Intelligent Analytics

Study of narcotic criminals using machine learning

INTERNSHIP REPORT

Submitted by

Guransh Kaur Ghai – 1RV18CS059

DEPARTMENT OF COMPUTER SCIENCE
RV College of Engineering

Under the guidance of

Prof. Smitha G R

Assistant Professor
Dept. of ISE
RV College of Engineering®

2020-2021

Acknowledgement

I am indebted to my mentor, Prof. Smitha G R, Assistant Professor, Department of Information Science and Engineering, RVCE, for her wholehearted support, suggestions and invaluable advice throughout the internship.

I would also express my sincere gratitude to Dr. Vijayalakshmi M N, Associate Professor, Department of MCA, RVCE, for her support and encouragement.

ABSTRACT

Study of narcotics using the machine learning algorithm-

The term narcotic refers to any psychoactive compound with sleep-inducing and euphoric properties. Different illegal drugs have different effects on people and these effects are influenced by many factors. This makes them unpredictable and dangerous, especially for young people.

This study concentrates on using machine learning algorithms to asses patterns of illegal drug consumption and sales based on various socio-economic attributes.

The dataset used is the US_narcotics dataset that is obtained from kaggle. The attributes used in the dataset are year ,population, nonUSborn, unemployment ,education_count , commute_count , GDP, income.The data is collected from 2011 to 2017 .The size of our dataset is 6953 KB. Programming language used is python (jupyter notebook).

This study use the SVM (support vector machine) supervised machine learning algorithm to find out the pre-dominant drug involved in crimes in different states and counties of US. This study is classifying the type of drug which is the dependent variable and the others are the independent variables. The 2 two types of drugs we are considering is methadone and heroine. The correlation matrix and SNS heatmap are also used to explain the patterns and dependencies of each to the attributes with respect to the variable this study predicts.

TABLE OF CONTENTS

1. INTRODUCTION 1

2. LITERATURE SURVEY 2

3. DATA DESCRIPTION 4

    3.1. Data Source 4

    3.2. Positive Data Sample Creation 4

    3.3. Negative Data Sampling 5

4. METHODOLOGY 6

    4.1. Model using Machine Learning 6

5. RESULTS AND DISCUSSION 9

6. CONCLUSION 12

INTRODUCTION

Leveraging the power of machine learning is a viable option to detect drug related patterns and set of population which is most vulnerable. It can help notify the authorities where drug consumption has a tendency to peak and help it to take the necessary precautions. This is done imputing the features and observing the patterns.
Machine Learning algorithms are broadly classified into Supervised and Unsupervised Learning Algorithms. Supervised learning algorithms are those which involve learning of a function that maps the input to output, forming input output pairs. They are generally of two types: Classification (use of categorical class variables) and Regression (predict real numbered outputs). This project makes use of the Classification technique. Unsupervised learning algorithms on the other hand are used to draw inferences from input data without labelled responses. They are divided into Clustering and Association. This project does not employ any unsupervised algorithms. Models of this kind can help save as many lives as quickly as possible and turn out to be an efficient and cost-effective solution. It can be further improved by the inclusion of features such as financial conditions.

Support Vector Machine (SVM) is a relatively simple Supervised Machine Learning Algorithm used for classification and/or regression. It is more preferred for classification but is sometimes very useful for regression as well. Basically, SVM finds a hyper-plane that creates a boundary between the types of data. In 2-dimensional space, this hyper-plane is nothing but a line.
In SVM, we plot each data item in the dataset in an N-dimensional space, where N is the number of features/attributes in the data. Next, find the optimal hyperplane to separate the data. So by this, you must have understood that inherently, SVM can only perform binary classification (i.e., choose between two classes). However, there are various techniques to use for multi-class problems.

We are also using the K-Nearest Neighboring Algorithm. KNN algorithm then uses its approach which assumes that similar things exist in close proximity and classifies new cases based on similarity measures. K Nearest Neighbor Classifier is a supervised machine learning algorithm useful for classification problems. It works by finding the distances between a query and all the examples in the data, selecting the specified examples that are closest to the query, and then votes for the most frequent label. It is not parametric which implies that it does not make any supposition on the primary data distribution. To put it in simple words, the model structure is decided by the data. It's pretty useful because in reality, most of the data does not follow the typical theoretical norms made. Hence, we decided to use K-Nearest-Neighbor Algorithm.

LITERATURE SURVEY

Paper Title and complete details: Crime Prediction Using K-Nearest Neighboring Algorithm

IEEE// A. Kumar, A. Verma, G. Shinde, Y. Sukhdeve and N. Lal, "Crime Prediction Using K-Nearest Neighboring Algorithm," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-4, doi: 10.1109/ic-ETITE47903.2020.155.

ML algorithm Used	Reason	Remarks
K Nearest Neighbor Algorithm	It is useful for classification problems and has high accuracy rate. It is not parametric which implies that it does not make any supposition on the primary data distribution and the model structure is decided by the data.	It works by finding the distances between a query and all the examples in the data, selecting the specified examples that are closest to the query, and then votes for the most frequent label. It is a greedy technique and stores all training data.

This research paper offers a way to foresee and predict crimes and frauds within a city. It basically gives us the hotspots of crime. The data is taken considering the time and type of crime that happened in the past.

KNN algorithm then uses its approach which assumes that similar things exist in close proximity and classifies new cases based on similarity measures.

Dataset: Obtained from police website of a city Indore in Madhya Pradesh.

The final dataset now has four attributes with hour, day of the year, longitude and latitude of the city.

It was processed multiple times and they dropped features such as police station, station number, Complainant name & address, accused name & address. Extra Trees Classifier is used to further drop year, month, week of the year from the dataset.

An SNS heatmap is generated to show how crimes vary according to days of a month.

The Elbow Method is very widely used method which helps in determining the optimal value of k. The elbow method runs k-means clustering on the dataset for a variety of values for k (e.g. 1-15) and then for every value of k, it works out an average score for all clusters.

The patterns of crime are not same every time patterns always changes after time to time. The system was trained to learn using some particular inputs. So, the method by itself learns different changes that come in the pattern of crime after analyzing them.

DATA DESCRIPTION

As shown in the below diagram, the process of obtaining and defining the data is mainly composed of three essential steps: Data Sourcing, Positive data sample creation, Negative data sample creation.

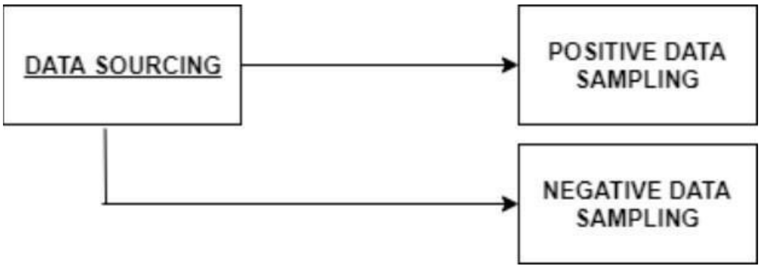


Fig 3.1 Flow diagram of the data description step

3.1 Data Source

The dataset used in the given project is obtained from kaggle. The dataset is open sourced in nature with a .csv format. The size of our dataset is 6953 KB.

The dataset has a total 128790 rows and 9 columns.All of the missing data is accounted for.

The train and test data were split in the following manner.

- Training dataset consists of 80% data.
- Testing dataset consists of 20% data.

The data will be classified into type of drug i.e. Methadone or Heroine where the algorithm takes into account the various attributes of different counties and states of the US to classify.

METHODOLOGY

This covers the technique and flow of events that were used to perform the detection process. The prediction methodology itself is composed of two integral steps: applying the descriptor/algorithm to the data model, the final prediction process.

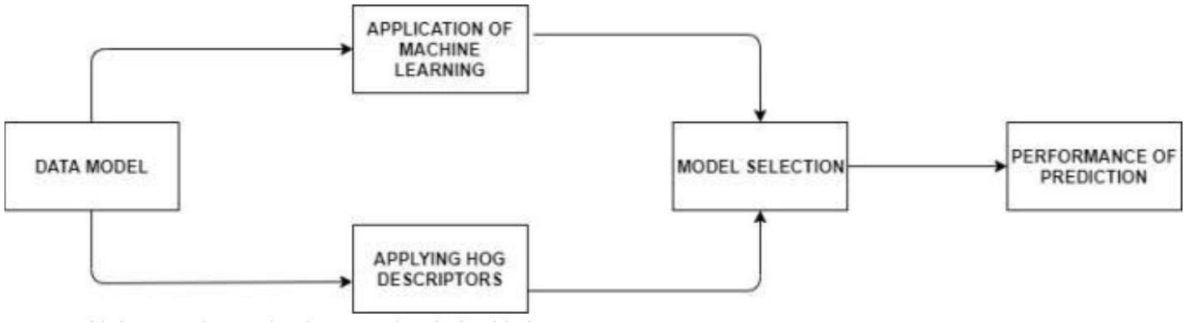


Fig 4.1 Flow diagram of the methodology step

4.1 Model using Machine Learning

For any given prediction/detection problem, there are numerous Machine Learning Algorithms that can be used. Thus, to find out the one most suitable to our purpose, all of them must be evaluated on the same data based on a suitable parameter.

For this given Data Model, we chose the accuracy_score metric that is provided by the sklearn library of python as our evaluation metric[10]. The accuracy score is calculated as follows

$$accuracy(y, y') = \frac{1 * \sum_{0}^{n\_samples-1} 1(y_i = y'_i)}{n\_samples} \quad (1)$$

In the above formula, y refers to the set of predicted values for a given test set, while y’ refers to the actual values of the test set that are being predicted. The number of samples in this case is n_samples. Both y and y’ contain the values -

1,1, that indicate if the predominant drug involved in crimes is Heroine or Methadone.

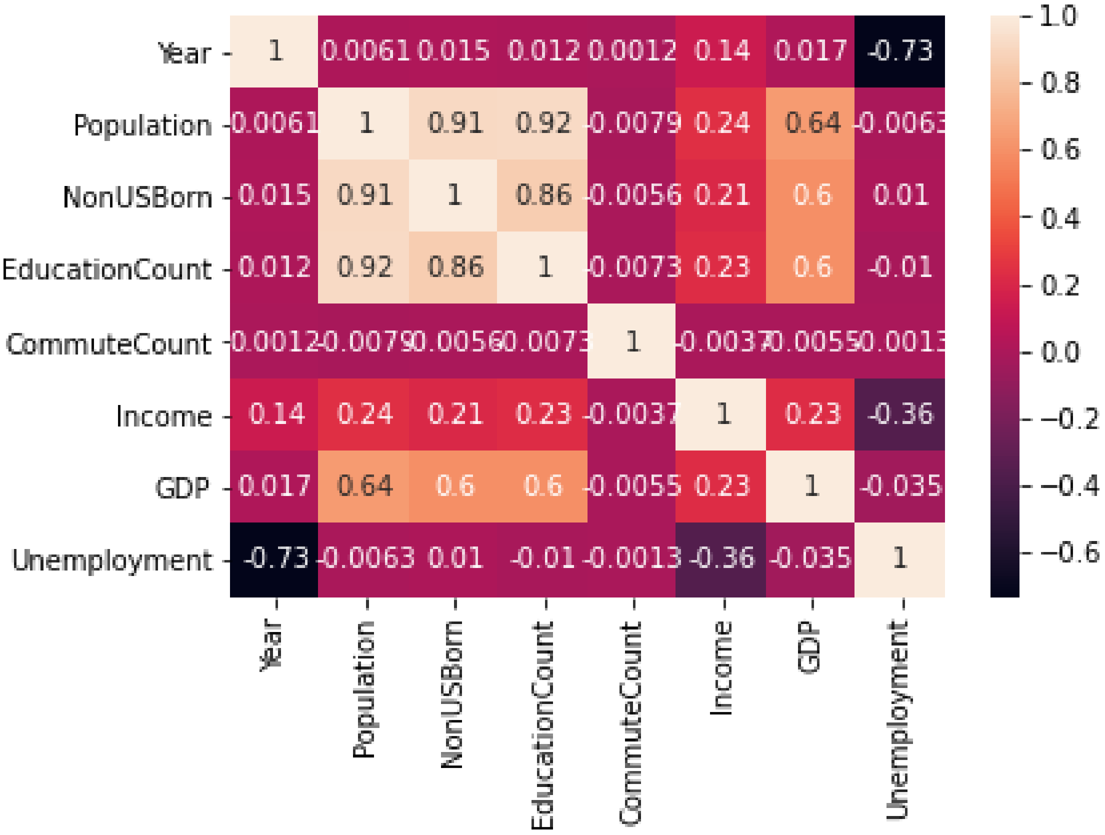
The y_Train and y_Test (The Training and testing set with all parameters) are two dimensional in nature. This was in sync with the fact that most Supervised Learning Algorithms tend to operate on two-dimensional y_Train and y_Test.

The standard X_train and X_test worked with the Support Vector Machine (provided as a part of the OpenCV library in Python). The accuracy of the algorithms was as follows

Algorithm	**Score**
Support Vector Machine	0.85
K-Nearest Neighbors (n=5)	0.98
K-Nearest Neighbors (n=10)	0.97

RESULTS AND DISCUSSION

A heat map is a data visualization technique that shows magnitude of a phenomenon as colour in two dimensions. The variation in colour may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.



Confusion matrix

It is a way of predicting the number of missclasifications.Since our dataset has 2 classes Heroine and Methadone. We have a 2x2 matrix. Each row is instances in the predicted class and each column is instances in the actual class.

TP	FP
FN	TN

1. **TN / True Negative:** when a case was negative and predicted negative
2. **TP / True Positive:** when a case was positive and predicted positive
3. **FN / False Negative:** when a case was positive but predicted negative
4. **FP / False Positive:** when a case was negative but predicted positive

Classification Report

Used to measure quality of prediction from algorithm. How many predictions are true and how many are false.

Precision – Accuracy of positive predictions.
Precision = TP/(TP + FP)

Recall: Fraction of positives that were correctly identified.
Recall = TP/(TP+FN)

F1 score –percent of positive predictions correct
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Support – Number of samples of the response that lie in that class.

```
           [[ 8225   984]
  [ 1202 53984]]
              precision    recall  f1-score   support

      Heroin       0.87      0.89      0.88      9209
   Methadone       0.98      0.98      0.98     55186

    accuracy                           0.97     64395
   macro avg       0.93      0.94      0.93     64395
weighted avg       0.97      0.97      0.97     64395
```

CONCLUSIONS

The model uses supervised machine learning algorithms Support vector machine and K-Nearest Neighbors to predict the predominant drug involved in crimes in US states and counties on basis of our dataset. K-nearest neighbors has higher accuracy in the predictive model. The accuracy was calculated using the sklearn provided accuracy_score function. The heatmap generated showed us the different dependencies and we chose the dependent variable as the type of the drug. Applications of this work can be used to predict the nature of drug crimes depending on attributes and data from the specific region.