**Predict Influencers in the Social Network**

The binary label represents a human judgement about which one of the two individuals is more influential. A label '1' means A is more influential than B. 0 means B is more influential than A. The goal of the challenge is to train a machine learning model which, for pairs of individuals, predicts the human judgement on who is more influential with high accuracy.

In this paper author is using Logistic Regresion, SVM (Support Vector Machine), naïve bayes, and MLP(Neural Network) Classifier

All the algorithms generate model from train dataset and new data will be applied on train model to predict it class. Random Forest algorithm is giving better prediction accuracy compare to all other algorithm.

**Logistic Regression:**

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression[1] (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a *logit*, from *logistic unit*, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

**Support vector machine:**

Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyper plane which separates the data into classes. In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. As a simple example, for a classification task with only two features (like the image above), you can think of a hyper plane as a line that linearly separates and classifies a set of data.

Intuitively, the further from the hyper plane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyper plane as possible, while still being on the correct side of it.

So when new testing data is added, whatever side of the hyper plane it lands will decide the class that we assign to it.

How do we find the right hyper plane?

Or, in other words, how do we best segregate the two classes within the data?

The distance between the hyper plane and the nearest data point from either set is known as the margin. The goal is to choose a hyper plane with the greatest possible margin between the hyper plane and any point within the training set, giving a greater chance of new data being classified correctly.

Both algorithms generate model from train dataset and new data will be applied on train model to predict it class. SVM algorithm is giving better prediction accuracy compare to ANN algorithm.

**Naïve Bayes Classifier Algorithm**

It would be difficult and practically impossible to classify a web page, a document, an email or any other lengthy text notes manually. This is where Naïve Bayes Classifier machine learning algorithm comes to the rescue. A classifier is a function that allocates a population’s element value from one of the available categories. For instance, Spam Filtering is a popular application of Naïve Bayes algorithm. Spam filter here, is a classifier that assigns a label “Spam” or “Not Spam” to all the emails.

Naïve Bayes Classifier is amongst the most popular learning method grouped by similarities that works on the popular Bayes Theorem of Probability- to build machine learning models particularly for disease prediction and document classification. It is a simple classification of words based on Bayes Probability Theorem for subjective analysis of content.

**Multilayer perceptron (MLP):**

A **multilayer perceptron**(MLP) is a class of feed forward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to refer to *any* feed forward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons (with threshold activation); see § Terminology. Multilayer perceptrons are sometimes colloquially referred to as "vanilla" neural networks, especially when they have a single hidden layer.

An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

To implement above all algorithms we have used python technology and ‘PeerIndex’ dataset.This dataset available inside dataset folder which contains test dataset with dataset information file.

Python Packages and Libraries used: Numpy, pandas, tkinter,

|  |  |  |
| --- | --- | --- |
| PyVISA | 1.10.1 | 1.10.1 |
| PyVISA-py | 0.3.1 | 0.3.1 |
| cycler | 0.10.0 | 0.10.0 |
| imutils | 0.5.3 | 0.5.3 |
| joblib | 0.14.1 | 0.14.1 |
| kiwisolver | 1.1.0 | 1.1.0 |
| matplotlib | 3.1.2 | 3.1.2 |
| nltk | 3.4.5 | 3.4.5 |
| numpy | 1.18.1 | 1.18.1 |
| opencv-python | 4.1.2.30 | 4.1.2.30 |
| pandas | 0.25.3 | 0.25.3 |
| pip | 19.0.3 | 20.0.1 |
| pylab | 0.0.2 | 0.0.2 |
| pyparsing | 2.4.6 | 2.4.6 |
| python-dateutil | 2.8.1 | 2.8.1 |
| pytz | 2019.3 | 2019.3 |
| pyusb | 1.0.2 | 1.0.2 |
| scikit-learn | 0.22.1 | 0.22.1 |
| scipy | 1.4.1 | 1.4.1 |
| seaborn | 0.9.0 | 0.9.0 |
| setuptools | 40.8.0 | 45.1.0 |
| six | 1.14.0 | 1.14.0 |
| sklearn | 0.0 | 0.0 |
| style | 1.1.6 | 1.1.6 |
| styled | 0.2.0.post1 | 0.2.0.post1 |

accuracy score, train\_test\_split, K-Fold, cross\_val\_score, Grid Search CV, LogisticRegression, SVC, naive\_bayes, and MLP Classifier.

Modules:

Module1: Upload the training data from the Dataset Folder

Module2: Preprocess the uploaded data. We are creating New feature, checking duplicates and removing, Scaling the data, Spliting the data into train and test samples

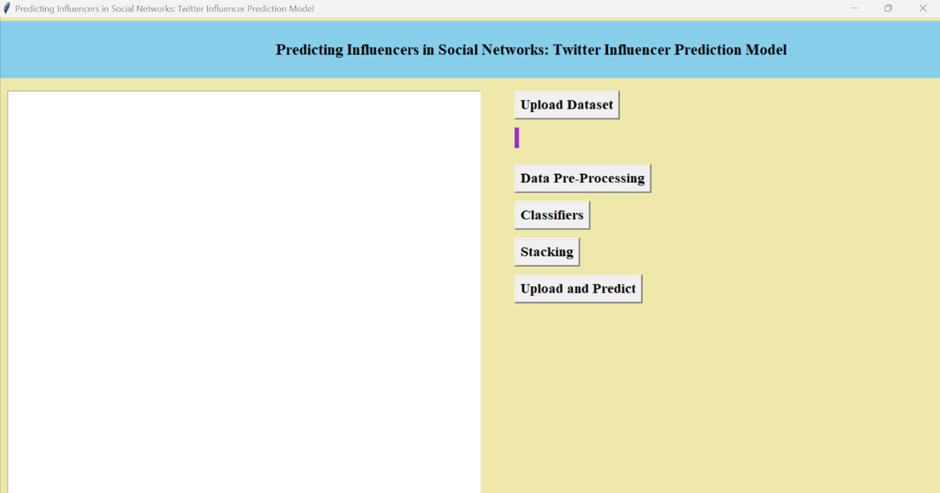
Module3: Build the model for SVM, Logistic Regression, Naive Bayes and Neural Network using MLP for all the featues.

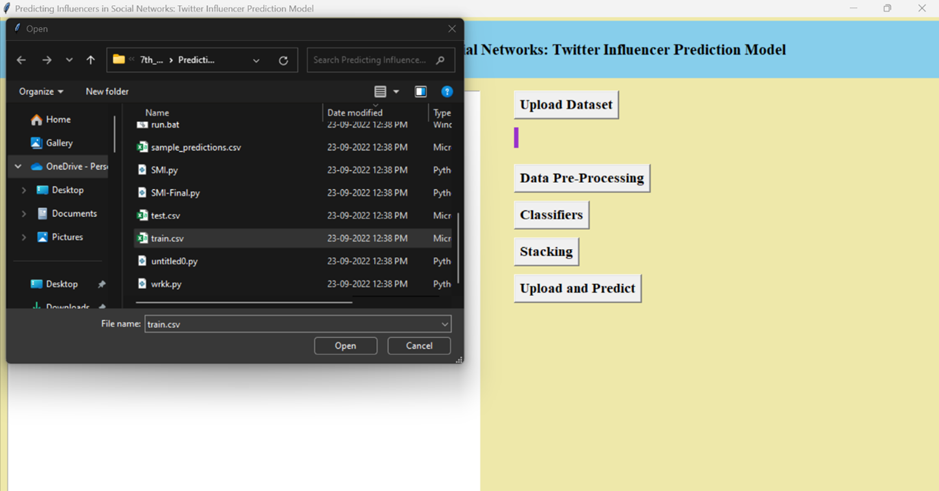
Module4: Now taking 6 Least significant features of data and Removing and creating New train and test sample then we are build above models

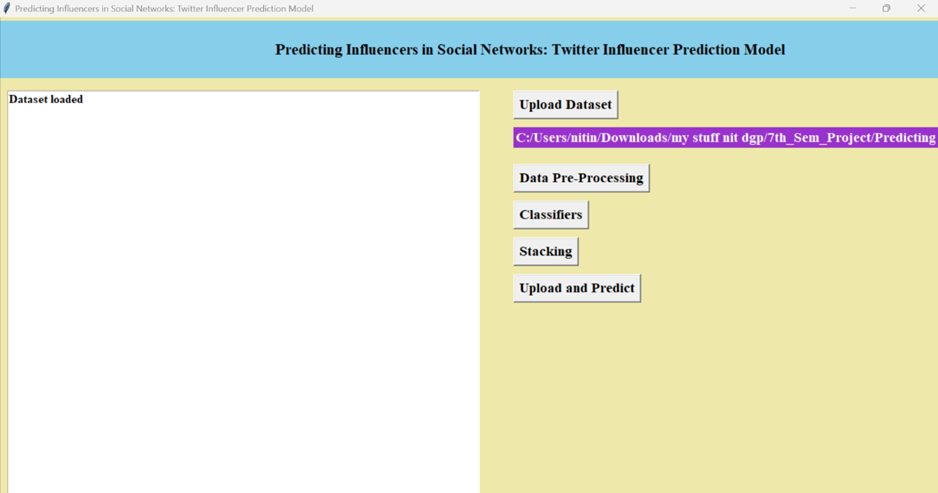
Module5: As we see in the module, SVM has best accuracy for testing data. Now we are predict new test samples.

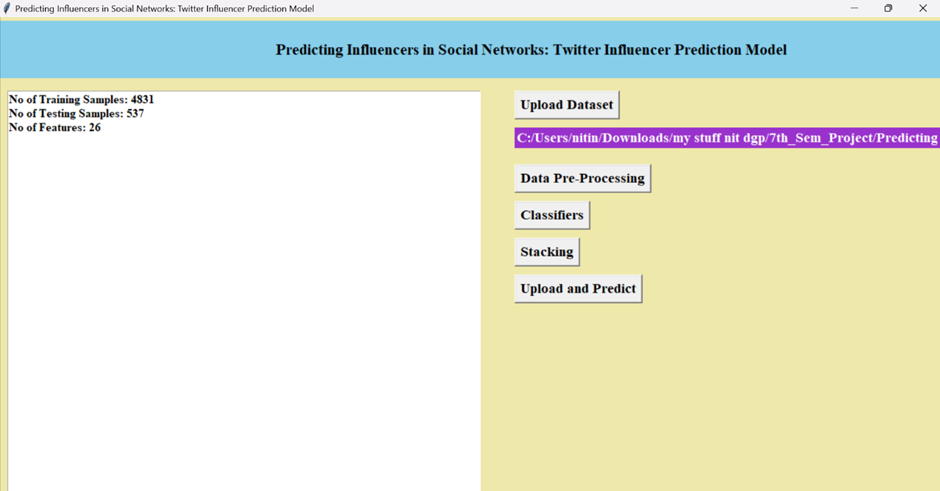
Screen shots

When we run the code it displays below window

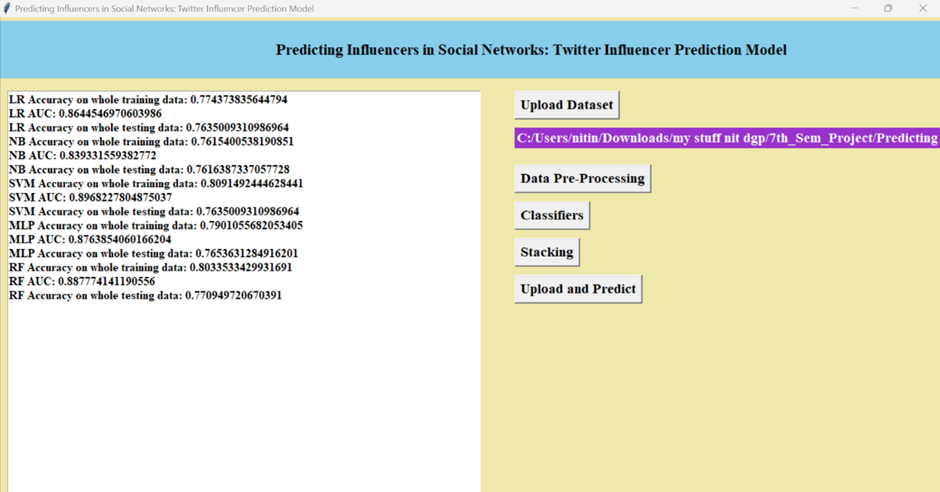


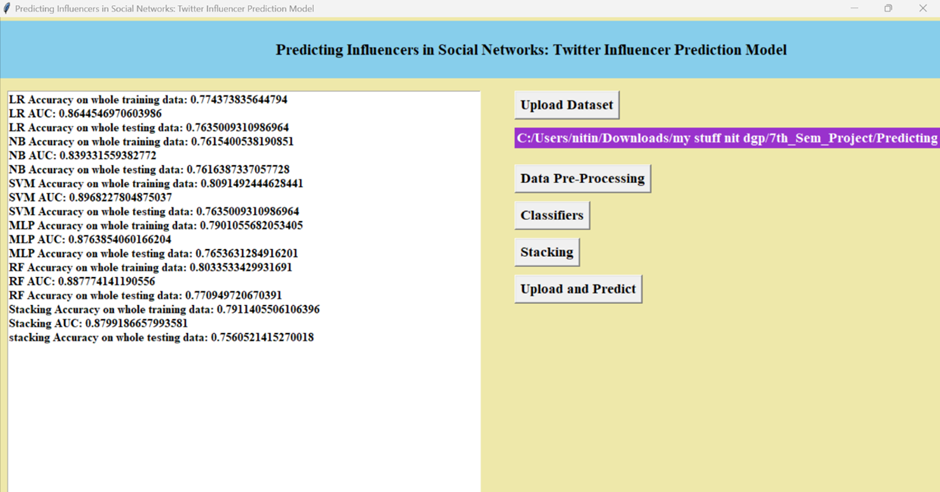
Now click on ‘upload dataset’ to upload the data



Now click on ‘Pre-Precossing’ it preprocess the data

Now click oClassifiers’ to classify the models with all the features



Now click on ‘Select best Features and Classifier’

Now click on Upload and Predict to Predict on test data.

