Construction site accident analysis using text mining and natural language processing techniques

In this paper author is describing concept to provide safety to workers at construction site from accidents by analysing past accident data by using machine learning algorithms and text mining technique such as TF-IDF (Term Frequency-Inverse Document Frequency) and natural language text processing to remove special symbols, stop words, stemming etc.

Past accident data contains details of accidents and by building machine learning algorithms model we can analyse data to identify cause of accident and can prevent future accident by giving test data of new work to predict causes of accidents and can avoid such causes. This machine learning algorithms can help in extracting dangerous objects such as misused tools, sharp objects nearby, damaged equipment etc.

In this paper to provide safety to workers author covering below points

1. Various texting mining and NLP techniques are explored with respect to construction site accidents analysis. Using this technique we will remove stop words, punctuations, special symbols and apply stemming technique to clean past accident data. After data cleaning we will convert all text data to numeric vector by using TF-IDF technique. TF-IDF contains frequency weight of each word in vector and using this vector we will build machine learning train model. Whenever we give new test data then that test data also convert to TF-IDF and then apply on train model to search for similar data and give output of similar data as prediction. Below example describe how to convert text to TF-IDF vector.

Suppose I have 3 sentences

Sentence 1: An apple a day keep doctor away

Sentence 2: apple good for health

Sentence 3: shipment of gold damage in fire

First we remove stop words such as ‘an, a, of, in’ from sentences and then take remaining words and form columns of vector. After forming columns put each word count as values of that vector. See below vector columns

Apple day keep doctor away good health shipment gold damage fire

Sentence1 1 1 1 1 1 0 0 0 0 0 0

Sentence2 1 0 0 0 0 1 1 0 0 0 0

Sentence3 0 0 0 0 0 0 0 1 1 1 1

So I convert all 3 sentences to TF-IDF vector just by putting count of each word as vector values, if sentence contains that column word then we will put its count, if sentence not contains work then we put 0 as that column values. Now to check similarity we can multiply one row with other and if multiply value greater than 0 then two sentences contains similarity otherwise not.

In above matrix if multiply sentence 1 row with sentence 2 row then we get value greater than 0 and similarity is there as both sentences contains 1 common word called ‘apple’. Similarly if we multiply sentence1 row with sentence3 row then we will get value 0 which means similarity not there between sentence 1 and 3 and we can see there is no common words in sentence 1 and 3.

1. Ensemble algorithm which has not been well studied in this field is proposed to classify the causes of accidents and SQP algorithm is utilized to search for optimal weighs of the ensemble model. In this technique we will use ensemble algorithms such as random forest and voting classifier with SQP (Sequential Quadratic Programming) to classify causes of accidents. Using SQP we can assign weight to the classifier which can help classifier in predicting correct causes of accident.
2. A rule based chunker is developed for dangerous objects extraction. Neither SQP optimization algorithm nor rule based chunker with regard to this field is found in the state of the art. Rule based chunker means getting Part Of Speech (POS) of each sentence to find dangerous object detection. When we apply POS on sentence then all dangerous objects will come under NOUN POS and by extracting noun phrases from sentences we can identify what are the dangerous objects which causes accidents.

To implement this project author using OSHA dataset and effectiveness of the proposed approaches is verified by the experiment results. OSHA dataset contains past accident data and by using this dataset we will analyse performance of various machine learning algorithms such as SVM, Decision Tree, Naïve Bayes, Logistic Regression, KNN, Ensemble Random Forest and Propose Voting Classifier which will build on all 5 base classifiers such as SVM, Naïve Bayes, Decision Tree, KNN and Logistic Regression. Voting classifier take all 5 classifier and then vote each classifier and whatever classifier give better accuracy then voting will choose that classifier for future data prediction.

Dataset saved inside ‘dataset/OSHA.csv’ folder and you can open and see the details and below are some data of new work and it has no details what accident can cause by doing that work but machine learning can predict and display future accident cause.

**cutting down a large horizontal pipe block**

**installing roof decking on a flat roof by carrying and placing decking material**

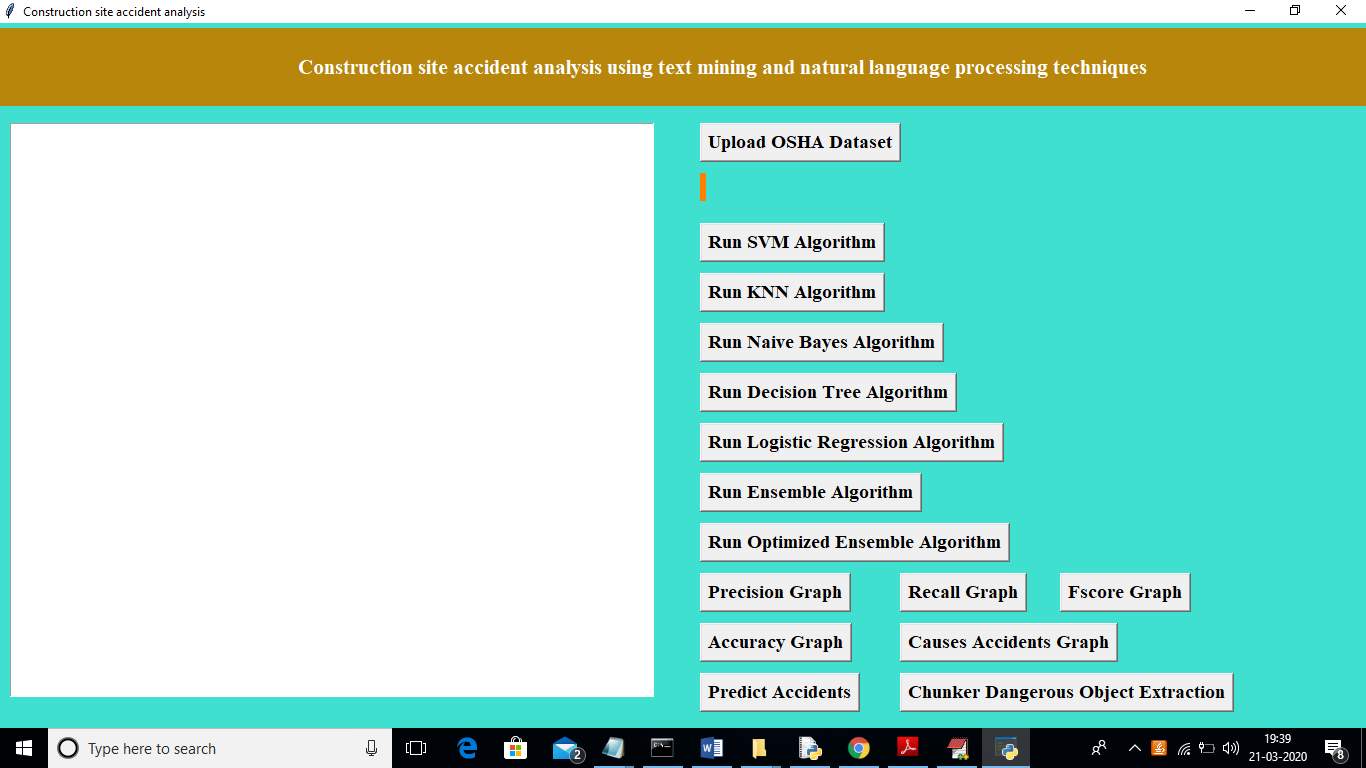
**portable storage tank and a running powered industrial truck Caught in or between**

**electrical transformer box distribution line electrocuted electric ladder work onto a 13800 volt power line work electrical parts**

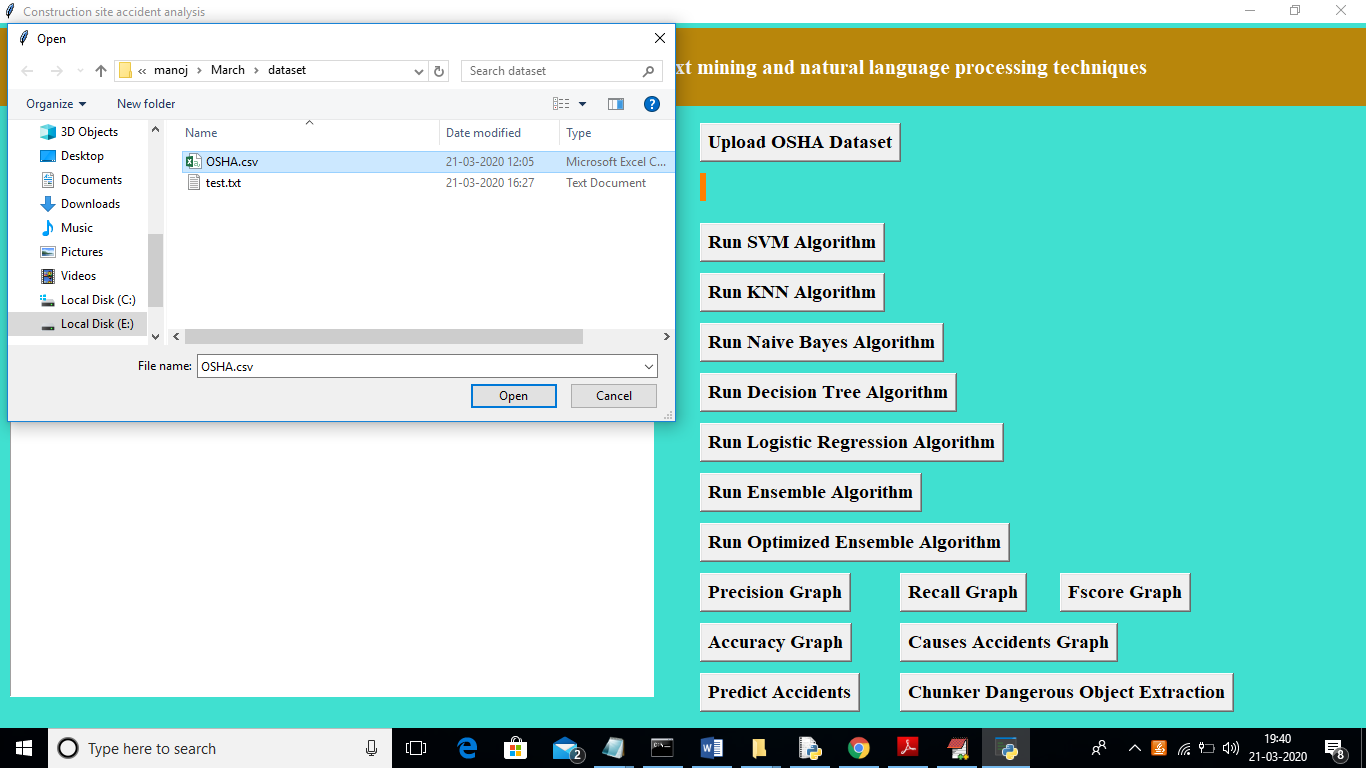
In above bold sentences some work details are there and while doing such work what accident can happen can be predicted with machine learning algorithms.

Screen shots

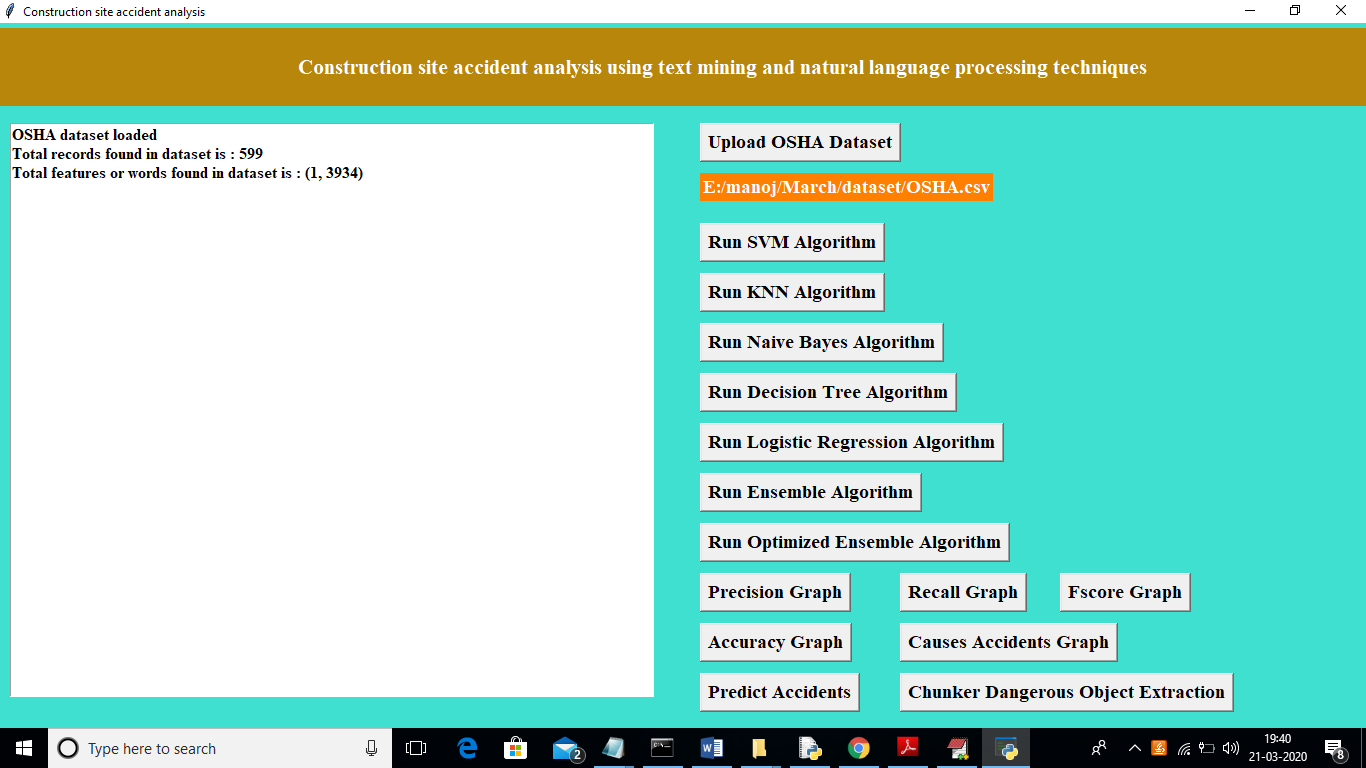
To run this project double click on ‘run.bat’ file to get below screen



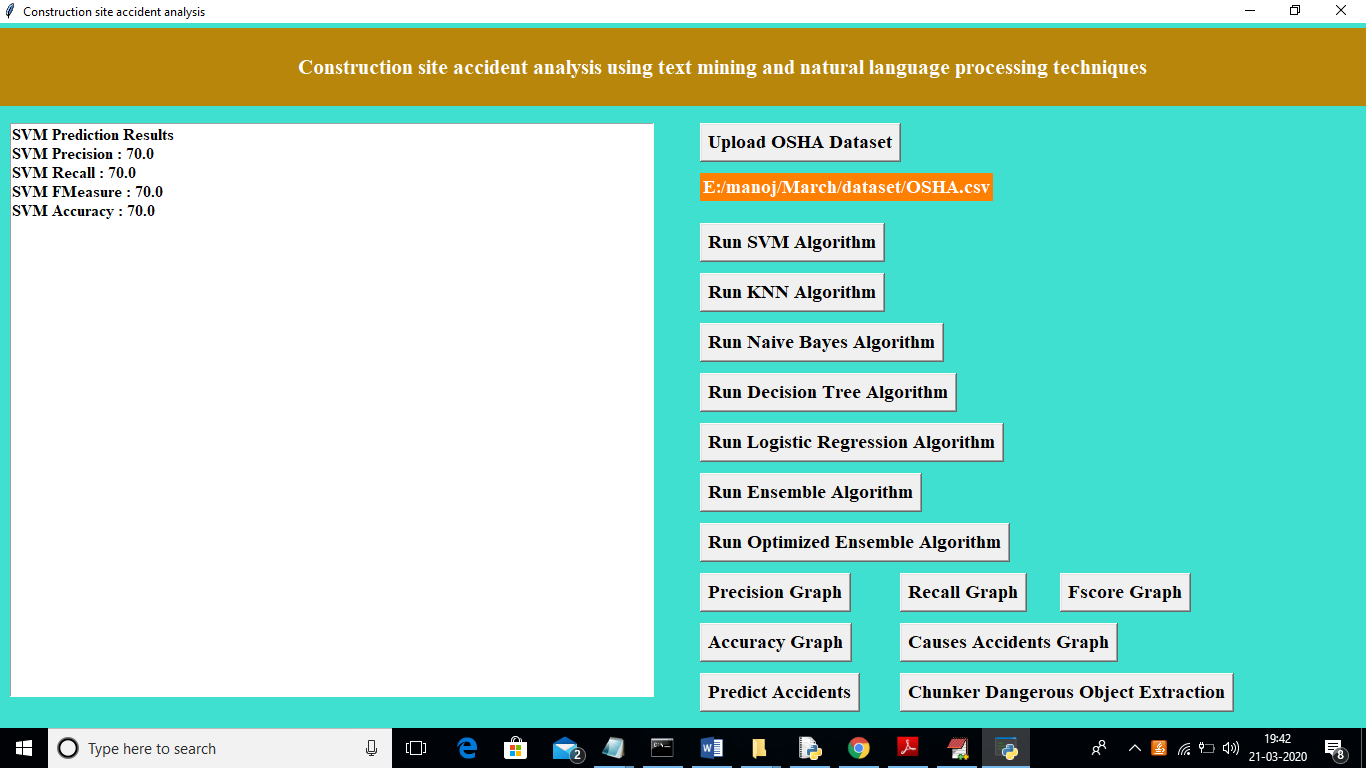
In above screen click on ‘Upload OSHA Dataset’ button and upload dataset



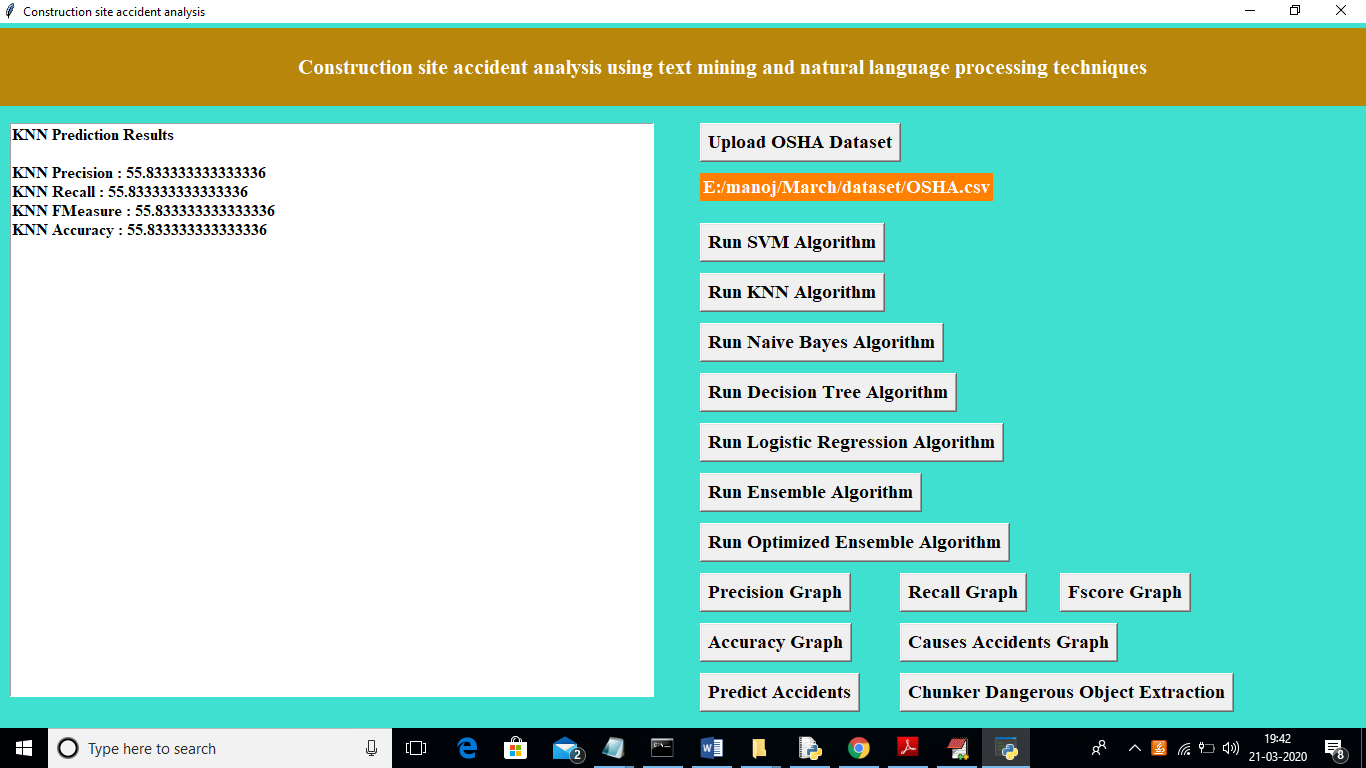
In above screen I am uploading ‘OSHA.csv’ dataset and after uploading dataset will get below screen



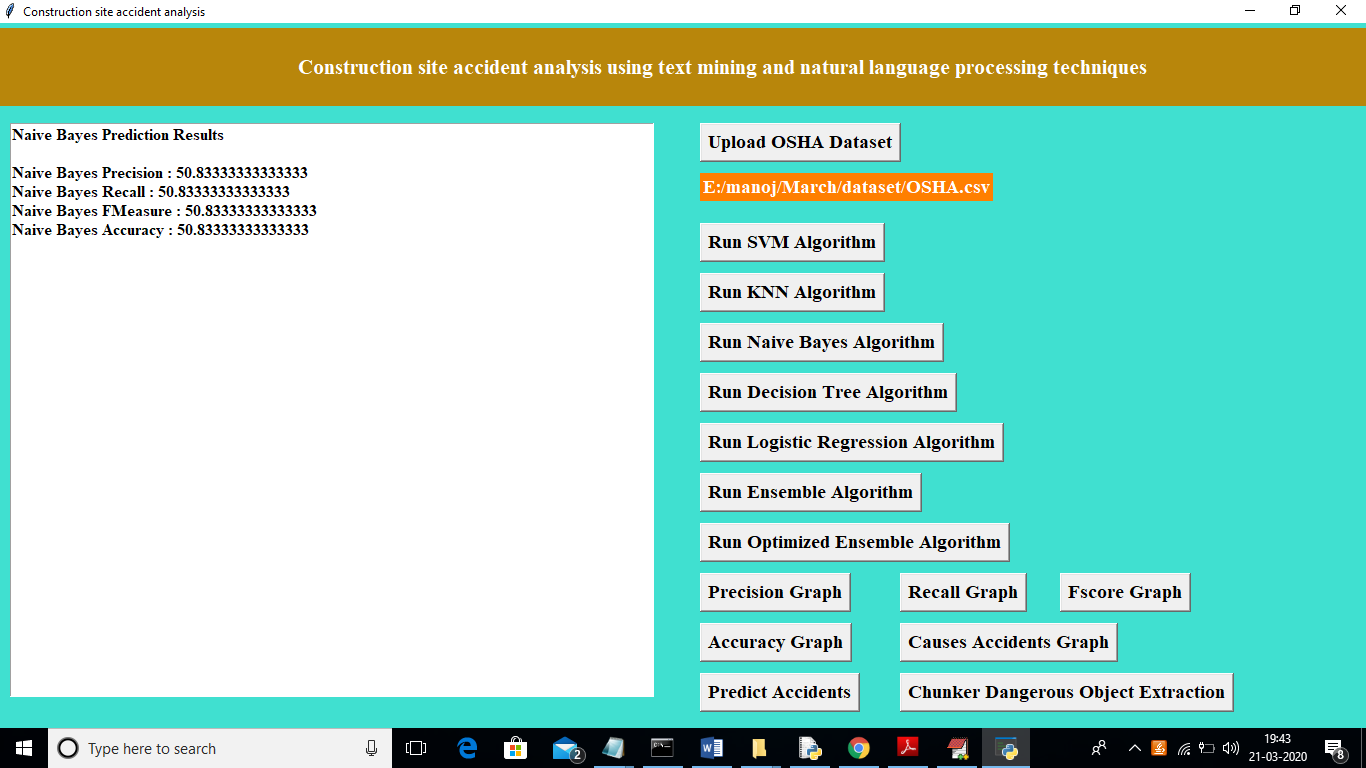
In above screen we can see dataset contains total 599 records and all records contains total 3934 words or features for vector. Now click on ‘Run SVM Algorithm’ button to build SVM model on uploaded dataset and calculate its prediction accuracy, precision etc.



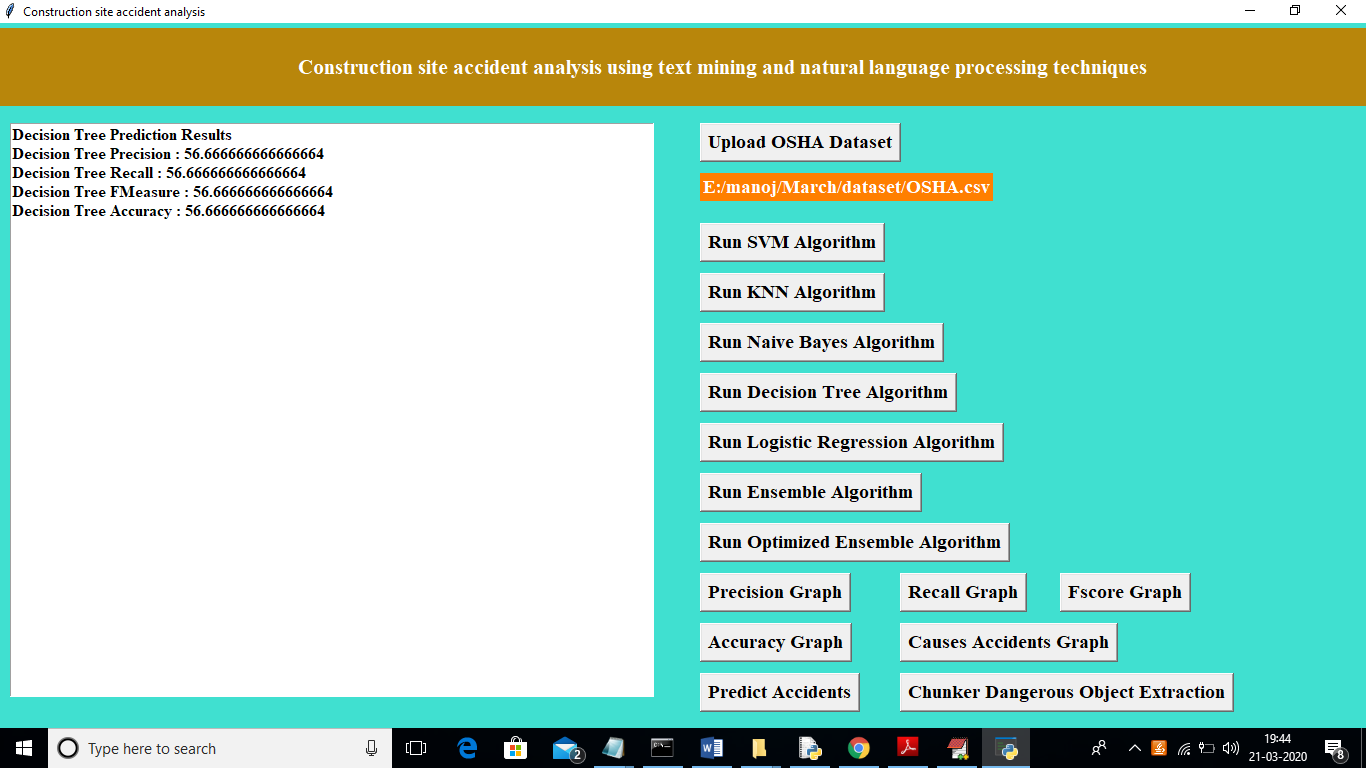
In above screen we got SVM prediction score as 70% and now click on ‘Run KNN Algorithm’ button to get its prediction accuracy



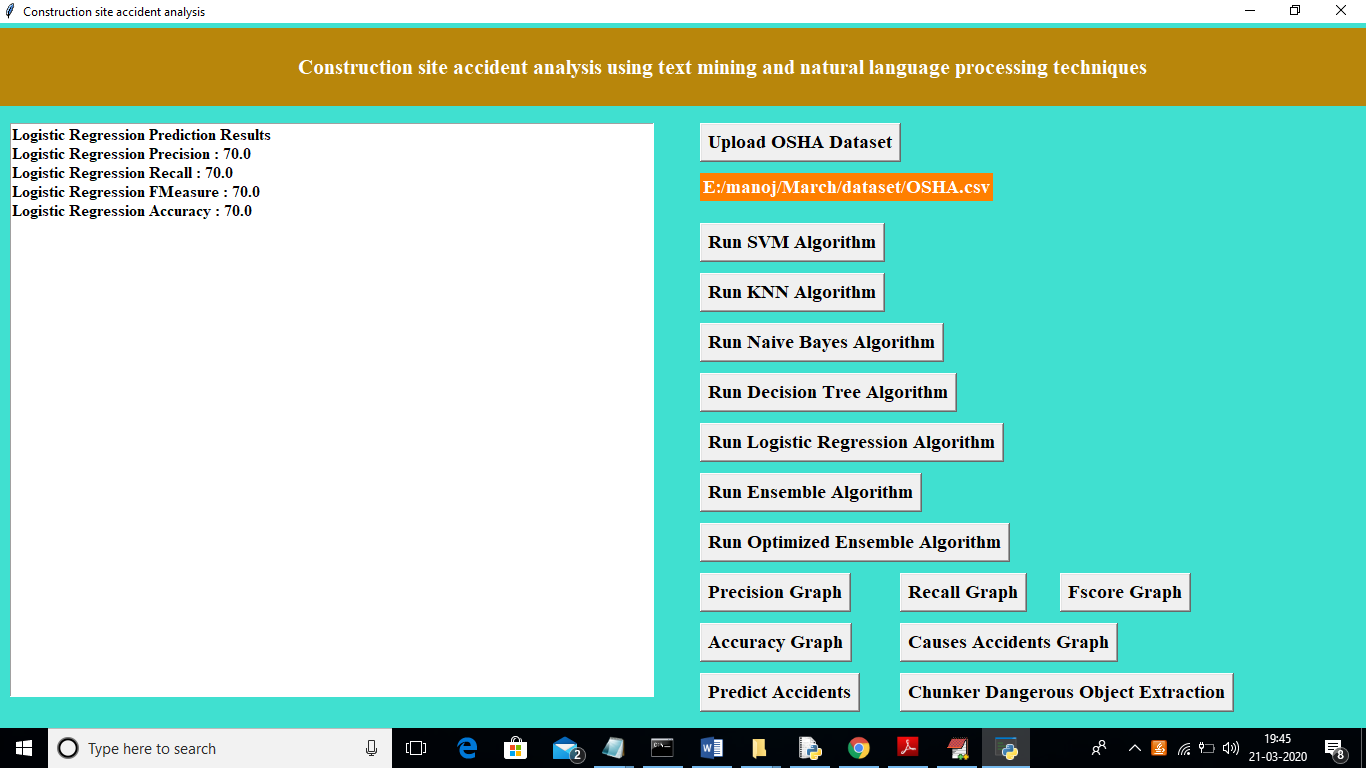
In above screen for KNN we got 55% prediction accuracy and now click on ‘Run Naïve Bayes Algorithm’ button to get its accuracy



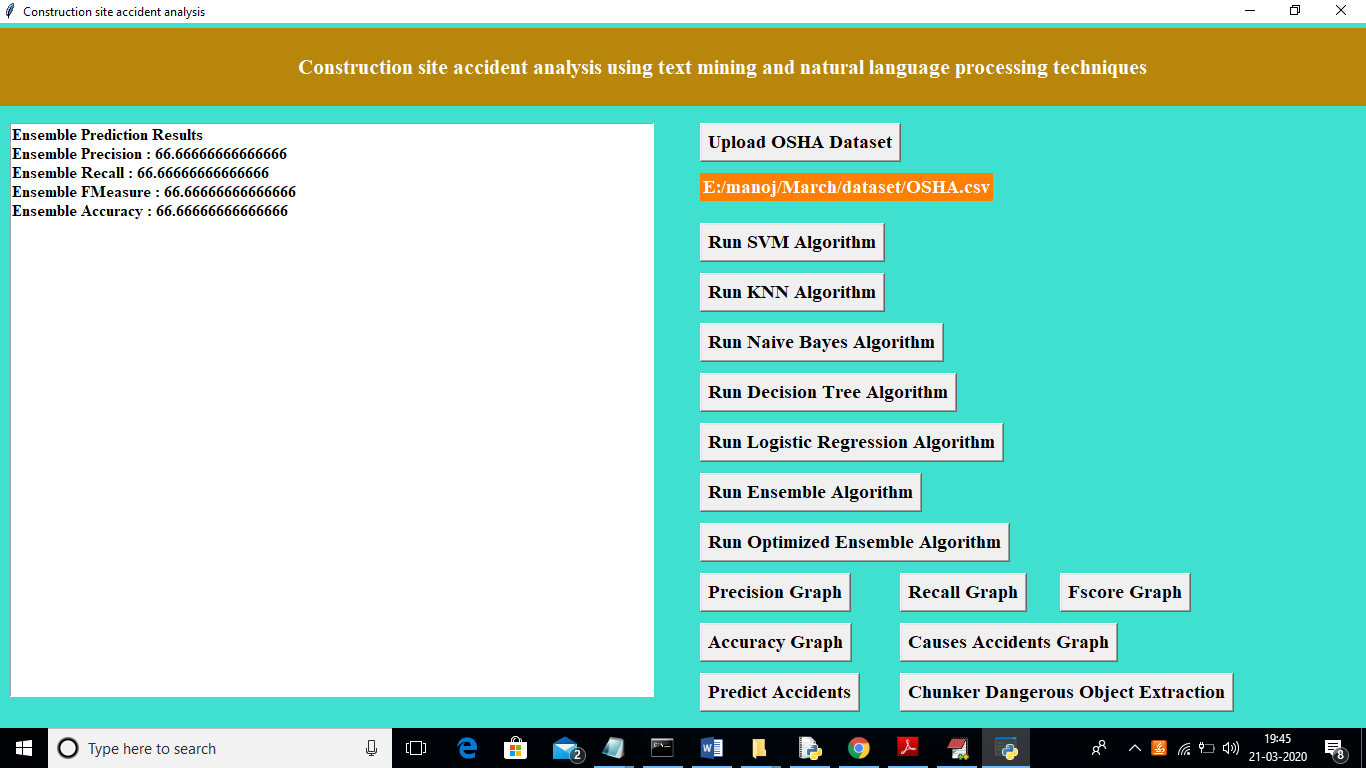
In above screen Naïve Bayes gave 50% accuracy and now click on ‘Run Decision Tree Algorithm’ button to get its accuracy



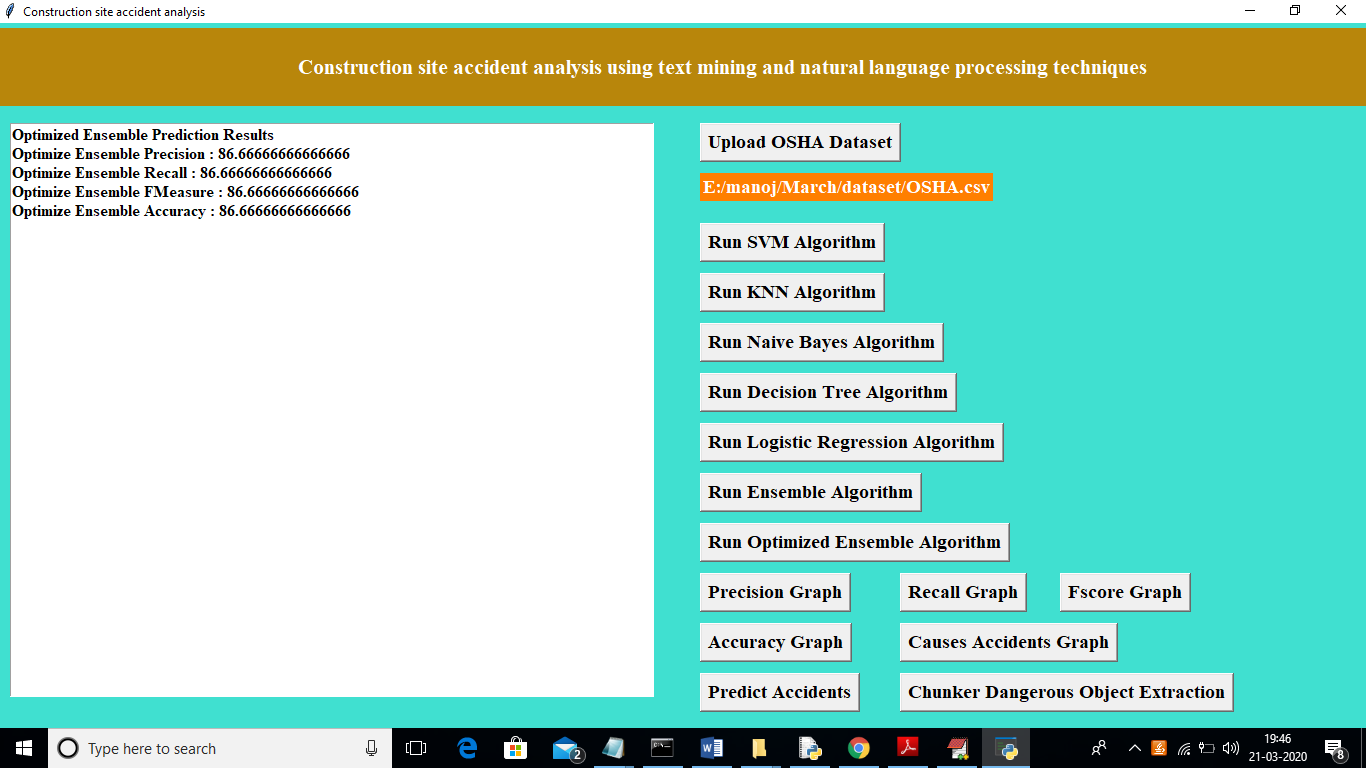
In above screen we got decision tree accuracy as 56% and now click on ‘Run Logistic Regression Algorithm’ button to get it accuracy



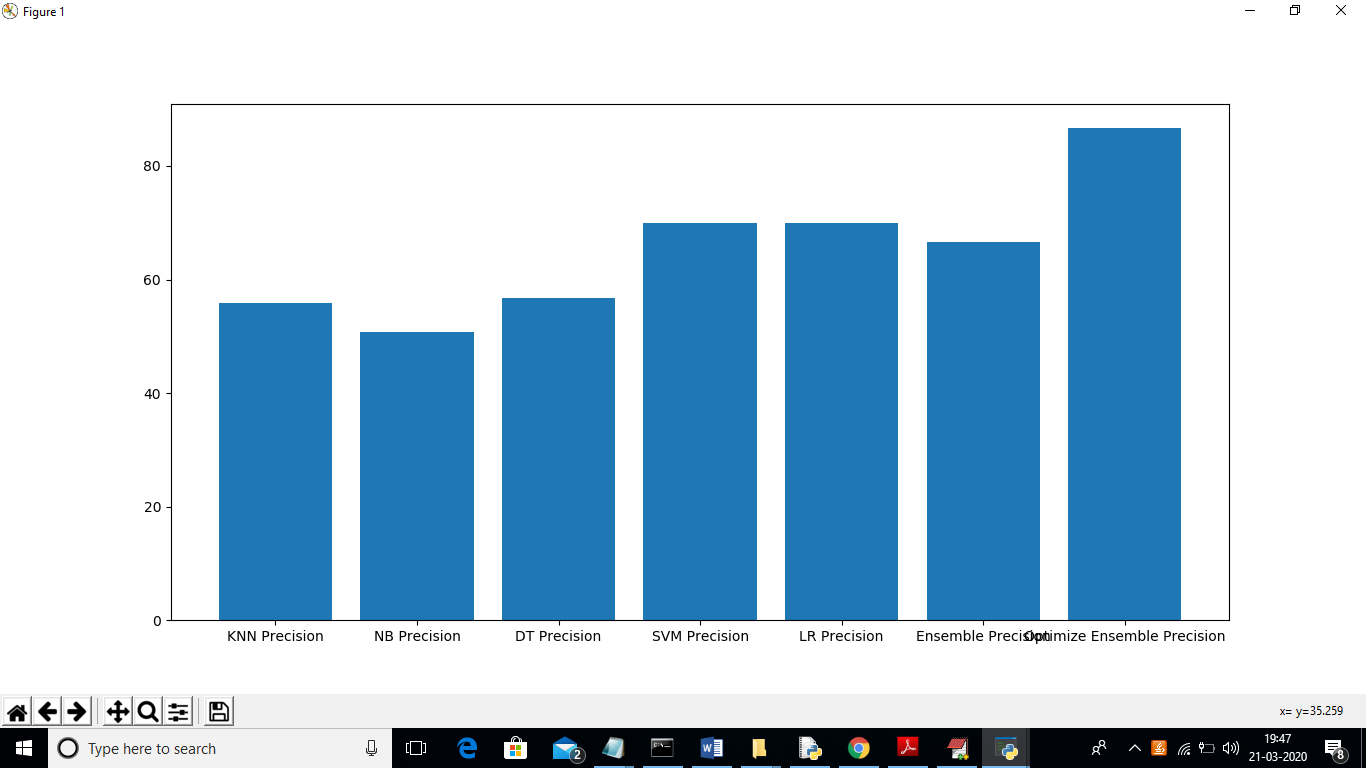
Similarly run ensemble algorithm



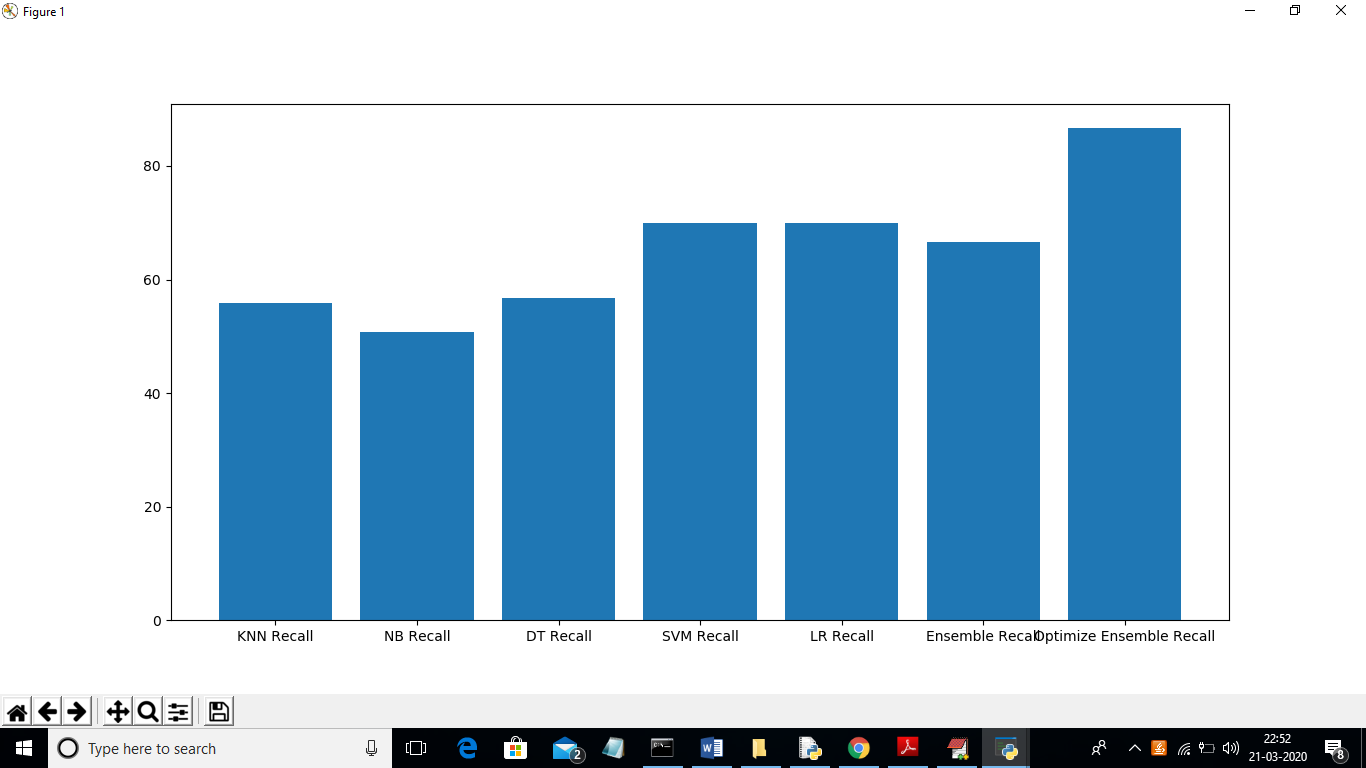
Similarly run ‘Optimized Ensemble Algorithm’ button to get accuracy of voting classifier. Optimize ensemble algorithm is also called as Voting Classifier



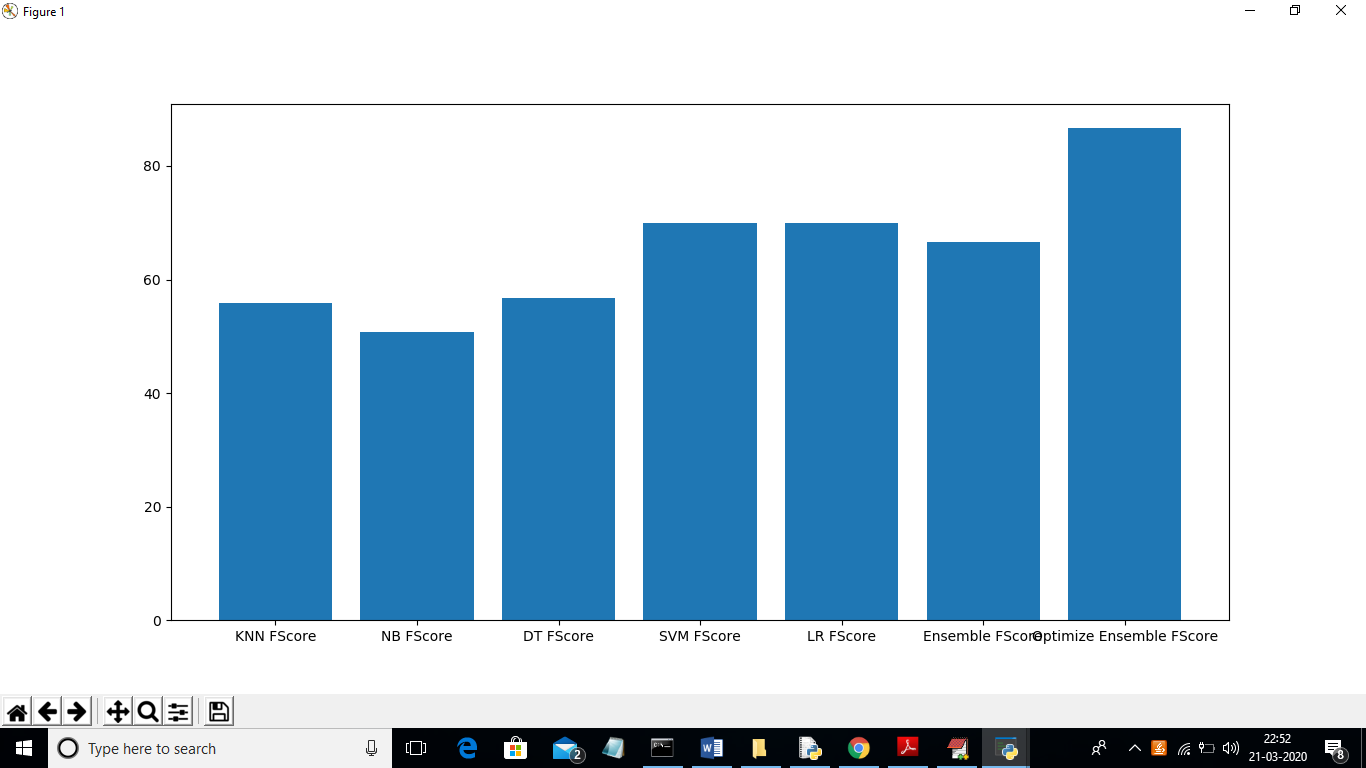
In above screen with Voting Classifier we got 86% accuracy. Now click on ‘Precision Graph’ button to view precision Comparison between all algorithms



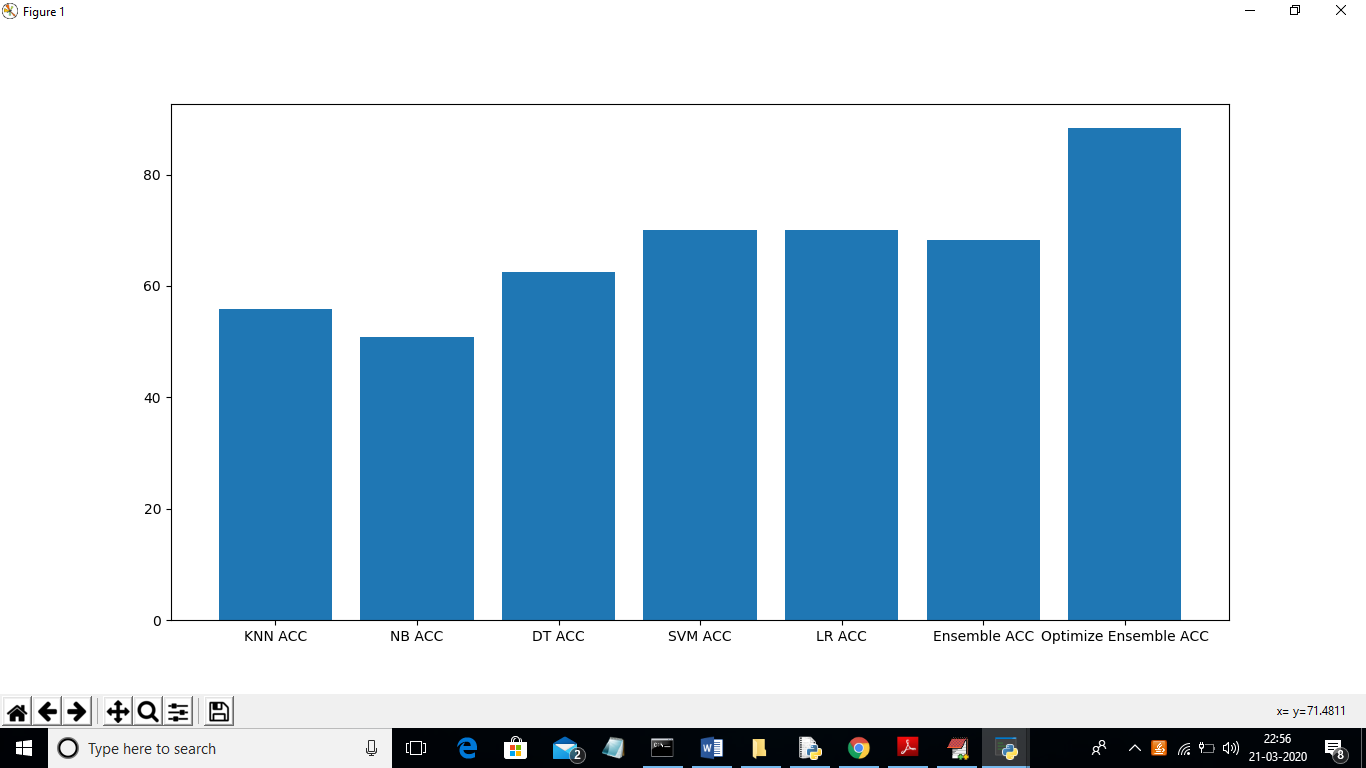
In above graph x-axis represents algorithm names and y-axis represents precision of those algorithms. In above graph we can see Propose Optimize Ensemble (Voting Classifier) gave better performance. Now click on ‘Recall Graph’ button to get below recall comparison graph in all algorithms



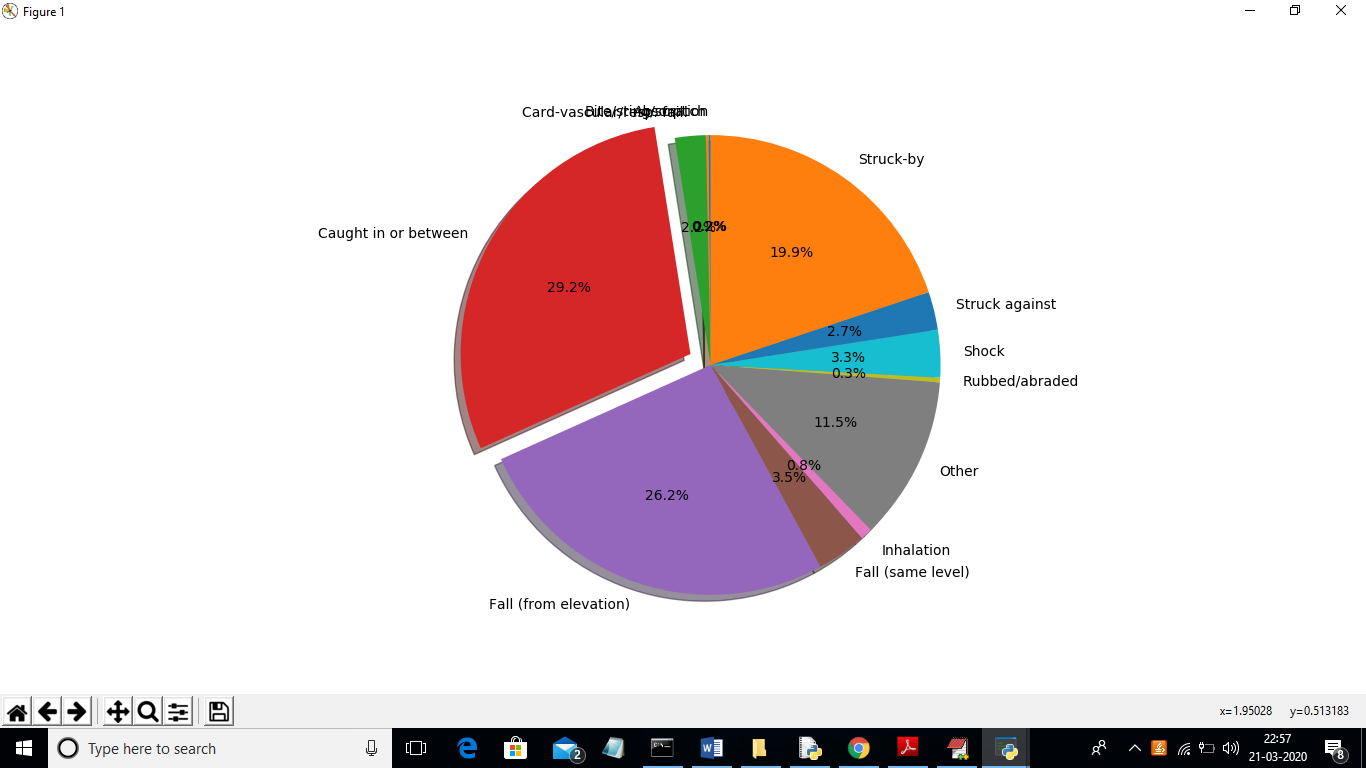
Below fscore graph



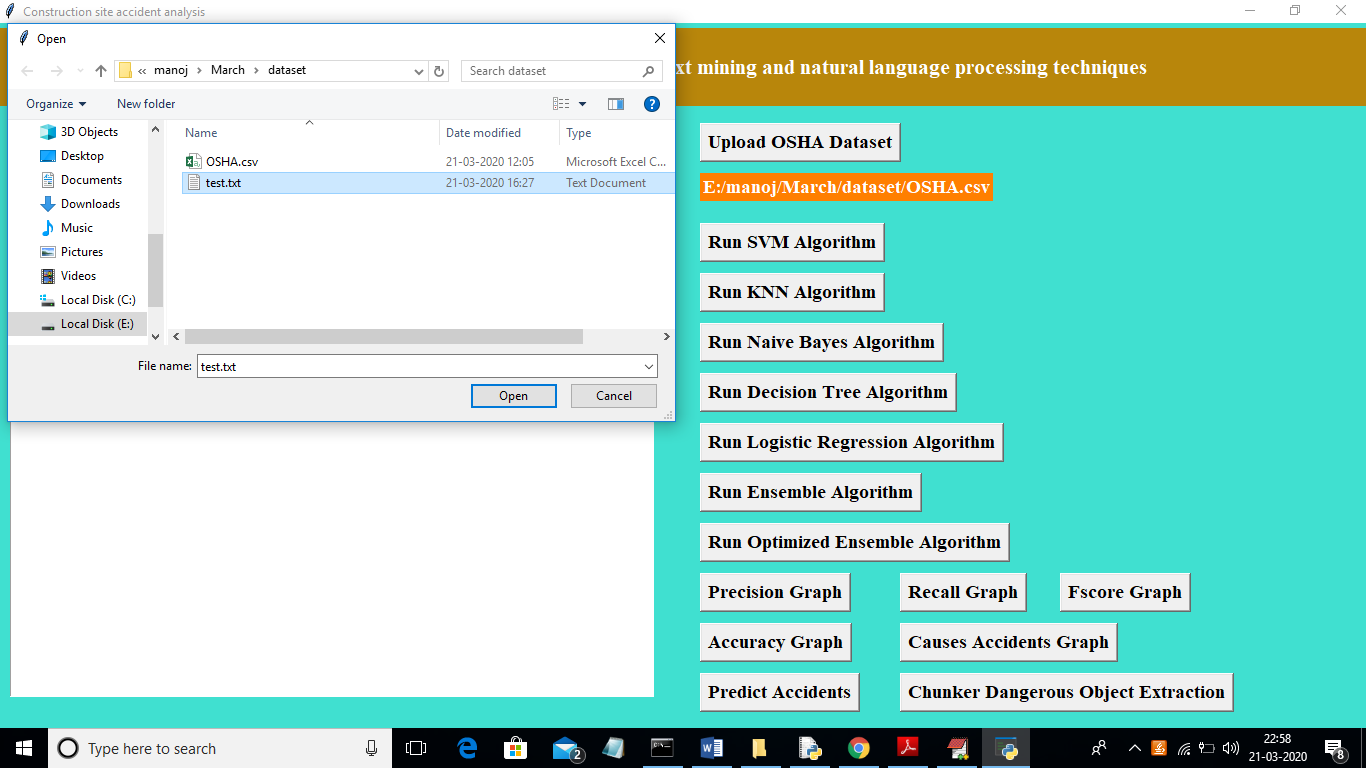
Accuracy graph comparison in all algorithms



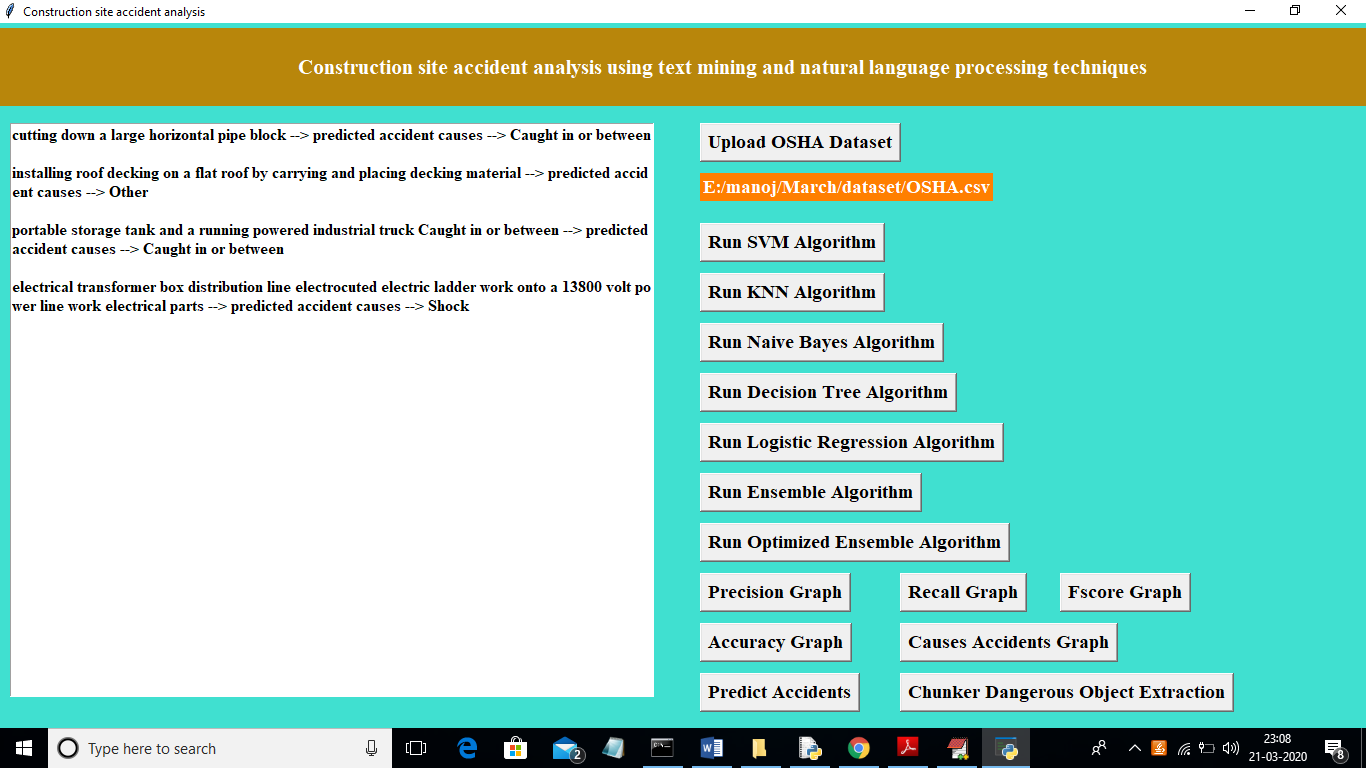
In above accuracy graph x-axis represents algorithm name and y-axis represents accuracy of those algorithms. Now click on ‘Causes Accident Graphs’ button to see various types of accidents causes in this dataset



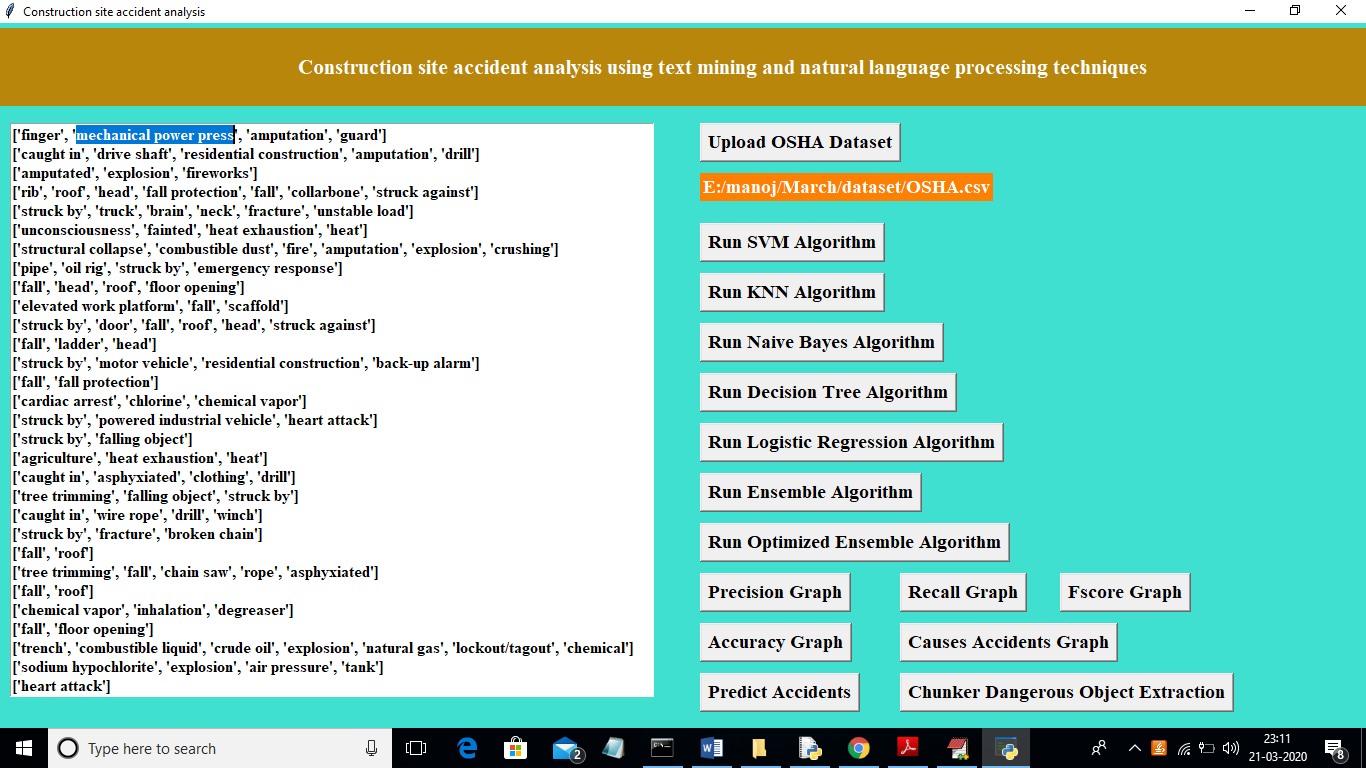
Now click on ‘Predict Accidents’ button to upload new work as test data and predict future accidents can occur in this work



In above screen I am uploading ‘test.txt’ which contains new work and application predict accidents which may occur in this work



In above screen before first arrow mark is the description of work and after second arrow mark is the predicted future accident that may occur while doing this work. Now click on ‘Chunker Dangerous Object Extraction’ button to extract all dangerous object which has to remove while doing work and this objects can be sharp blades or misused tools etc.



In above screen we can see all tools names as noun phrases. Selected text in above screen is one of the example of tool extracted using noun phrases. Scroll down text area to get all tool names