Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 7423 - 7428

Research Article

Using Machine Learning Algorithms to Analyze Visualize and Detect Crime Data in Three Dimensional Directions

[1]B.Geetha Kumari, [2]Kandukuru Sneha Reddy

[1]Assistant Professor, Department of Computer Science & Engineering, G.Narayanamma Institute of Technology & Science, Shaikpet, Hyderabad, India.

[2]Assistant Professor, Department of Computer Science & Engineering, G.Narayanamma Institute of Technology & Science, Shaikpet, Hyderabad, India.

ABSTRACT- Crime is a huge problem in our society, and fighting it is crucial. On a daily basis, there are several crimes done on a massive scale. For future reference, a database needs to be kept of all criminal acts and acts of misconduct, as these events are unlikely to be reported later. One of the issues now encountered is keeping up with crime statistics as well as analyzing this data to assist in preventing and solving future crimes. The aim of this project is to gather and analyze datasets containing a multitude of criminal events, and to determine which kinds of criminal activity may occur in the future based on various factors. We will use Machine learning algorithms for crime prediction in Chicago by applying machine learning algorithms to the Chicago criminal data set. Decision Tree, Gaussian Naive Bayes, k-NN, Logistic Regression are all excellent supervised classification tools for this task. With this method, we are attempting to predict crimes by classifying, recognizing patterns, and using effective technologies and tools. We can correlate aspects which can aid our comprehension of future crime patterns by looking at crime data trends that have occurred in the past. Machine learning and visualization approaches are applied to estimating that crimes dispersion over an area in this project. Once the primary records are analyzed and shown based on the need, the procedure shifted to the next stage. A three-dimensional modeling of Cologne's city is constructed using crime statistics. Result visualization in geovirtual environments and real-time situation monitoring in a social protection context result from exploring crimes data analysis in geovirtual environments. Three-dimensional measurements are rapidly becoming the norm in the documenting procedure for crime scenes. By utilizing 3D measuring tools, a richer investigation is provided, revealing the presence of each piece of evidence in relation to the rest of the crime scene.

Keywords: Nature of a criminal, determining criminal reasons, predicting criminal behavior, problem solving, and tree-based modeling.

1. INTRODUCTION

Currently, ongoing criminal cases in India are expanding swiftly as crimes are on the rise. In order to solve a case based on certain data, an investigation and analysis should be done within the company. Deciding these criminal cases is extremely difficult for the investigators because there is so much crime data available in India now. The purpose of this paper is to make a positive change to the decision-making of crime [4]. In learning algorithms, computers make decisions without human assistance. The phrase "computerized or self-driving cars" is an excellent illustration of the increasingly broad application of machine learning in recent years. Predicting specific consequences based on our inputs using machine learning algorithms gives us the ability to come up with a solution to solving incidents of crime in India. Machine learning might be a form of artificial intelligence that looks for patterns by employing data analysis. Using machine learning, a computer can discover patterns in data and generate predictions from it without needing to be programmed directly. Machine learning is most typically split into three categories: supervised, unsupervised, and semi-supervised. The supervised learning technique Machine Learning with Unsupervised Learning Just several inputs are provided to the system, and reinforcement learning is utilised. This article shows that supervised learning algorithms cannot accurately identify criminal types [9]. Comparing the multiple classification and regression models to learn which model performs best and yields a more accurate forecast of the data in the dataset is part of the classification and regression process. Here, we're talking about two inputs: our typed input, and the places where the crime occurs. Both of these are applied using machine learning methods.

Both classification and regression are common forms of predictive models. Filtering is performed in this crime statistics prediction area. Preprocessing is an important prediction approach, and it has been applied in a wide range of areas, including predicting stock trends, the pharmaceutical industry, and more. The major purpose of this paper is to identify certain algorithms that can be utilized to accurately assess and estimate crime data. The goal is to teach the model how to predict future the training samples by checking the accuracy of the test data. Here, Logistic Regression, Decision Tree Classification, and Random Forest Classification are all being employed as the models.

2. METHODOLOGY

A lot of scholars have encountered this difficulty when it comes to criminal cases going unsolved for an extended length of time. Several crime prediction algorithms were proposed. Different models yield different results due to the different data sets and feature options selected regarding the data preprocessing. On the Mississippi crime data set, linear regression and Decision stump model methods were utilized to find crime prediction results of 83%, 88% and 67% respectively. However, because each machine learning method is deployed on various datasets with differing features, predictions vary for all scenarios. The data set that was used was sourced from Kaggle.com and the models were those of logistic regression, K-Nearest Neighbors (KNN), and Decision tree classification. The null values are discarded, and the unknown values are filled in. When utilizing these machine learning algorithms, the accuracies are: For KNN, the percentage of successful classifications is 78.73%, but for decision tress classifier, the percentage of correct classifications is 78.60%. Support Vector Classifier (SVC) has a 31% accurate estimate while Gaussian Naïve Bayes has a 64.6% accurate estimate. Data cleaning and pre-processing is completed before model training in order to attain an accuracy of 78.73%.

Logistic Regression

One of its strategies to nonlinear classifications is regression analysis. This is a classification instead of a predictor because it ends with the letter "regression." Data will be categorized using linear bounds in Regression analysis into multiple categories. When doing a multi - class collection, the approach is to use a one vs remaining methodology, in which a distinct binary classifier is trained for every class. By making this assumption, each class is assessed against all the other classes.

Supervised classification estimates are primarily used in regression analysis model estimation. Logistic regression offers a prediction of whether or not an instance belongs to each of classifications, by calculating the linearization of the possibility of observations.

Decision Tree

Continuous values targeting products can be approximated using the decision tree. A tree hierarchy is used to express the concept of knowledge in this representation. In this method, the trees may be expressed as if-then statements, and they can be better understood by Humans. Inductive inferencing algorithms are frequently used for learning approaches that are particularly suited for inductive inference.

The 2 categories of tree structure are Classifiers Trees and Random Forest. We utilised the Decision Tree Model as our model. This is employed when findings that are considered 'yes' or 'no' fall into discrete or categorized categories. The strategy of divide and rule is utilised in a Decision Tree. The data is then subdivided into smaller regional subsets. Select a root node and segment the data on the basis of the value of the information gain (IG).

We keep performing this procedure until all of the sub-nodes are of that class, i.e. it classifies instances by putting them in a specific order down the tree and only keeps the root node, which results in It will be executed for all of the sub trees at this new node. Labeling is determined randomly with regard to the distribution of labels between branches. Impurity is calculated by adding together the product of the probability p and the chance of misclassifying an item. Information Gain assists in making a selection which characteristic to separate the next time a new step is taken. The quantity of knowledge gained throughout the course of the experiment is commonly calculated utilizing entropy that may be related to calculating the arithmetic average at each phase.

This approach has the drawback of outliers because of the dividing of data into subsets; to prevent this problem, a depth limit needs to be established for the tree. To solve this issue, a random forest could be able to help.

Random Forest

A forest of many decision trees is referred to as a random forest. The resulting system includes all of these decision trees, and trains each one. The final prediction is the arithmetic mean of the expectations of each decision tree. The ensemble comprises of several decision trees.

One of the most essential methods adopted in random forest or bootstrapped aggregation is known as Bootstrap aggregation. It is both simple and strong. The ensemble technique includes many machine learning models; for example, it will produce better quality prediction than a single product.

The more flexible your data is, the better it fits your training data. Since we can see from that the algorithm not only is studying the real results and also any disturbance existing in it, the model must also be learning noise values. It is because random forest classifiers is employed to help prevent the disease. Using the decision tree is delicate. The input constantly alters depending on the change in the input.

This system was built upon the foundation of research that explored diverse resources. Almost majority of the crimes are dependent on location and type of crime. Because Linear Regression, Decision Tree, and Random Forest all have an excellent prediction performance, several techniques are applied in this paper in order to forecast criminal activities. This study is based on the dataset available on data.world.com. There are various forms of criminal offences occurring in India, based on the state and year. This document takes crime types as input and outputs the location of crimes. Data processing that precedes analysis is required, and includes cleaning, feature selection, null value removal, data scaling, and normalization and standardization. Removing missing value before modeling may dramatically impact the accuracy of the model.

Feature extraction is used to ensure that just the important elements are selected that don't compromise the model's accuracy. Logistic Regression, Decision Tree, and Random Forest are each trained on the splits between training and test data. Categorical classification models are employed in this example since the output required is a categorical value. In Python, we use data prediction to work with the data.

3. RESULT AND DISCUSSION

This dataset consists of the number of crimes that occurred in India in the year 2001-2018. The geospatial classification used in this visualization classifies regions of India on the basis of several factors, such as the regions of India as well as the district of each and every country where the crimes are carried. In addition, it lists the type of criminal offences that are being committed, such as kidnapping, rape, robbery, theft, fraud, and various other types of criminal misconduct. [6].In order to eliminate the null values and the useless features or attributes, the first and most important step is to pre-process the data. There are 9,000 data entries that are used in this [8]. Null values are eliminated. In order to employ machine learning methods, you must first convert your string values to float.

When all features and attributes in a specific data set have been identified, then any alterations in accuracy are avoided or a greater level of accuracy is achieved by just picking needed features and attributes. By getting rid of the unneeded features, the model's accuracy is improved.

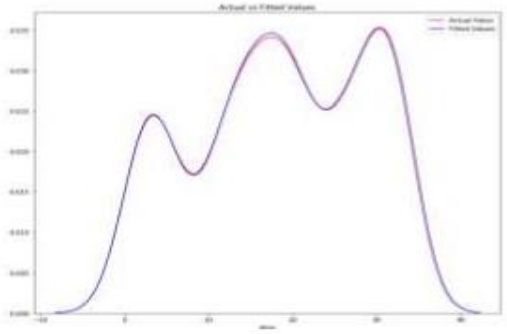
Model	Accuracy (%)
Random forest	95.122
Logistic Regression	78.955

Decision Tree	51.065

It shows the accuracies of the respective algorithms used in this Paper to predict the crime data. Following is the table which summarizes the accuracies of all the models used to predict the results.

By observing the above results it's clear that Random Forest Classifier method has giving the best accuracy among all the methods used and can be chosen or taken as the best model to predict the data for the given dataset with an accuracy of 95.122%.

Sample graph:



The above graph depicts the closeness of actual to the predicted values when we use above three Classification.

4. CONCLUSION

Using certain machine learning techniques, this article shows causes of crime in a specific location. In this work, it is demonstrated that Random Forest Classifier is a good method for predicting outputs, as it is more accurate and is easier to use. Decision Tree categorization has an accuracy of 51.068 percent. It appears that the outputs of the predictions vary depending on the algorithms, and the Random Forest Classification was found to have a near-perfect accuracy of 95.122%.

5. REFERENCES

1. This is done by applying machine learning techniques to crime data. MLAIJ 2.1 (2015): machine learning and apps: an international conference (Journal), volume 2, issue 1, pages 1–12.

2. Dr. Sarvanaguru RA. K, Dr. Alkesh Bharati, "Crime Detection and Analyses Applying Machine Learning" are the two authors of this paper, which is included in the latest issue of the Worldwide Research Article of Science and Engineering (IRJET).

3.Don McClendon and Neetha Meghanathan (2015). Conducting computer analysis using machine learning methods to examine criminal data. The Machine Learning and Applications Journal, volume 2, issue 1, article 1 to 12.

4. Sichun, Hsinchun, et al. "Crime data mining: A method for research & common findings."

5. Our researchers are: Chen, H. and Chung-Dafu, W., and our technicians are: Xu, J. J., Wangsac, G., Qin, Y., and Chauascas, M. (2004). General framework: A crime data mining framework is presented here. Additionally, several examples are provided. a computer, 50-56.

6. As crime data analysis: a broad framework and some examples, Chen, Hsinchun, Wingyan Chung, Jennifer Jie, Xu, Gang Wang, Yi Qin, and Michael Chau.

7. The authors of this study are Chen, H., Chungsasda, W., Xu, J.J., Wang, G., Qin, Y., and Chau, M. a general framework and several examples for crime data mining are found in this article.

8. The best analysis of crime trends and crime prediction are aided by data mining. The First International Conference on Networks and Soft Computing was held in November 2014. (ICNSC2014). The IEEE, in 2014

9. Venkatachalam, V. (2014, August). Data mining for criminal analysis and prediction this year's ICNSC is focused on soft computing and networking technology (pp. 406-412). IEEE.

10. Sathyadevan, Shiju. "Using data mining techniques to study crime trends and forecast." In 2014, the First International Conference on Networks and Soft Computing (ICNSC2014) featured 406-412 in its conference proceedings. 2014.