CRIME PREDICTION USING DECISION TREE (J48) CLASSIFICATION ALGORITHM RESEARCH SUMMARY

I. Publication and researchers

The research paper **"Crime Prediction using Decision Tree (J48) Classification Algorithm "** was published in the **International Journal of Computer and Information Technology , Vol.6 , Issue 03** in the Month of **May 2017.** The researchers were a group of four students from **University of Nairobi , Kenya** and **Kabale University , Uganda.**

II. Introduction

The amount of increase in the occurrence of criminal activities in the recent days is alarming. The society has ceaselessly been a victim to unlawful acts , crime and violence. Crimes are a common social problem controlling people , their way of living and ultimately affecting the economy of the nation. Researchers and Scientists have been dealing with various crime datasets in order to extract knowledge and identify criminal behavioural patterns. The use of Data Mining techniques and machine learning algorithms can contribute to a great extent in the field of criminology. There are various data mining techniques that can be implemented on datasets such as classification , clustering , association etc. WEKA , RapidMiner and R Studio are great platforms for experimenting data mining techniques and machine learning algorithms. The prime aim of this research paper had been to implement decision tree (J48) classification algorithm to predict the likelihood of future crimes based on crime data using WEKA.

III. Dataset

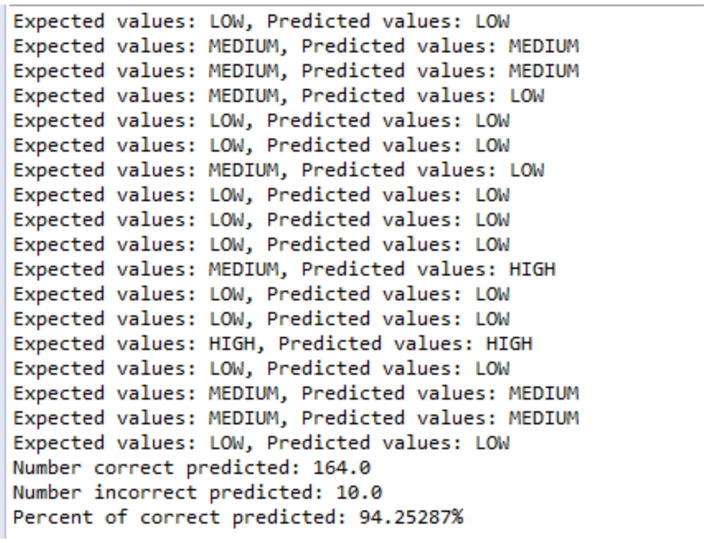
The research students have made use of a crime dataset from the UCI Machine Learning repository website titled **'Crime and Communities'**. This dataset comprises a total number of 128 attributes and 1994 instances. They have stated that the dataset is original and authentic . It was prepared using real socio-economic data from 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR . It is confirmed that all the data elements are numeric and normalised. As the main goal is to predict criminal happenings , they have not taken the whole dataset into consideration. The data was preprocessed and reshaped according to the user's research needs. 12 attributes were selected : State, population ,Median household income , Median family income , Per capita income , Number of people under the poverty level , Percentage of people 25 and over with less than a 9th grade education , Percentage of people 25 and over that are not high school graduates, Percentage of people 25 and over with a bachelor's degree or higher education , Percentage of people 16 and over, in the labor force, and unemployed , Percentage of people 16 and over who are employed , Total number of violent crimes per 100K population. They have introduced a new nominal attribute , Crime Category with three values (low , medium , high) to represent the level of criminal activities in each state.

IV. Technique

Technically , classification in data mining is a supervised learning technique that results in predicting class types for various unknown objects in the dataset based on its similarity to previous examples .The classification process comes under the predictive method. There are a number of classification algorithms such as Support Vector machines , K Nearest Neighbours , Artificial Neural Networks , decision trees , Multilayer Perceptron and Naive Bayes. Evidentially , it is stated that the J48 decision tree performed better on crime data than Naive Bayes , Multi Layer Perceptron and Support Vector Machine methods with a faster and higher performance accuracy .Therefore , J48 decision tree classifier was fixed for this study. For any classification algorithm , the dataset is initially split into two , training set (80%) which is used to create the model and test set (20%) to assess its performance , in such a way to avoid common problems like overfitting and underfitting. Sometimes the model can turn out to be abnormally perfect that it learns the training dataset very well and negatively impacts the future performance of the model , when new data is introduced.

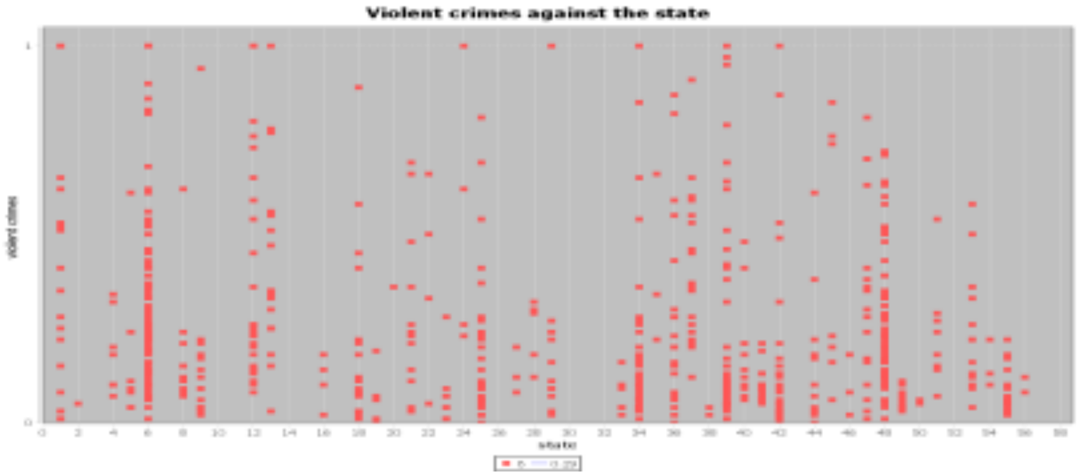
V. Major Findings

After the model had been subjected to initial steps , training and testing the crime data from unknown categories were fed into the model to estimate the prediction of crimes. The result was surprisingly brilliant with an accuracy of 94.25%. In total , 174 different datasets were checked upon , out of which , 164 were predicted correctly.

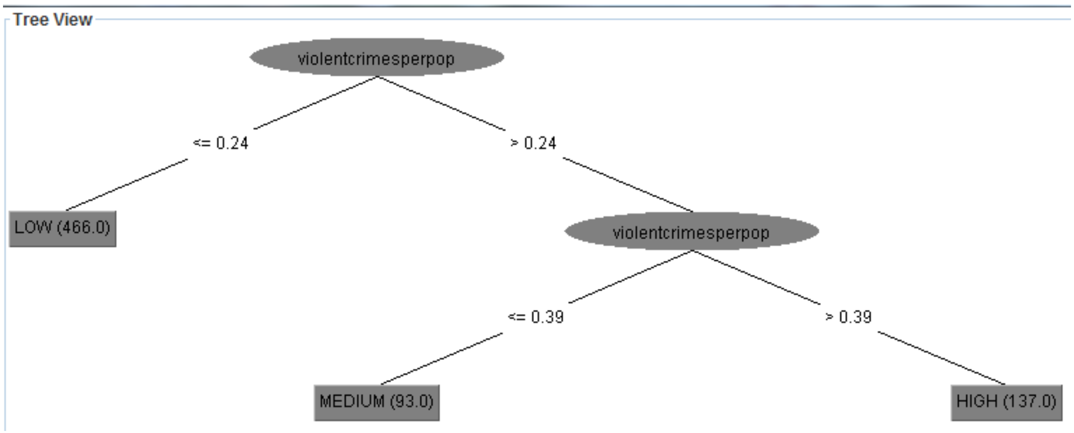


```
Expected values: LOW, Predicted values: LOW
Expected values: MEDIUM, Predicted values: MEDIUM
Expected values: MEDIUM, Predicted values: MEDIUM
Expected values: MEDIUM, Predicted values: LOW
Expected values: LOW, Predicted values: LOW
Expected values: LOW, Predicted values: LOW
Expected values: MEDIUM, Predicted values: LOW
Expected values: LOW, Predicted values: LOW
Expected values: LOW, Predicted values: LOW
Expected values: LOW, Predicted values: LOW
Expected values: MEDIUM, Predicted values: HIGH
Expected values: LOW, Predicted values: LOW
Expected values: LOW, Predicted values: LOW
Expected values: HIGH, Predicted values: HIGH
Expected values: LOW, Predicted values: LOW
Expected values: MEDIUM, Predicted values: MEDIUM
Expected values: MEDIUM, Predicted values: MEDIUM
Expected values: LOW, Predicted values: LOW
Number correct predicted: 164.0
Number incorrect predicted: 10.0
Percent of correct predicted: 94.25287%
```

Violent crimes were taken into account for further analysis. A scatterplot was used to illustrate the distribution of violent category crimes in different states with respect to the dataset. The plot comprehensibly classifies the states into different levels based on the violence in criminal happenings. The premise is simple , more the plots , more the violence in those states and vice versa.



The decision tree , below clearly classifies the states as low , medium , high ; based on the value of Violent crimes per population. The algorithm uses this attribute to group the states in terms of crime, *low if violentcrimes < 0.24 , medium violentcrimes > 0.24 & < 0.39 and high , violentcrimes > 0.39* . When data is visualised as a structure , we can derive at a better understanding and extrapolate more meaningful insights.



VI. Potential Relevance

Considering a number of various classification algorithms and review of literature , the J48 classifier was chosen for this research as it outperformed the other techniques. The WEKA tool was used to handle the crime data after different levels of preprocessing . J48 was found to be successful in the prediction of crimes with a fair accuracy percentage of 94.25 . We found this document highly relevant for our work on the Wine dataset to predict quality of wine based on its points, make , country, price and winery. The J48 classifier model created in WEKA for the wine dataset resulted in an extraordinary accuracy of 99.96% with an absolute error of just 0.0003. The J48 algorithm achieved significantly high accuracy rates on both the datasets used. All the necessary research and background work done , guided us to decide that J48 would possibly be the best suitable classification technique for our dataset.