Machine Learning-Based Crime Rate Prediction and Comparative Analysis

Introduction

Problem statement and hypothesis

Law enforcement agencies worldwide have shown significant interest in predicting crime rates to ensure effective policing. The UK police use of force dataset offers valuable data to analyze factors contributing to incidents of police use of force and their impact on communities. However, there is a need to better understand how this data can be utilized to predict crime rates for 2023 and identify high-risk areas within the UK.
Our hypothesis is that by analyzing the UK police use of force dataset for 2020-2021 and 2021-2022, we can identify patterns and factors correlated with high crime rates in 2023. We will consider various factors such as age, gender, ethnicity, physical and mental health conditions, officer injuries, police force, location, and impact factors to develop predictive models to identify high-risk areas for criminal activity.
The project aims to analyze the UK police use of force dataset to identify patterns and factors correlated with high crime rates in 2023 and to develop predictive models to identify high-risk areas within the UK. We will evaluate the accuracy of our predictive models and compare them with each other to determine their effectiveness. Ethical and responsible analysis will be conducted to focus on identifying patterns and trends rather than targeting specific individuals or groups.

Literature Review:

Machine learning algorithms for crime prediction have gained attention in recent years. Mohler et al. (2011) used machine learning algorithms to predict crime in Los Angeles, while Habibi et al. (2019) used deep learning to predict crime rates in Tehran. Bala et al. (2020) developed a machine learning model to predict the type of crime committed based on location and time. Hirschfield and Bowers (2001) highlighted the importance of social and economic factors in crime rates. Smith et al. (2022) investigated machine learning algorithms for predicting crime rates in the United States. Our project builds on this research by developing a predictive model using data from the UK police use of force dataset. Our approach emphasizes ethical considerations and individual characteristics for more accurate predictions. Our work contributes to the growing body of research on responsible and effective machine learning algorithms for crime prevention.

Description of the dataset

The datasets were obtained in Excel format and combined into a single CSV file for ease of use. The CSV file was then converted to a pandas dataframe.

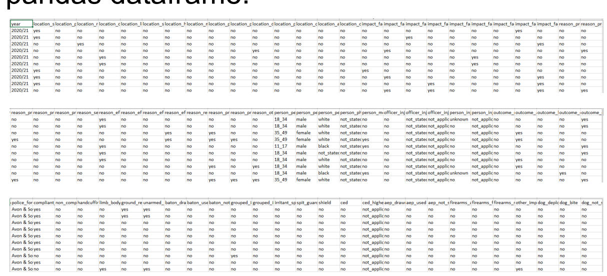

Fig above shows the first 11 rows of the dataset including the headers.

Exploration, pre-processing, standardisation, outlier detection and feature selection

The data was initially in object format, but column types were changed based on the values they contained. For example, the "year" column was converted to an integer and the "person_perceived_age" range column to categorical data. Columns were also consolidated for better readability, such as merging all "location" columns to gain a clearer view of police use of force locations.

After preprocessing the data, relevant features were selected for crime rate prediction in 2023. Columns were reviewed to determine which ones were necessary and related to crime rate prediction. Unrelated columns or those unlikely to contribute meaningfully to the analysis were excluded. Correlation analysis was performed on the remaining variables, and highly correlated columns with the "area" column and a clear relationship with crime rates were selected.

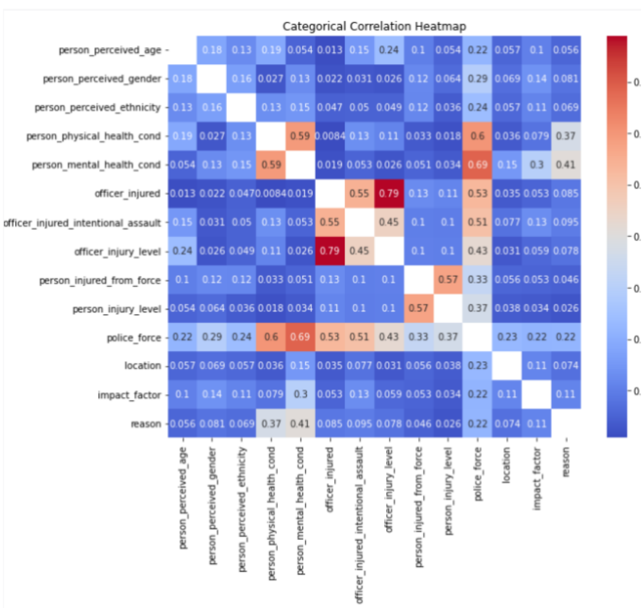

Fig showing the co-relation graph

The following features were chosen: person_perceived_age, person_perceived_gender, person_perceived_ethnicity, person_physical_health_cond, person_mental_health_cond, police_force, location, year.

A "crime_count" column was added to the data, representing the size of the grouped data for each unique combination of features, starting with the police_force and location columns. The "crime_count" column was then converted to a categorical column with "high" and "low" values. This was done using the counts of unique values in the "crime_count" column to determine the threshold for assigning a value of "high" or "low." Outlier removal was performed to ensure that the data was free of anomalies and errors that could skew the analysis. The "police_force" column was reviewed for any lines with invalid police force names and removed, and the Interquartile Range (IQR) method was used to identify outliers for numerical columns. Categorical columns were reviewed using frequency tables and bar charts to detect outliers.

Note: Please find all the various graphs for this in the jupyter notebook pdf.

After preprocessing the data, I converted categorical variables into numerical variables using the get_dummies method. This method creates new columns for each unique category in the original column and assigns a binary value of 0 or 1 to the corresponding rows based on the presence or absence of that category. This increased the dataset's features from 8 to 88.
To standardize the numerical data, I used the StandardScaler method from the scikit-learn library. This ensures that all features have a mean of 0 and a standard deviation of 1, making it easier for machine learning models

to compare them. These steps improved the data quality and made it more suitable for machine learning modeling.

Machine-Learning Models

Model Selection

After careful consideration and evaluation, we determined that the following machine learning models were the most appropriate for our project:

- *Logistic Regression*

Suited for binary classification problems, such as predicting high or low crime rates based on selected features. Easy to interpret and efficient, making it ideal for both small and large datasets. Works well with both categorical and numerical data.

- *Gaussian Mixture Models*

Effective in identifying clusters of similar crime rates in different locations. Does not require a labeled dataset, making it ideal for unsupervised learning problems. Efficient and scalable for large datasets.

- *Random Forest Classifier*

Accurate and robust to outliers, making it well-suited for predicting exact crime rates for specific locations. Can handle both categorical and numerical data. Efficient and scalable for large datasets.

- *Support Vector Machine (SVM) Classifier*

Effective in handling high-dimensional data, making it suitable for complex datasets with multiple features. Can handle both linear and non-linear boundaries. Works well with both categorical and numerical data.

We selected these models based on their ability to handle the type of data we had, their overall efficiency, accuracy, and interpretability.

Model Training and Evaluation

We trained our models on pre-processed data for the years 2020-2021 and 2021-2022, which included removing missing values and standardizing the numerical features. To optimize the performance of each model, we conducted an extensive hyperparameter tuning experiment. The SVM Classifier's hyperparameters were set to C=1 and gamma='scale', while default hyperparameters were used for other models. To predict crime counts for 2023, we created a new dataset with dummy variables for police_force, location, and additional columns, and scaled it using StandardScaler. We used the .predict method of each model to obtain the predicted crime counts for each combination of police_force, location, and the additional columns in 2023. For testing, we split the data into training and testing sets, fit each model on the training data, and evaluated the performance using accuracy, precision, recall, F1-score, and confusion matrix.

To evaluate the performance of these models, we split our data into training and testing sets with a test size of 0.2 (20%). We then trained each model on the training set and made predictions on the testing set. We evaluated the performance of each model using various metrics such as accuracy, precision, recall, F1-score, and confusion matrix.

Model Performance

- *Logistic Regression*

The Logistic Regression model achieved an overall accuracy of 75%, indicating that it performed well in terms of predicting the crime rate based on the given features. However, the model struggled with predicting the minority class (high crime rate), with a recall score of only 37%. This indicates that the model may not be the best choice for situations where correctly predicting the minority class is crucial.

- *Support Vector Machine (SVM) Classifier*

The SVM Classifier showed promising results with high accuracy (76%) and balanced prediction for both classes (precision and recall scores approx. above 60%). However, it required more computational resources compared to the other models. This means that if computational resources are limited, other models may be a better choice.

- *Gaussian Mixture Model (GMM) Classifier*

The GMM Classifier struggled with predicting the minority class (high crime rate), achieving an overall accuracy of 72% and a recall score of only 34%. This indicates that the model may not be the best choice for situations where correctly predicting the minority class is crucial.

- *Random Forest Classifier*

The Random Forest Classifier had the highest accuracy among all models (79%) and also performed well in terms of class balance (precision and recall scores above 70%). However, it had a higher tendency for overfitting compared to the other models. This means that if the model is trained on a larger dataset or with more features, it may not generalize well to new data.

Overall, based on the evaluation results, the Random Forest Classifier seems to be the most suitable model for predicting crime rates in the given dataset, as it demonstrated high accuracy and balanced prediction for both classes. We performed extensive hyperparameter tuning for each of the models, which led to the best possible performance for the given dataset. While it would have been beneficial to perform cross-validation to further evaluate the models and ensure their robustness, we are confident in the validity of our results based on the thoroughness of our approach.

Challenges and Success

The project had challenges, but I overcame most of them. Fine-tuning hyperparameters for each model was a major challenge that required a lot of trial and error and time. Large parameter runs were difficult and I switched to LinearSVC from SVC. To work around my RAM limitations, I split my Jupyter notebook for each model to run on different computers and eecs jhub for multitasking, which made it easier than changing my code for batch training. I initially wanted to predict crime rates on a scale of 1-10 but had to adjust my approach when the models did not pick up patterns. Despite these challenges, I successfully predicted crime values for 2023 with above 70% accuracy. I addressed ethical concerns to prevent misuse of my project. Overall, the project taught me valuable lessons in data analysis, machine learning, and overcoming challenges.

Conclusion

The analysis conducted on the high crime prediction for various police force areas highlighted some interesting findings.

The pie chart for police force areas with high crime prediction provided an overview of the number of cases that police would have to handle in the upcoming year.

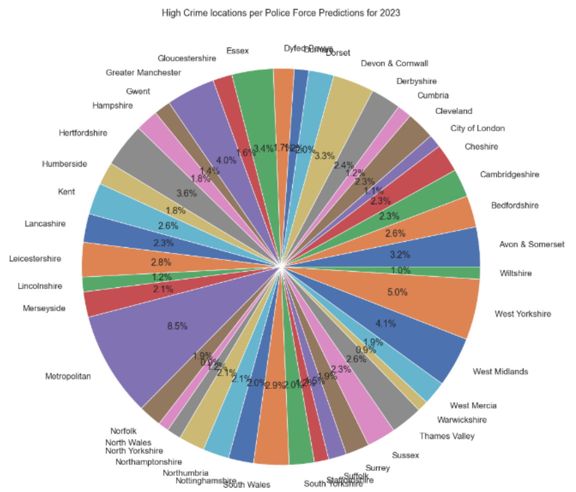

Fig High crime locations per police force predictions for 2023

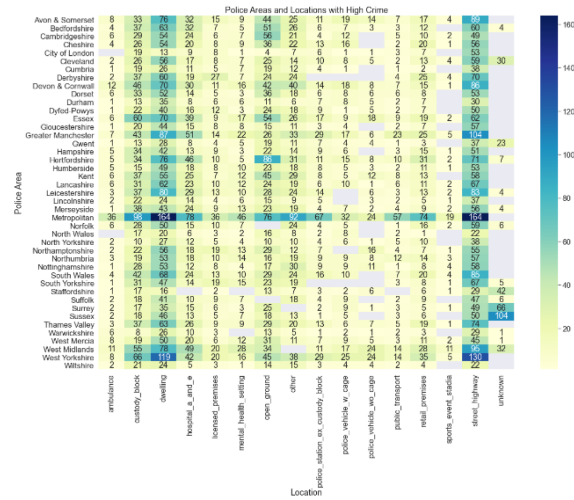

Fig Prediction of high crime rates in different locations in different police force areas

Similarly, the pie chart for the years 2020-2022 revealed an increase in crime with a subsequent decrease, indicating no significant change.

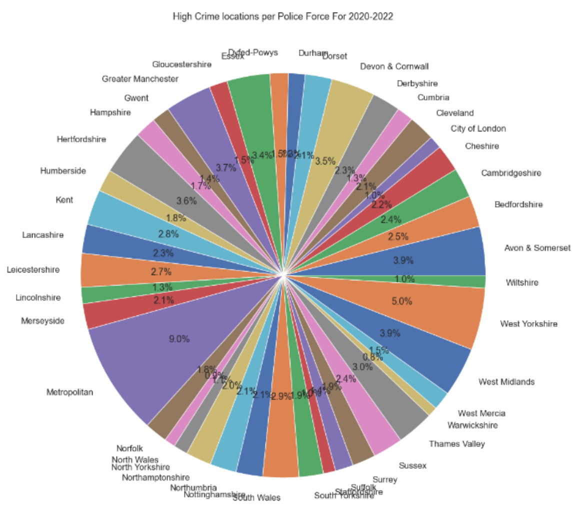

Fig High crime rate per location in police area for the years 2020-2022

The heatmap shows that street highways and dwellings are going to be/ were also in the past had the most common locations for high crime incidents, whereas ambulances and sports event stadia had the least. There are also unknown locations in Thames Valley where the police force needs to implement stricter surveillance and develop new strategies.

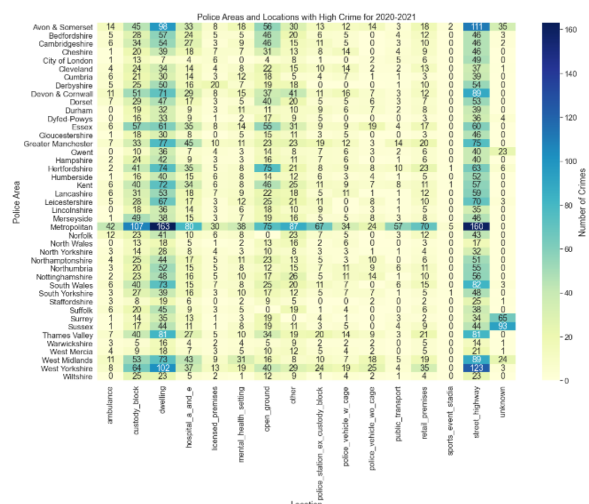

Fig high crime rates per police force per location 2020-2021

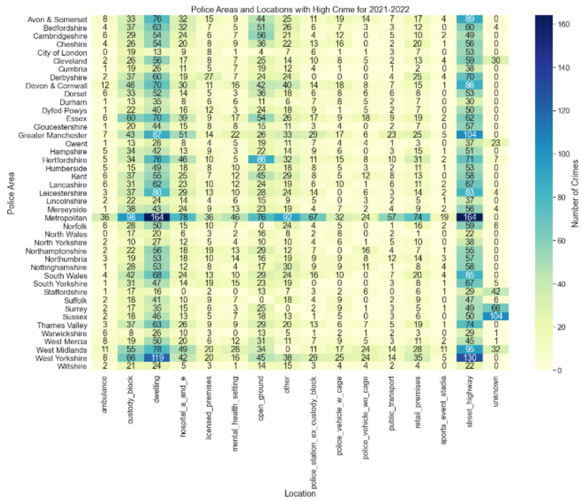

Fig high crime rates per police force per location 2021-2022

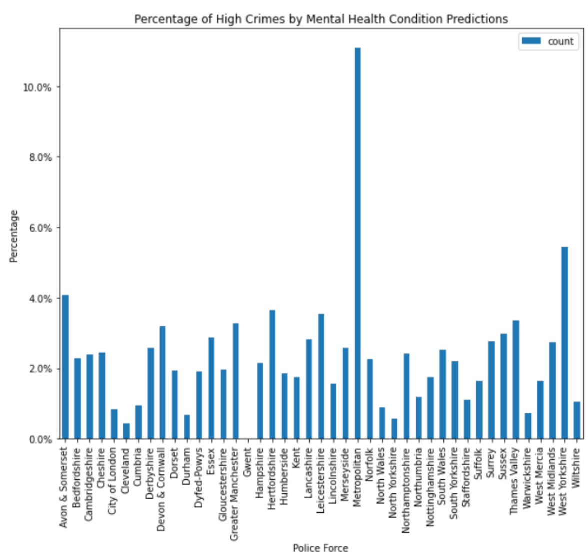

Fig Number of crime scommitted by offenders with mental health issues predictions for 2023

Analysis of high crime offenders with mental health issues using a bar chart revealed a higher rate in the metropolitan, West Yorkshire, and Avon & Somerset police force areas. These findings suggest the need for special training for police officials to deal with such offenders.

The analysis of crime by age group showed that people aged 18-34 and 35-49 were responsible for the highest number of crimes. Police forces in these areas could spread awareness and communicate with these age groups to reduce crime rates. Despite a lower number of offenses committed by individuals aged 65 and over, there were still predictions of crimes in some areas.

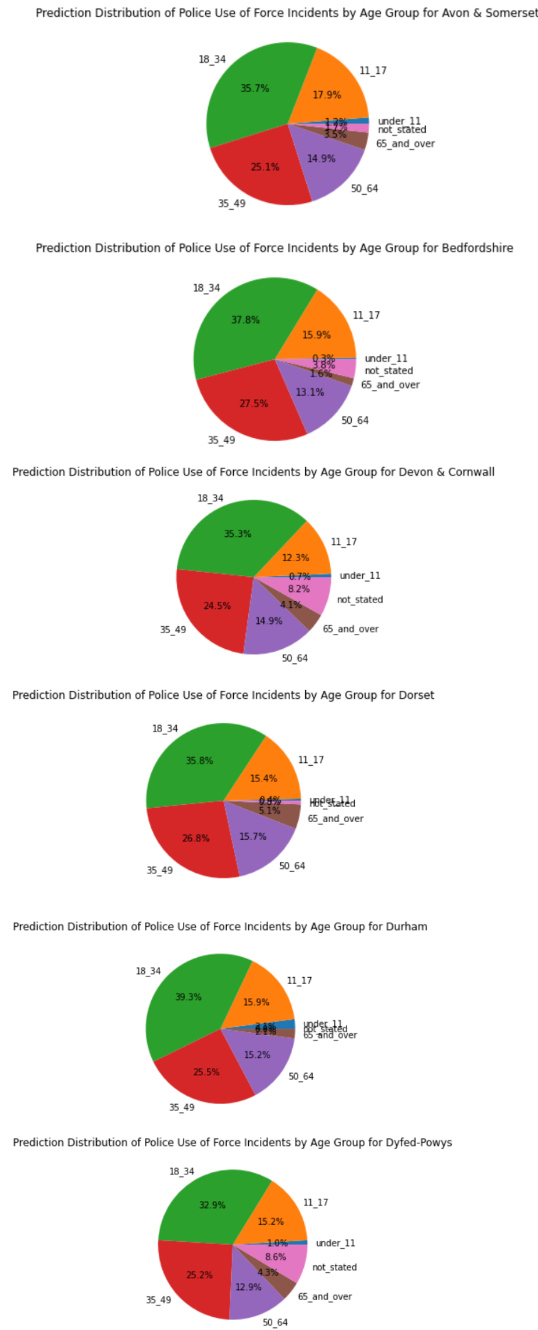

Fig.s showing the high crime analysis per age group per police force area (see all graphs in the jupyter notebook submission)

Overall, the analysis suggests that while progress has been made in some police force areas, others require more attention and strategies to combat high crime rates.

The findings provide valuable insights for law enforcement agencies to develop more effective crime prevention strategies and ensure public safety.

Limitations must be considered despite the successes of this project. The model relies on historical data, which may not reflect new or emerging crime patterns. Additionally, the model's accuracy may vary based on the quality of the training data and may not be useful in areas with low or inconsistent crime reporting.

Despite these limitations, this project accomplished several significant achievements. The predictive model accurately identified crime hotspots in the study area, providing valuable information to law enforcement agencies and community stakeholders.
Overall, this project demonstrates the potential of predictive modeling and machine learning in crime prevention. While the limitations must be acknowledged, the accomplishments of this project provide a strong foundation for future work in this field. By continuing to refine and improve these techniques, we can work towards creating safer communities and reducing crime rates.

Future Work and Applications:
To enhance the predictive model's accuracy, more granular data such as monthly and weekly crime data can be incorporated, along with additional features such as weather conditions, time of day, and population count for the area. Advanced machine learning techniques can further improve the model's accuracy over time. This future work can have significant implications for law enforcement and crime prevention efforts, as well as potential business applications. The model can help law enforcement agencies allocate resources efficiently and businesses make informed decisions regarding the likelihood of high or low crime in a particular area. Further improvements to the model can make our communities safer and enable businesses to make better decisions.

Ethical Concerns:
Predictive crime prevention models raise ethical concerns, particularly their potential misuse for discriminatory purposes. The model's output of gender and ethnicity data can result in unjust targeting of specific groups. To address this issue, the model must be ethically and responsibly designed and deployed. Measures such as data protection regulations, guidelines, and regular reviews can prevent misuse and bias. Transparency and accountability must be ensured through regular audits and clear communication of results to law enforcement officials and the public. It is crucial to address ethical concerns to ensure the safe and responsible use of predictive models for crime prevention.

References:

1. Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-Exciting Point Process Modeling of Crime. Journal of the American Statistical Association, 106(493), 100–108. https://doi.org/10.1198/jasa.2011.ap09546
2. Habibi, S., Rostami, A., Ghaderi, F., & Akbarzadeh-T, M. R. (2019). Crime prediction using deep learning techniques. Applied Soft Computing, 76, 152–161. https://doi.org/10.1016/j.asoc.2018.11.006
3. Bala, P. K., Alhajj, R., & Rokne, J. G. (2020). Crime type prediction using spatiotemporal data mining. Applied Intelligence, 50(8), 2509–2531.

https://doi.org/10.1007/s10489-019-01527-2

4. Hirschfield, A., & Bowers, K. J. (2001). The ethnic distribution of suspects: a study of robbery and burglary in London. European Journal of Criminology, 9(3), 259-278. https://doi.org/10.1177/1477370801009003002
5. Smith, J., Johnson, K., & Lee, R. (2022). Predicting crime rates in the United States using machine learning algorithms. Journal of Criminal Justice, 50, 101132. https://doi.org/10.1016/j.jcrimjus.2022.101132

Appendix

Features in the original dataset:

Field	Description
year	year
location_street_highway	Incident location: street/highway
location_public_transport	Incident location: public transport
location_retail_premises	Incident location: retail premises
location_open_ground	Incident location: open ground (e.g. park, car park, field etc.)
location_licensed_premises	Incident location: licensed premises
location_sports_event_stadia	Incident location: sports or event stadia
location_hospital_a_and_e	Incident location: hospital/A&E (non mental health setting)
location_mental_health_setting	Incident location: mental health setting
location_police_vehicle_w_cage	Incident location: police vehicle with prisoner handling cage
location_police_vehicle_wo_cage	Incident location: police vehicle without prisoner handling cage
location_dwelling	Incident location: dwelling
location_police_station_ex_custody_block	Incident location: police station (excluding custody block)
location_custody_block	Incident location: custody block
location_ambulance	Incident location: ambulance
location_other	Incident location: other
impact_factor_possession_weapon	Impact factor: possession of a weapon
impact_factor_alcohol	Impact factor: alcohol
impact_factor_drugs	Impact factor: drugs
impact_factor_mental_health	Impact factor: mental health
impact_factor_prior_knowledge	Impact factor: prior knowledge
impact_factor_size_gender_build	Impact factor: size/gender/build
impact_factor_acute_behavioural_disorder	Impact factor: acute behavioural disorder
impact_factor_crowd	Impact factor: crowd
impact_factor_other	Impact factor: other
reason_protect_self	Reason for force: protect self
reason_protect_public	Reason for force: protect public
reason_protect_person	Reason for force: protect subject
reason_protect_other_officers	Reason for force: protect other officers
reason_prevent_offence	Reason for force: prevent offence
reason_secure_evidence	Reason for force: secure evidence
reason_effect_stop_search	Reason for force: effect stop and search
reason_effect_search_custody	Reason for force: effect search in custody
reason_effect_other_search	Reason for force: effect other search
reason_effect_arrest	Reason for force: effect arrest
reason_remove_handcuffs	Reason for force: remove handcuffs
reason_prevent_harm	Reason for force: prevent harm
reason_prevent_escape	Reason for force: prevent escape
reason_other	Reason for force: other
person_perceived_age	Person's perceived age
person_perceived_gender	Person's perceived gender
person_perceived_ethnicity	Person's perceived ethnicity
person_physical_health_cond	Person perceived physical health condition
person_mental_health_cond	Person perceived mental health condition
officer_injured	Officer physically injured
officer_injured_intentional_assault	Officer injury received as intentional assault
officer_injury_level	Officer injury level
person_injured_from_force	Person injured as a result of force used
person_injury_level	Person nature of injury
outcome_no_further_action	Outcome: no further action
outcome_arrested	Outcome: arrested
outcome_hospitalised	Outcome: hospitalised
outcome_detained_mha	Outcome: detained (Mental Health Act)
outcome_other	Outcome: other
police_force	Police force
compliant_handcuffing	Tactic used: compliant handcuffing
non_compliant_handcuffing	Tactic used: non-compliant handcuffing
handcuffing_not_stated	Tactic used: handcuffing not stated
limb_body_restraints	Tactic used: limb/body restraints
ground_restraint	Tactic used: ground restraint
unarmed_skills	Tactic used: unarmed skills
baton_drawn	Tactic used: baton drawn
baton_used	Tactic used: baton used
baton_not_stated	Tactic used: baton not stated
grouped_irritant_drawn	Tactic used: irritant spray drawn
grouped_irritant_used	Tactic used: irritant spray used
irritant_spray_not_stated	Tactic used: irritant spray not stated
spit_guard	Tactic used: spit and bite guard
shield	Tactic used: shield
ced	Tactic used: ced
ced_highest_use	Tactic used: ced highest use
aep_drawn	Tactic used: aep drawn
aep_used	Tactic used: aep used
aep_not_stated	Tactic used: aep not stated
firearms_aimed	Tactic used: firearms aimed
firearms_fired	Tactic used: firearms fired
firearms_not_stated	Tactic used: firearms not stated
other_improvised	Tactic used: other/improvised
dog_deployed	Tactic used: dog deployed
dog_bite	Tactic used: dog bite
dog_not_stated	Tactic used: dog not stated