1.Frame the problem and look at the big picture.

1. Define the objective in business terms.

Ans: In business terms it will be beneficial as the doctors won’t have to do too many traditional tests & so simply our Drug Discovery System will be providing time benefits to the doctors & the health of the patients can be determined easily & prescription will also be done in a proper time without too much delay.

a.With this, cost of various apparatus will be cut down.

b.Time will be saved as Drug Discovery System will generate the desired results properly & swiftly.

c.Doctors will be able to deal with large number of patients as their time for treating each patient will be decreased.

Thus, overall business of the hospital will be enhanced.

2. How will your solution be used?

Ans: It will help the doctors to analyze the health of the patients & so accordingly they can take the decisions more swiftly as compared to the traditional methods.

4. How should you frame this problem (supervised/unsupervised, online/offline,

etc.)?

Ans: Supervised because we have defined prediction labels and features in this system.

5. How should performance be measured?

Ans: We’ll use various kind of models of Machine Learning to measure the performance of the Drug Discovery System.

6. Is the performance measure aligned with the business objective?

Ans: Yes, as this system will analyze the severity of the disease and accordingly will provide the rating so as to treat the patient as soon as possible.

Ex: Diagnosis scale ranges from 1 to 5 based on the severity of the disease, 5-represents critical condition, 4-represents severe requires immediate treatment, and 3-represents moderate requires further investigation, 2-represents normal, 1-represents within control

7. What would be the minimum performance needed to reach the business objective?

Ans: firstly we will try to reach the accuracy level of 70% and then proceed further.

8. What are comparable problems? Can you reuse experience or tools?

Ans: Yes we can reuse the experience & tools for various other comparable problems like detecting the chances for getting Heart Attack, detecting dermatology prescription & we can also prescribe diet chart to the patients.

9. Is human expertise available?

Ans: Yes, we have many experts who can suggest us about the results & the implementation of the model inside the system.

10. How would you solve the problem manually?

Physician notes, medical history, medical prescription, lab and scan reports will be helpful in solving the problem manually.

2.Get the Data

Note: automate as much as possible so you can easily get fresh data.

1. List the data you need and how much you need.

Dataset of the patients which will be helpful in recognising that which drug has been taken by the patient & till what time it’s effect will be there on the body.

2. Find and document where you can get that data

We get our dataset from ---.

3. Check how much space it will take.

The dataset is of approx 36MB and some other space required by the program.

4. Check legal obligations, and get authorization if necessary.

It is legally provided.

5. Get access authorizations.

It is publically accessible.

6. Create a workspace (with enough storage space).

7. Get the data.

We get our data from these links https://www.ebi.ac.uk/chembl/

8. Convert the data to a format you can easily manipulate (without changing the

data itself).

The dataset in .CSV form so after that we can import it easily and can also go though with the analysis of data in that format by just looking to the data.

9. Ensure sensitive information is deleted or protected (e.g., anonymized).

Ans: Yes

10. Check the size and type of data (time series, sample, geographical, etc.).

Ans: Approx 36 MB.

11. Sample a test set, put it aside, and never look at it (no data snooping!).

3.Explore the Data

Note: try to get insights from a field expert for these steps.

1. Create a copy of the data for exploration (sampling it down to a manageable size

if necessary).

Ans: Done

2. Create a Jupyter notebook to keep a record of your data exploration.

Ans: Done

3. Study each attribute and its characteristics:

• Name

• Type (categorical, int/float, bounded/unbounded, text, structured, etc.)

% of missing values

• Noisiness and type of noise (stochastic, outliers, rounding errors, etc.)

• Possibly useful for the task?

• Type of distribution (Gaussian, uniform, logarithmic, etc.)

4. For supervised learning tasks, identify the target attribute(s).

5. Visualize the data.

Ans: Done

6. Study the correlations between attributes.

Ans: Done

7. Study how you would solve the problem manually.

Ans: By reading various manual reports generated in real time without applying Deep Learning.

8. Identify the promising transformations you may want to apply.

9. Identify extra data that would be useful

Ans: Measures such as Classification error, Computational cost, Accuracy, Loss, EOR (enrichment over random) can be used for calculating the accuracy of drug discovery using neural network.

10. Document what you have learned.

Ans: We’ve learned that how we can implement the CNN to discover drugs. Self-Organizing Map (SOM) or Kohonen neural network usage.

**4.Prepare the Data**

Notes:

• Work on copies of the data (keep the original dataset intact).

We will collect the dataset of very large amount and then split into the 4 parts 2 training and 2 testing and test it and this will be implemented on all the models.

Write functions for all data transformations you apply, for five reasons:

* [General data manipulation and preparation](https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/data-transformation-manipulation): Merging datasets, cleaning missing values, grouping and summarizing data, changing column names and data types, or indicating which column is a label or a feature.
* [Sampling and splitting datasets](https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/data-transformation-sample-and-split): Divide your data into training and test sets, split datasets by percentage or by a filter condition, or perform sampling.
* [Scaling and reducing data](https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/data-transformation-scale-and-reduce): Prepare numerical data for analysis by applying normalization or by scaling. Remove or replace outliers.

So you can apply these transformations in future projects?

We will apply these basic transformation in our future projects also for better accuracy of model and better understanding of machine.

To clean and prepare the test set?

1. Data cleaning:

• Fix or remove outliers (optional).

• Fill in missing values (e.g., with zero, mean, median…) or drop their rows (or columns)

We have these methods for filling missing values-:

1.Deleting the rows.

2.Replacing with mean/median/mode.

3.Assign an unique category.

4.Predicting the missing values.

5.Use of algorithm which support missing values like- k-nearest, random forest etc.

we apply all these and then calculate the accuracy and after that we proceed further for the best possible method

2. Feature selection (optional):

We will do feature selection later now we are computing accuracy on different dataset.

• Drop the attributes that provide no useful information for the task.

3. Feature engineering, where appropriate:

• Discretize continuous features.

**6. Decompose features (e.g., categorical, date/time, etc.).**

. Feature scaling: standardize or normalize features.

**Feature Scaling** is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing. we will do feature scaling if required.

• If the data is huge, you may want to sample smaller training sets so you can train

many different models in a reasonable time (be aware that this penalizes complex

models such as large neural nets or Random Forests).

Our data is huge and we will try to implement this on the different models and try to find the accuracy from all the models and which gives the best result then we will select and try to proceed further.

• Once again, try to automate these steps as much as possible.

For better implementation we made function which takes parameter of dataset set and inside that all models implementations are there which gives its accuracy rate.