of new compounds that not only bind, but also have all the other properties required for a drug (solubility, oral absorption, lack of side effects, appropriate duration of action, toxicity, etc.)

3.6 Reusability:

The code written and the components used should have the ability to be reused with no problems.

3.7 Application compatibility: The different components for this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of the Python code to ensure proper transfer of information.

3.8 Deployment: Streamlit, Flask, Django (Images of all these)

4. Conclusion:

The Project will help to detect the activity of the drug compound, which will in turn be useful for the researchers in order to design new compounds that not only bind, but also have all the other properties required for a drug (solubility, oral absorption, lack of side effects, appropriate duration of action, toxicity, etc.). This project will be very helpful in the process of a new drug discovery.