Module 7

XN PROJECT: PROJECT SCOPE

ALY 6080 Integrated Experiential Learning

Instructor: Prof. Allen Bryce

Northeastern University

College of Professional Studies, Northeastern University, Boston, MA

Submitted By: Group 2

Alok Madamanchi

Kashika Tyagi

Md Tajrianul Islam

Sandeep Karapa Srivatsav

Spurthi Patnam

Sweta Mankala

XN: PROJECT SCOPE

The purpose of this assignment is to have a well-defined scope of the project which is necessary to ensure the successful completion of the project. The project we are working on is based on Leukemia disease. We aim to analyze the top 10 targets of Leukemia Indication for our sponsor - Silicon Therapeutics.

I. Project Needs:

Due to the recent developments in technology and the availability of huge amounts of data in the drug discovery and bioactivity domain, it has become possible to store and retrieve data required for our research and analysis. With the help of publicly available data on various drug discovery databases such as ChEMBL, we can discover significant patterns to identify which target or drug should our Sponsor (Silicon Therapeutics) target next or consider for drug discovery".

II. Scope of the Project

Leukemia Target Discovery project aims to analyze and predict the top 10 targets best- suitable for leukemia drug discovery. Though leukemia drugs work quite well in treating children, they are not very effective on adults as it seems to be recurring in them. This concern is the principal intention of screening the targets that are best suitable for drug discovery.

III. Scope Description

For data preparation, an understanding of the objective, target types, potential mechanisms that need to be described is important, which can help in guiding the preparation and transformation of the data. Thus, we began with understanding the data by performing descriptive statistics in Python such as identification of variables and their data types, missing values, count of unique values in the Target Type variable and their distribution.

Python provides great flexibility and readability, and R has many packages that assist in delivering statistical analysis. We plan to use the combination of Python, R, and Tableau for our analysis. We performed data extraction, cleaning, univariate and multivariate analysis applying the python's pandas and numpy package. Also, generated few visualizations such as the top 10 mechanisms of actions, Max phase of indication vs. targets, etc. using python and tableau.

IV. Expectations and Acceptance

Our sponsor, Silicon Therapeutics, is a physics-based drug discovery organization. It expects us to analyze the CHEMBL database to analyze and predict the targets for leukemia. This analysis will be effective and valuable for the sponsor, as we will be delivering an accurate analysis of this data by applying techniques such as hypothesis testing, logistic regression, and random forests. The process to narrow down and identify targets would require identification of constraints to establish a strong relationship between small molecules and target compounds.

V. Constraints:

One of the first constraints of the project will be prioritization of targets. After going through the ChEMBL database we have observed that there are over 400 small molecules that have been tested for around 300+ different types of targets. Although, there are some open source platforms like OpenTarget that use the same database to rank these targets in order after calculating some certain values like evidence, association score and target safety etc. We have to either ensure the trustworthiness of the platform or follow similar procedure, which could be a lengthy procedure for this project.

Also, as we try to predict or suggest a small molecule for further research it is important for us to early identify a compound’s potential off-target effects. It is not just desirable - it is essential for proper strategic planning as this might halt our entire progress in the project.

VI. Necessary Changes:

Our analysis revolves around using Python and Tableau for exploratory data analysis and R for statistical analysis. We have to ensure our analysis is not repeated on both the programming languages. We will have to segregate our analysis and visualize our results on Tableau. The reports generated will depict insights of the targets for the disease 'Leukemia' and generate outcomes pertaining to the target molecules and help us identify the most appropriate targets we should focus on to perform further analysis.