Is the UK Police Biased?

Abstract—The proposed work used to analyse the crime data and investigate discrepancy made by UK police based on black and other colour ethnicity in the number of arrests. The primary research question that is tackled includes. Correlation based mechanism to determine number of arrests that are made due to colour ethnicity. Next research question is how to distinguish data on the basis of certain attributes. The pre-processing-based mechanism is required while using this dataset for correlation analysis. Pre-processing mechanism is used to handle any noise in the form of missing values. The result accuracy will be increased using pre-processing based mechanism. To determine number of arrests proving discrimination by UK police, correlation-based mechanism is used. Result section indicates that positive correlation is obtained between the number of arrests and people other than white UK people. This positive correlation mechanism indicates that number of arrests increases with increase in black or other coloured peoples. Visualization tools such as matplotlib is used for simplifying the result description.

Index Terms—Pre-processing, UK police discrepancy, Correlation, Visualization tool

F

1 INTRODUCTION

THE primary purpose of this work is to identify the number of persons other than white arrested by UK police to determine discrepancy of the police. [1]The motivation for the same is to gather better knowledge of data analysis through machine learning. Correlation analysis is conducted to authenticate the results. [2], [3][4]The overall process of identification is portioned into phases. The first phase in the identification includes pre-processing.

1.1 Pre-processing

Pre-processing mechanism eliminates the noise if any from the dataset. the noise present within the dataset is in the form of missing values. In addition, numeric field within the dataset “Number of arrests” contains string values represented with “- “. This value causes a problem in the identification of correlation values. To overcome the issue, “-“is replaced with “0”. Also making number of instances within each column of the dataset equal is accomplished using fill method. This is required for achieving better results.

1.2 Categorization

In the next phase, categorization based on different attributes takes places. At first, categorization is based upon Population by ethnicity. The information is grouped and then printed. In the next categorization, geography attribute serves as primary attribute and in the third categorization, time attribute serves as key. All this categorization has the purpose of identifying number of arrests made during time, geography, and ethnicity to identify any discriminatory made by UK police.

1.3 Distinguishment

To identify the discrepancy, number of White people arrested at different time and geography are distinguished from the people who are arrested with another colour. The discrepancy is found and plotted with visualization tool of matplotlib.

1.4 Correlation

In the last phase, correlation is used to authenticate the result. High positive correlation is obtained indicating the people other than white peoples are arrested heavily by UK police and hence proves discrepancy.

The organization of this paper is as under section 2 gives the literature survey describing the work done in analyse data for crime prediction. Section 3 describes the methodology, section 4 gives the results , section 5 gives the discussion of achieved results and section 6 gives the conclusion.

2 Literature Survey

This section gives described the machine learning mechanism used for the prediction of crime. The machine learning mechanism discussed in [5] for the detection of phishing attack. To detect the attack support vector machine was used. High classification accuracy of 95.66 percentage. The dataset for performing the operation was derived from UCI machine learning website. Social media platform also presents threat as discussed in [6]. Tweeter tweets serve as key element for data analysis in this case. Multimodal based approach including support vector machine, k nearest approach and support vector machine was employed. The result of this approach was presented in the form of classification accuracy that was better as compared to single modal based approach. Detection of crime and predicting it is critical but reducing it could be difficult. To this end, mechanism proposed though [7] provides approach to reduce crime rates within India. Clustering mechanisms, optimization and statistical mechanisms were discussed through this approach. The result of the approach was presented through classification accuracy. different machine learning algorithms and data mining approaches discussed in [8] for crime prediction. Learning approaches including supervised and unsupervised both were discussed, and supervised learning approach produced better result according to this literature. The result of crime prediction was presented in the form of classification accuracy. Crime analysis through past 15 years data of Vancouver was analysed in [9]. Machine learning based framework was used for the detection process. Machine learning based mechanism utilised pre-processing, segmentation, and classification phase. The result of this approach was expressed in the form of classification accuracy. The classification accuracy was increased from 39 to 44 percentage proving worth of study. Model for the prediction of DDOS was prepared in [10] using pre-processed convolution neural network. Entire model was portioned into layers. First layer in the operation includes input layer. This layer receives the raw data from the dataset. pre-processing mechanism eliminate the noise if any from the data. After this process, refined data passed to the processing layer. The processing layer acquire the training from the data and last layer known as classification layer receive the test data for result prediction. Entire process validity was expressed in the form of classification accuracy.

The mechanisms discussed within this work deals with the machine learning mechanisms to detect the crime at early stage. Next section discussed the methodology to be followed within the proposed system.

3 Methodology

The methodology of work is portioned into different phases. The phase of operation is given within figure 1. The methodology of work demonstrates the process of achieving the answers to research questions as described below. RQ1 : How correlation based mechanism can be used to determine number of arrests that are made due to colour ethnicity by UK police? To answer this question first pre-processing to eliminate null values and make dataset symmetrical. The categorization is then performed to determine the number of arrests made by UK police. The categorization phase based on arrests of White and other coloured people suggests the discrepancy of UK police. To authenticate the results correlation analysis is conducted. Correlation mechanism shows high positive correlation and hence first research question is answered through this methodology. RQ2: to answer RQ2, the categorization is accomplished with the use of aggregate functions. Group by function is used to group the similar identities and to present the specific result to the user. The result of both the approaches including RQ1 and RQ2 is given within result section.

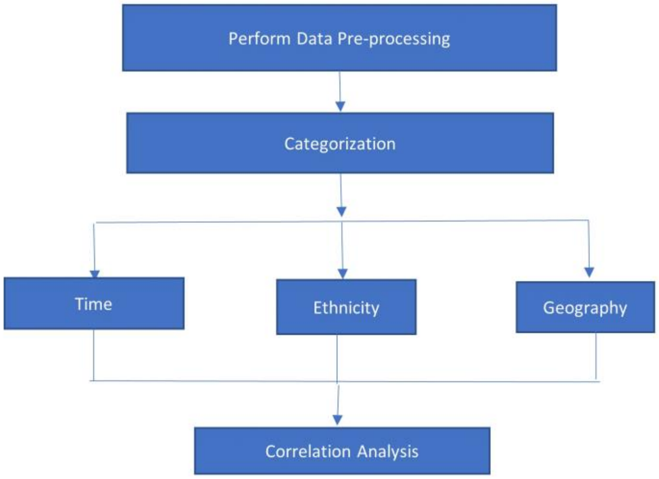


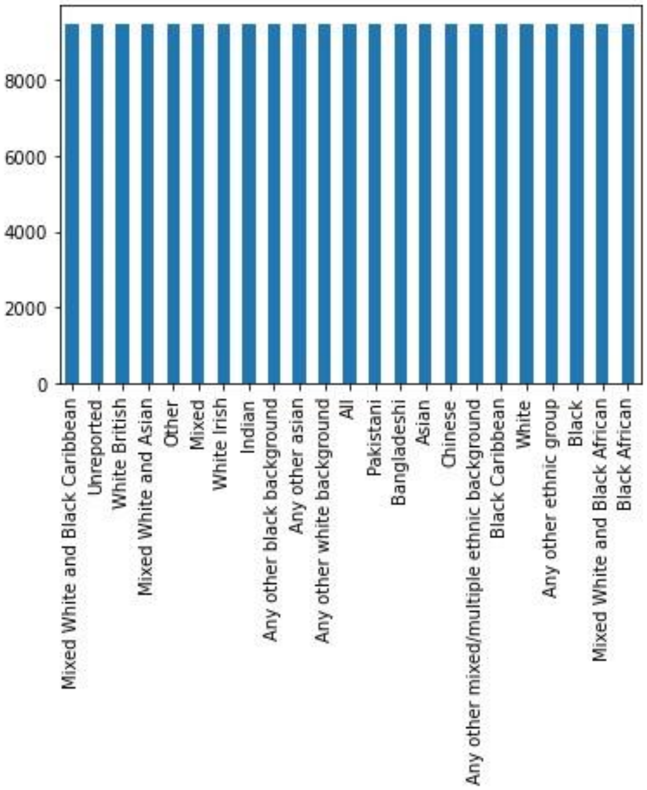
Fig. 1. Methodology of research

4 Results

Data for operation is required and fetched from benchmark website data.gov.uk. the link to the dataset is specified within appendix. At first, pre-processing takes place. Pre-processing mechanism first drop all the null values from the attribute ‘Notes’ . In addition, numeric field ‘Number of arrests’ containing ‘-‘ is replaced with ‘0’ values. The statement used for the same is given as under $df[{}^0Notes^0].fillna({}^0N/A^0,inplace \quad = \quad True)$ (1) $df[{}^0Numberofarrests^0].str.replace({}^0-{}^0,{}^0\,0^0)$ (2)

here ’df’ indicates the data frame in which csv file is being extracted. Result obtained after performing pre- processing of ‘Notes’ attribute is given as under 0 N/A 1 N/A 2 N/A 3 N/A 4 N/A Name: Notes, dtype: object Result obtained after performing pre-processing of ‘Number of arrests’ attribute is given as under 0 10 1 25 2 9 3 0 4 4

Name: Number of arrests , dtype: object 3rd row in Number of arrests attribute contains ‘-‘ which is replaced with ‘0’ value. The plots corresponding to



categorization by ethnicity is given in the following figure 2 The categorization made through geographical location is plotted and result is given within figure 3 the next categorization is on the basis of time. The year-wise categorization of the data is given within figure 4. To prove the discrimina-

Fig. 2. Categorization by ethnicity

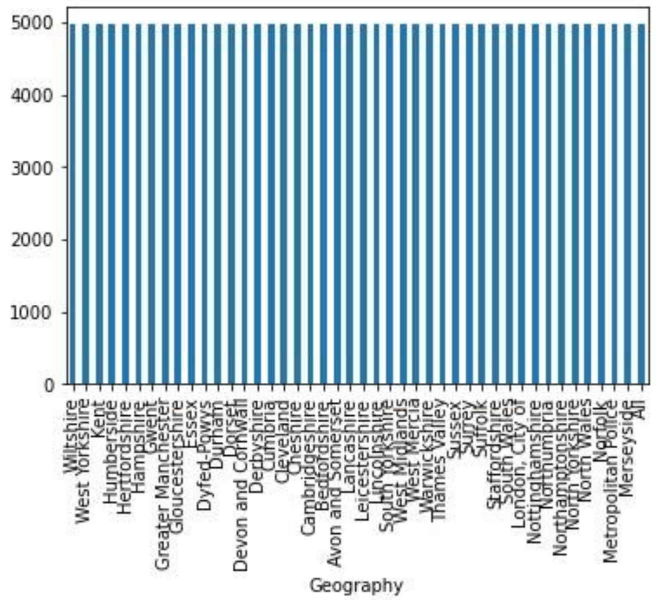


Fig. 3. Categorization by Geography

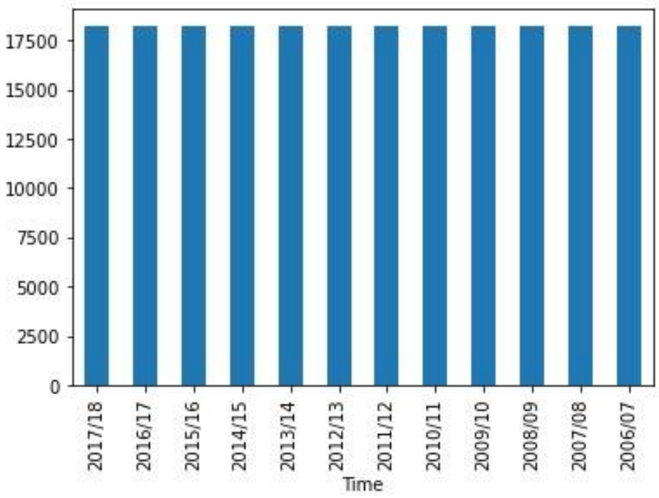


Fig. 4. Categorization by ethnicity

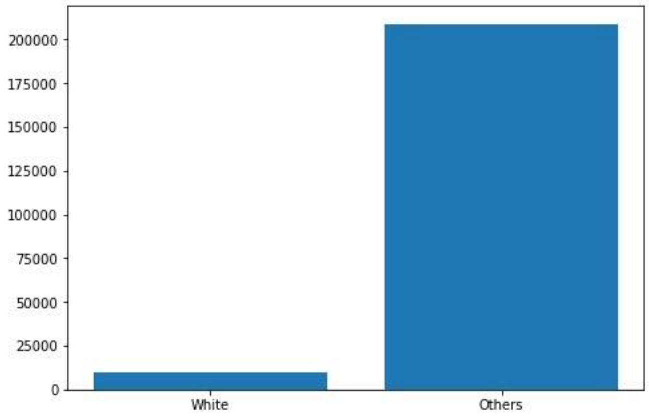


Fig. 5. Categorization by ethnicity

tion performed by UK police on the basis of ethnicity, a group by instruction is used on the basis of number of arrests of White and other colored people. Figure 5 shows this result The correlation is the last phase that is performed to check the validity of the approach. correlation value 1 is obtained showing high positive correlation indicating number of arrests made by UK police is directly proportional to colour of people. The equation used for corrleation evaluation is given as under

$$Ethnicity_{c}ount1.corr(method =^{0} kendall^{0}) \qquad (3)$$

’Ethnicity-count1’ variable contains number of arrests associated with white and non white people. ’kendall’ method is used for correlation calculation since it will evaluate column-wise correlation.

5 DISCUSSION

The entire process of crime data analysis conducted through this approach validated the discrimination made by UK police in terms of ethnicity. Research question 1 indicating how to validate discrimination is answered through correlation. Research question 2 is answered through the categorization process. Within categorization result is obtained using geography, time, and ethnicity grouping. Number of arrests of white people through the time period are much less as compared to other coloured people. To authenticate the result, correlation is obtained. High positive correlation is obtained between number of arrests and ethnicity attribute. Correlation based mechanism indicates the validity of the discrimination made by UK police against non-white people.

6 Conclusion

The entire process of crime data analysis begins with the data acquisition. Data is gathered from data.gov.uk. the link is provided within appendix. The mechanism of discrimination detection is divided into phases. In the first phase, pre-processing is performed. This phase eliminates any noise from the dataset. the noise can in the form of missing values or special characters. The rows that are completely empty are rejected from the data frame. After this phase, categorization is performed. Categorization is based upon three different parameters including time, geography, and ethnicity. After categorization, correlation is performed to check validity of discrimination.

Appendix Dataset

Link :https://data.gov.uk/dataset/f92e60cdea9d-4561-b8df-ba979bda82eb/arrests-byethnicity

References

[1] [1] T. Siddiqui, A. Y. A. Amer, and N. A. Khan, “Criminal Activity Detection in Social Network by Text Mining: Comprehensive Analysis,” in 2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019, Nov. 2019, pp. 224–229, doi: 10.1109/ISCON47742.2019.9036157.

[2] S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma, and N. Yadav, “Crime pattern detection, analysis prediction,” in Proceedings of the International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017, 2017, vol. 2017-January, pp. 225–230, doi: 10.1109/ICECA.2017.8203676.

[3] A. Mary Shermila, A. B. Bellarmine, and N. Santiago, “Crime Data Analysis and Prediction of Perpetrator Identity Using Machine Learning Approach,” in Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, ICOEI 2018, Nov. 2018, pp. 107–114, doi: 10.1109/ICOEI.2018.8553904.

[4] B. Panja, P. Meharia, and K. Mannem, “Crime Analysis Mapping, Intrusion Detection-Using Data Mining,” Jun. 2020, doi: 10.1109/TEMSCON47658.2020.9140074.

[5] J. Rashid, T. Mahmood, M. W. Nisar, and T. Nazir, “Phishing Detection Using Machine Learning Technique,” in Proceedings - 2020 1st International Conference of Smart Systems and Emerging Technologies, SMARTTECH 2020, Nov. 2020, pp. 43–46, doi: 10.1109/SMARTTECH49988.2020.00026.

[6] Z. Abbass, Z. Ali, M. Ali, B. Akbar, and A. Saleem, “A framework to predict social crime through twitter tweets by using machine learning,” in Proceedings - 14th IEEE International Conference on Semantic Computing, ICSC 2020, Feb. 2020, pp. 363–368, doi: 10.1109/ICSC.2020.00073.

[7] S. R. Bandekar and C. Vijayalakshmi, “Design and analysis of machine learning algorithms for the reduction of crime rates in India,” in Procedia Computer Science, Jan. 2020, vol. 172, pp. 122–127, doi: 10.1016/j.procs.2020.05.018.

[8] S. Mahmud, M. Nuha, and A. Sattar, “Crime Rate Prediction Using Machine Learning and Data Mining,” in Advances in Intelligent Systems and Computing, 2021, vol. 1248, pp. 59–69, doi: 10.1007/978-981-15-7394-1-5.

[9] S. Kim, P. Joshi, P. S. Kalsi, and P. Taheri, “Crime Analysis Through Machine Learning,” in 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2018, Jan. 2019, pp. 415–420, doi: 10.1109/IEMCON.2018.8614828.

[10] M. Ghanbari, W. Kinsner, and K. Ferens, “Detecting a distributed denial of service attack using a pre-processed convolutional neural network,” in 2017 IEEE Electrical Power and Energy Conference, EPEC 2017, Feb. 2018, vol. 2017-October, pp. 1–6, doi: 10.1109/EPEC.2017.8286243.