

# Research on Victim Genders in LA based-on Machine Learning

Yuan Feng

Department of Mathematics, University College London, London, UK

yuan.feng.21@ucl.ac.uk

**Abstract.** Recent years, more and more victims in crime in Los Angeles, especially after COVID-19. And women have more fear of crime and feel that they are more vulnerable than men. This article studies the future victim gender prediction by using algorithms in machine learning, like KNN, logistic regression and neural network. Machine learning has become more and more popular in the past few years. The results show that females are not seem as more vulnerable in crime than males in LA region. In addition, neural network method performs better than KNN and logistic regression, showing higher accuracy and f1-score. The accuracy of KNN is 59.82% and the f1-score is 65.36%. The accuracy of LR is 61.31% and f1-score is 67.76%. The accuracy of neural network is 62.58%, the f1-score is 70.22%. The result in this article can help the government to mitigate the heavy anxiety of females about being a victim in crimes.

**Keywords:** Machine Learning, Crime Victim, Gender, KNN, Logistic Regression, Neural Network.

## 1. Introduction

After over three years since the start of COVID-19 pandemic, in 2022, the aggregate crime rate in Los Angeles witnessed an 11% surge. Last year, there were 60 reported crimes per 1000 residents, whereas in 2019, the figure was 54 per 1000 residents [1]. In Los Angeles, women expressed a higher tendency to believe that crime was escalating and felt slightly less secure compared to men. However, conversely, men seemed to experience criminal incidents more frequently than women [2].

Machine learning has become more and more popular in the past few years and using machine learning to predict crime captivate high volumes of considerations [3]. Many studies have utilized machine learning techniques for prediction purposes. One such example is the work of Paidipati Dinesh and Kalyanasundaram.P, who employed KNN and logistic regression algorithms to predict disease diagnosis. Their models achieved an accuracy of 90% and 95% respectively [4]. Mohammed Marouane Saim and Hassan Ammor also used machine learning techniques, KNN and logistic regression, predicting potential cardiovascular diseases with f1-scores are 82.25% and 62.27% respectively [5].

However, the majority of research has focused on analyzing victim patterns in recent years rather than making predictions for the future, so that would be the main purpose of this written article. The primary objective is to discuss which genders are more vulnerable in crimes in Los Angeles. This article can be split into five sections, the background and the purpose in the first section while the second and the third section are descriptions of data and methods (KNN, logistic regression and neural network) used for prediction.

## 2. Data Methods

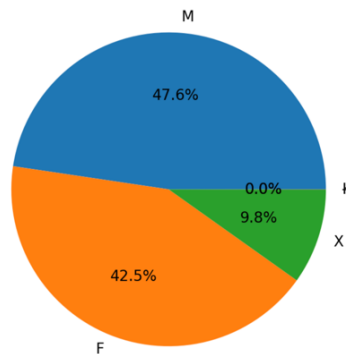
### 2.1. Data

The parameters applied in this research and their definitions are shown in table 1.

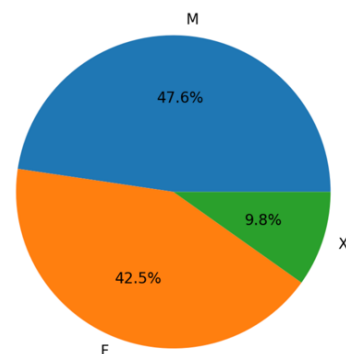
Sex data from the dataset from Kaggle which has licensed from U.S. Government Works had been summarized, which has 313598 males, 279785 females, 64859 unknowns, 87 'H' and 1 '-', visualizing them by using a pie chart to indicate their proportions clearly. From Fig.1, the most apparent parts are 'F', 'M' and 'X' (42.5%, 47.6% and 9.8% respectively), so it seems sound to drop 'H' and '-' data since they only accounts for tiny parts. Therefore, Fig.2 shows the distribution of victim gender after filtering.

**Table 1.** Parameter Description

| Variables   | Definitions                                                                    |
|-------------|--------------------------------------------------------------------------------|
| Vict Sex    | Gender of each victim, 'M' means male, 'F' means female and 'X' means unknown. |
| Crm Cd      | Code of each crime type committed.                                             |
| Crm Cd Desc | Description of different crime type with corresponding code (Crm Cd).          |
| DATE OCC    | Crime occurring date in MM/DD/YYYY form.                                       |
| TIME OCC    | Crime occurring time in 24-hour military time                                  |

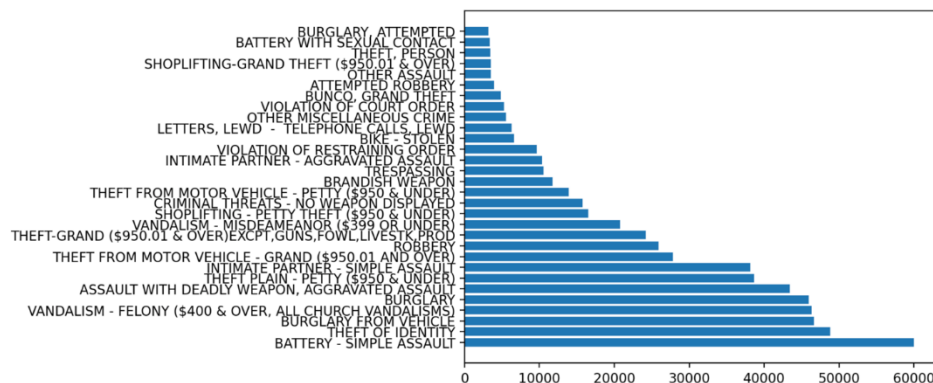


**Figure 1.** Gender Distribution



**Figure 2.** Gender Distribution (after filtering)

After the analysis about genders, the crime type provided in the dataset is also crucial. Due to the massive number of crime type, the top 30 kinds of crime have been selected and illustrated in the bar chart (Fig.3) in descending order (from bottom to top).



**Figure 3.** Crime Type Distribution

Then, in order to indicate the gender distribution in each crime type, the highest 15 types of crime have been utilized to build bar chart (Fig.4) showing the gender differences. From that, in 13 out of 15 crime type, there were more male victims which disobeys our intuition.

Total crime report (Fig.5) shows that the number of crimes had reached a summit from February 2022 to December 2022.

From above data exploring, this research is obviously a classification problem, so the predicting methods which will be used are KNN (K-Nearest Neighbor), LR (Logistic Regression) and Neural Network.

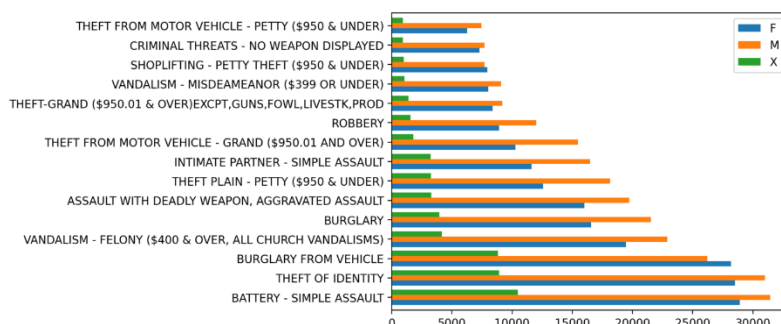


Figure 4. Crime Type Distribution with Genders

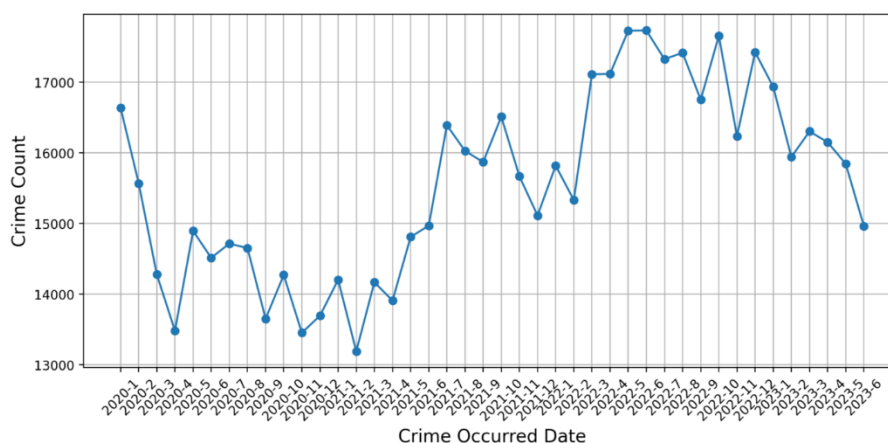


Figure 5. Total Crime Report

## 2.2. Methods

### 2.2.1 KNN

The K-nearest neighbors (KNNs) classifier is a method that calculates the similitude or dissimilarity between two cases in a dataset by measuring the distance between them [6]. The most widely used and simplest distance measure is the Euclidean distance [6]. It identifies the K examples in the dataset that are most similar to the observation being predicted [5]. The KNN classifier then calculates the variable y-value for the observation being predicted based on the variable y's output of these K similar examples [5]. This is a supervised machine learning method that is suitable for both classification and regression tasks, making it a good fit for this research [5].

In this research, KNN had been applied due to the ease of usage, the accuracy of KNN is 59.82% and the f1-score is 65.36%.

### 2.2.2 Logistic Regression (LR)

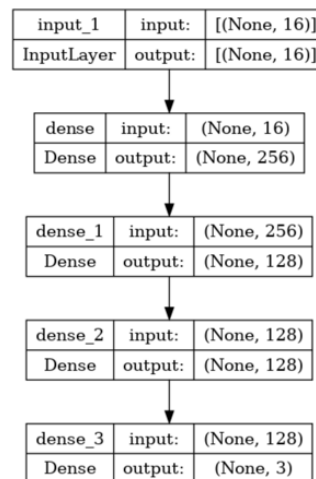
Logistic regression is majorly used in binary classification, but this project has three kinds of genders (M, F and X) need to be classified. Therefore, a layer has been added and adjusted to the LR algorithm, aiming at doing multi-class classification. That method is often called one vs rest (OvR) logistic regression [7]. The main idea behind OvR is transforming a task about classification into various binary tasks, then the log loss is inverted to cross-entropy loss [7]. OvR method is just based on the assumption that each classification task is separated and isolated [8]. In this article for instance, male was used to compare with female then compare with unknowns.

LR got accuracy of 61.31% and f1-score is 67.76%.

### 2.2.3 Neural Network

The configuration of the neural networks utilized in this study is keras, a deep learning framework. Keras is one of neural network libraries based on python that offers high level of modularity. Tensor flow has been chosen in this research as a backend. In addition to this, the activation function used is Sigmoid function and the optimizer method is ‘adam’. The network model was built and the crucial parameters were enhanced by using these modules [9].

The structure of this model is shown in Fig.6.



**Figure 6.** Model Structure

The accuracy of this method is 62.58% while the f1-score is even higher (70.22%).

## 3. Results

This article has used KNN, logistic regression and neural network methods to predict the future 50 victim genders, showing them in the table 2.

**Table 2.** Prediction Results Summary

|                     | Male | Female | Unknown |
|---------------------|------|--------|---------|
| KNN                 | 15   | 31     | 4       |
| Logistic Regression | 28   | 19     | 3       |
| Neural Network      | 30   | 16     | 4       |

## 4. Discussion

By comparing the accuracy and f1-score neural network method shows the better prediction than logistic regression and KNN. Their performance has been sorted into Table 3.

Accuracy: The accuracy is a measure of how well a model predicts the correct outcome [10]. It can be calculated by equation (1), where parameters are explained in Table 4.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

F1-score: In f1-score, precision and recall are combined to evaluate the performance of different models [10]. Its calculation equation is equation (2).

$$F1 = \frac{2TP}{2TP+FP+FN} \quad (2)$$

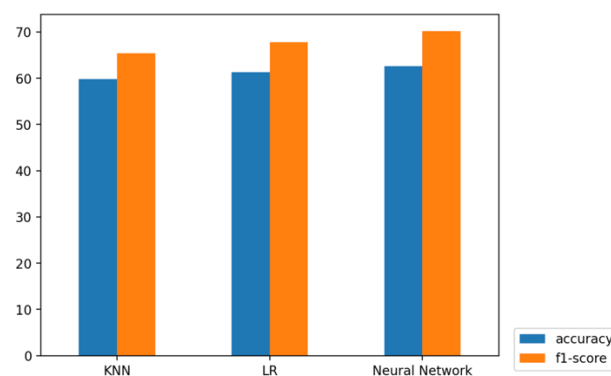
**Table 3.** Performance

|                     | Accuracy | F1-score |
|---------------------|----------|----------|
| KNN                 | 59.82%   | 65.36%   |
| Logistic Regression | 61.31%   | 67.76%   |
| Neural Network      | 62.58%   | 70.22%   |

**Table 4.** Parameters in Equations

| Parameters | Description    |
|------------|----------------|
| TP         | True Positive  |
| TN         | True Negative  |
| FP         | False Positive |
| FN         | False Negative |

The comparison can be seen more clearly via bar charts (Fig.7).



**Figure 7.** Performance Comparison

Only the KNN results shows that females are more vulnerable than males, and KNN has the lowest prediction accuracy. Therefore, from data analyzed in this article, there is a trend that males are more likely to be victims in the future.

## 5. Conclusion

This article utilized three machine learning methods (KNN, LR and neural network) to do some research on future victim genders, the results contained the future 50 victim genders and the performance accuracy of each method. For KNN method, the result shows that there will be 15 male victims and 31 female victims as well as 4 unknown gender victims, but its accuracy is only 59.82% and f1-score is 65.36%. As for logistic regression one, male victim number would be 28 while female victim number is 19, and there will be only 3 unknown-gender victims. The accuracy and f1-score of logistic regression methods is slightly higher than KNN, which are 61.31% and 67.76% respectively, not to mention the accuracy and f1-score of neural network which is the highest among these three methods. Its accuracy is 62.58% and its f1-score even reach 70.22%. The predicted result of neural network method indicates that the number of potential male victims will be 30 and the number of potential female victims would be 16. The government of Los Angeles can use these results to relieve the anxiety of females on crimes.

However, there are still some limitations in this research since only 3 machine learning methods been used. In the future, more methods other than KNN, LR and neural network could be induced to predict victim genders. Also, the criteria of performance measurement can be altered by using other functions like MSE.

## References

- [1] Bianca Pallaro, Tami Abdollah. Which Los Angeles neighborhoods are safest? See the latest trends in the LA crime rates. 2023.03.20. Retrieved at: <https://www.usatoday.com/story/news/2023/03/20/los-angeles-crime-rates-map-safest-neighborhoods/11032761002/>.
- [2] Raqota Berger. Crime in Greater Los Angeles: Experiences and Perceptions of Local Urbane Residents. Center for Criminal and Psychological Studies, Los Angeles, USA, 2018.
- [3] Varun Mandalapu, Lavanya Elluri, Piyush Vyas, Nirmalya Roy. Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions. 2023.
- [4] Paidipati Dinesh, Kalyanasundaram P. Medical Image Prediction for Diagnosis of Breast Cancer Disease Comparing the Machine Learning Algorithms: SVM, KNN, Logistic Regression, Random Forest, and Decision Tree to Measure Accuracy. ECS Transactions, 2022, 1(107).
- [5] Mohammed Marouane Saim, Hassan Ammor. Comparative Study of Machine Learning Algorithms (SVM, Logistic Regression and KNN) to Predict Cardiovascular Diseases. ERSC, E3S Research Center, Mohammed V University of Rabat, Morocco. 2022.
- [6] Elias Dritsas, Maria Trigka. Stroke Risk Prediction with Machine Learning Techniques. Department of Computer Engineering and Informatics, University of Patras, 26504 Patras, Greece. 2022.
- [7] A. B. Amjoud and M. Amrouch, Transfer Learning for Automatic Image Orientation Detection Using Deep Learning and Logistic Regression, in IEEE Access, 2022, (10), 128543-128553.
- [8] Aygyil Epimakhova. Using machine learning to classify and interpret wordplay. DATA LAB du master BIHAR de l'ESTIA, ESTIA BERRI-Technopole Izarbel 90 allée Fauste d'Elhuyar, BIDART, 64210, France.
- [9] Z. Jiang and G. Shen, Prediction of House Price Based on The Back Propagation Neural Network in The Keras Deep Learning Framework, 2019 6th International Conference on Systems and Informatics (ICSAI), Shanghai, China, 2019, 1408-1412.
- [10] Mall PK, Yadav RK, Rai AK, Narayan V, Srivastava S. Early Warning Signs of Parkinson's Disease Prediction Using Machine Learning Technique. Journal of Pharmaceutical Negative Results. 2022 Dec 31: 4784-92.