# Programming Part

## Web UI for sell medicine

---

## Machine learning for diagnosis

### Data collection

we collect the data to make the dataset from a study of university of Columbia performed at New York Presbyterian Hospital during 2004. <https://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html>

### Library Used

In this project standard libraries for database analysis and model creation are used. The following are the libraries used in this project.

#### Tkinter

Tkinter is a standard GUI library of python. Python when combined with tkinter provides fast and easy way to create GUI. It provides powerful object-oriented tool for creating GUI. It provides various widgets to create GUI some of the prominent ones being :

* 1. Button
  2. Canvas
  3. Label
  4. Entry
  5. Check Button
  6. List box
  7. Message
  8. Text
  9. Message box

Some of these were used in this project to create our GUI namely messagebox, button, label,

Option Menu, text and title. Using tkinter we were able to create an interactive GUI for our

model.

#### Numpy

Numpy is core library of scientific computing in python. It provides powerful tools to

deal with various multi-dimensional arrays in python. It is a general purpose array processing

package.

Numpy’s main purpose is to deal with multidimensional homogeneous array. It has tools ranging

from array creation to its handling. It makes it easier to create a n dimensional array just by using

*“np.zeros()”* or handle its contents using various other methods such as replace, arrange, random,save, load it also helps I array processing using methods like sum, mean, std, max, min, all, etc.

Array created with numpy also behave differently then arrays created normally when they are

operated upon using operators such as +,-,\*,/.

All the above qualities and services offered by numpy array makes it highly suitable for our

purpose of handling data. Data manipulation occurring in arrays while performing various

operations need to give the desired results while predicting outputs require such high operational capabilities.

#### Pandas

pandas is the most popular python library used for data analysis. It provides highly

optimized performance with back-end source code purely written in C or python. Now we choose python

Data in python can be analysed with

1. Series is one dimensional array defined in pandas used to store any data type.
2. Dataframes are two-dimensional data structure used in python to store data consisting of rows and columns.

Pandas dataframe is used extensively in this project to use datasets required for training and

testing the algorithms. Dataframes makes it easier to work with attributes and results. Several of

its inbuilt functions such as replace were used in our project for data manipulation and

preprocessing.

#### Sklearn

Sklearn is an open source python library with implements a huge range of machine-

learning, pre-processing, cross-validation and visualization algorithms. It features various simple

and efficient tools for data mining and data processing. It features various classification,

regression and clustering algorithm such as support vector machine, random forest classifier,

decision tree, gaussian naïve-Bayes, KNN to name a few.

In this project we have used sklearn to get advantage of inbuilt classification algorithms like

decision tree, random forest classifier, KNN and naïve Bayes. We have also used inbuilt cross

validation and visualization features such as classification report, confusion matrix and accuracy

score.

### Model

There are four different kind of models present in our project to predict the disease these are

#### *Decision tree*

Decision tree is classified as a very effective and versatile classification technique. It is used in

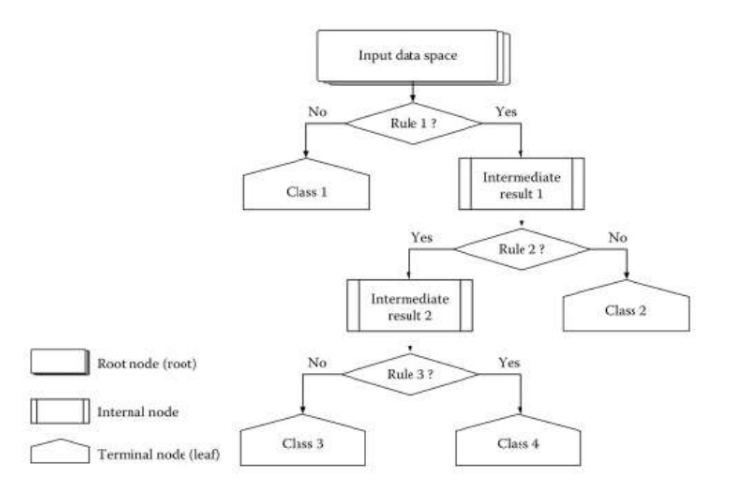
pattern recognition and classification for image. It is used for classification in very complex

problems dew to its high adaptability. It is also capable of engaging problems of higher

dimensionality. It mainly consists of three parts root, nodes and leaf.

Roots consists of attribute which has most effect on the outcome, leaf tests for value of certain

attribute and leaf gives out the output of tree.



Decision tree is the first prediction method we have used in our project. It gives us an accuracy

of ~ 95%.

#### Random Forest tree

Random Forest Algorithm is a supervised learning algorithm used for both classification and

regression. This algorithm works on 4 basic steps –

1. It chooses random data samples from dataset.
2. It constructs decision trees for every sample dataset chosen.
3. At this step every predicted result will be compiled and voted on.
4. At last most voted prediction will be selected and be presented as result of classification.

In this project we have used random forest classifier with 100 random samples and the result

given is ~ 95% accuracy.

#### K Nearest Neighbour

K Nearest Neighbour is a supervised learning algorithm. It is a basic yet essential algorithm. It

finds extensive use in pattern finding and data mining.

It works by finding a pattern in data which links data to results and it improves upon the patter

recognition with every iteration.

We have used K Nearest Neighbour to classify our dataset and achieved ~ 92% accuracy.

#### Naïve Bayes algorithm

Naïve Bayes algorithm is a family of algorithms based on naïve bayes theorem. They share a

common principle that is every pair of prediction is independent of each other. It also makes an

assumption that features make an independent and equal contribution to the prediction.

In our project we have used naïve bayes algorithm to gain a ~ 95% accurate prediction.