# Research Proposal

# Crime reduction with machine learning: a comparative analysis

# Requirements of an MSc research project

- Understanding of knowledge
- Application of techniques
- Original
- Complex
- Self-direction
- Self-evaluation
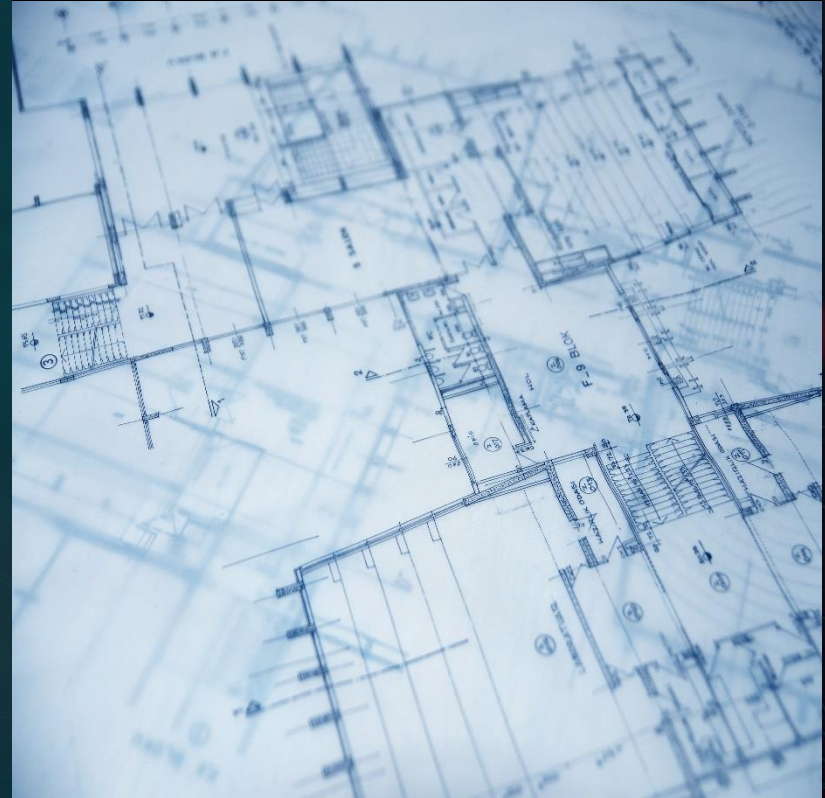
# Significance to the discipline

- Machine learning for crime prediction has been explored academically, but does not have widespread adoption

- Potentially due to academic focus on model performance over practicalities of implementation

- Challenges include:
  - Insights are often not actionable
  - Lack of interpretability*
  - Lack of computational efficiency

- Study seeks to combine performance, interpretability and computational efficiency metrics to evaluate true "usefulness"

- The "usefulness" metric would be applicable to domains outside of crime prediction



*"A machine learning model is interpretable if you can inspect the actual model and understand why it got a particular answer for a given input" (Russel & Norvig, 2021: 739)

# Research questions

- Which metrics should be used to measure machine learning model performance, interpretability, and computational efficiency, to provide an overall blended score?

- Which machine learning models provide the best overall performance, interpretability and computational efficiency when predicting locations of crime?

# Aims and objectives

## Aims

Comparing machine learning performance for crime prediction is difficult due to different datasets and metrics. This study will define a methodology and use a single dataset to compare multiple solutions

## Objectives

1. Propose a methodology to measure machine learning model usefulness by incorporating performance, interpretability and computational efficiency metrics

2. Use the proposed methodology to assess the solutions from three papers using the same dataset to validate the methodology and recommend the best model for predicting crime
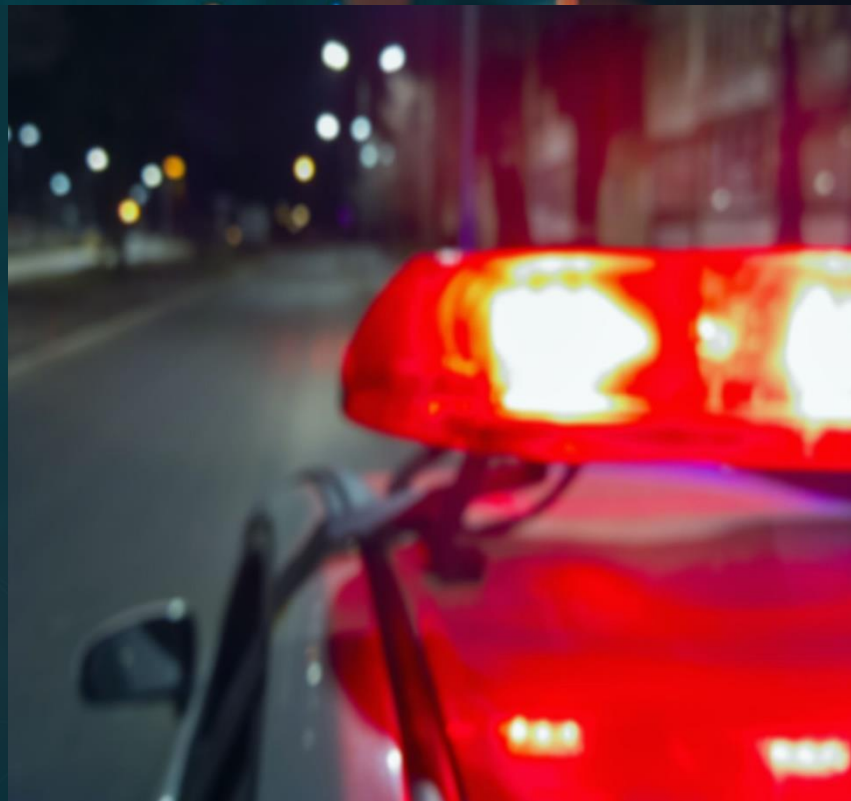
# Key literature

- Literature review (Feaviour, 2024) identified various solutions

- Challenges (varied by paper) included:

  - Different performance metrics used

  - Lack of real-world applicability

  - Lack of attention to interpretability

  - Lack of attention to computational efficiency

- Three solutions selected with real-world applicability to test against each other

| Reference | Machine learning model(s) | Solution overview | Granularity | Relevence to day-to-day policing | Limitations |
|---|---|---|---|---|---|
| Jain et al. (2017) | K-Means | Cluster by crime type and location | Variable; typically a few streets | Low | Not a predictor. More appropriate for strategic resource planning |
| Iqbal et al., (2013) | Decision Tree Naïve Bayes | Demographic data to predict if a state has high, medium or low violent crime | State | Low | Predictions at state level so limited use in day-to-day policing |
| Zaidi et al., (2020) | Random Forest Support Vector Machine | Demographic data to predict if a state has high, medium or low violent crime | State | Low | Predictions at state level so limited use in day-to-day policing |
| Ahishakiye et al. (2017) | Decision Tree | Demographic data to predict if a county has high, medium or low violent crime | County | Low | Predictions at county level so limited use in day-to-day policing |
| Safat et al. (2021) | Logistic regression Decision Tree Random Forest Multilayer Perceptron Naïve Bayes Support Vector Machine GXBoost K Nearest Neigbour | Compare classifier algorithms using two datasets then test time series prediction using ARIMA | District | Potential to inform short-term resource planning | Longer-term accuracy is questionable |
| Sivaranjani et al. (2016) | K Nearest Neighbour K-Means Agglomerative Clustering DBSCAN | Classify crimes with KNN then cluster crimes of the same category to give each city a high/medium/low preopensity per crime type | City | Potentially useful for city-level planning | Lacks granularity to inform day-to-day policing |
| Hajela et al. (2020) | K-Means Naïve Bayes Decision Tree | Cluster crime hot spots then classify crime type at each hot spot | Variable based on cluster size; typically district but centre point is visible | Hot spots are based upon coordinates, so potential for targetted interventions | Each cluster only has one class, so interventions would be based upon most prevelent crime only |
| Cichosz (2020) | Logistic Regression Support Vector Machine Decision Tree Random Forest | Points of interest (POI) aggregated into cells in a geographic grid to predict hotpots of different crime types based upon POI atributes. Performed well when trained on one area and tested on another | 300m² | Predicting crime types based upon POI could enable targetted policing | The grid doesn't give a precise prediction of where crime will occur |
| Rummens et al. (2017) | Logistic Regression Artificial Neural Network | Hotspots by crime type plotted in a grid, split by day and night. Predictions using rolling time window | 200m² | Predictive patterns by day and night could help with proactive policing | Only three crime types, but more could be added |
| Stalidis et al. (2021) | CNN LSTM | Time series maps with crime plots passed through CNN to create feature vectors then LTSM to extract feature vectors to predict emerging crime hotspots | 500m² | Predicted emerging crime hotspots could be used to target police activity | Reliatively high compute required in pre-processing the maps and running the algorithms |

*Machine learning models for crime prediction (Feaviour, 2024)*

# Research methodology

- Three solutions to be assessed from the literature review (Feaviour, 2024)

- Primary research and typical quantitative and qualitative analysis not required

- Machine learning performance metrics:
    - Accuracy
    - Precision
    - Recall
    - F1-score
    - Area Under Curve (AUC)

- Additional literature review to assess:
    - Interpretability
    - Computational efficiency
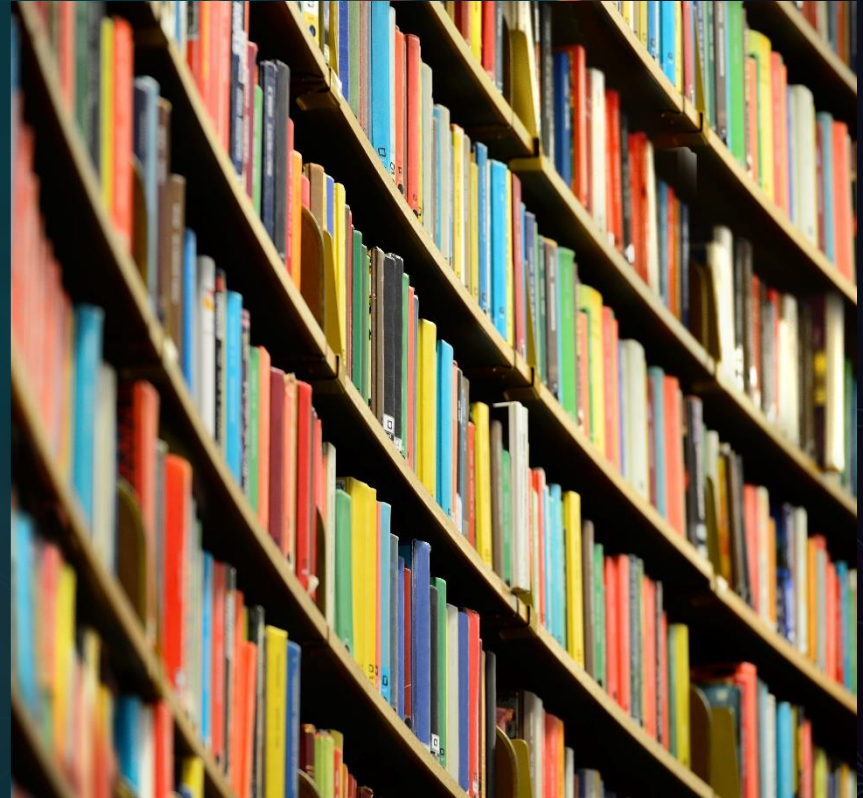
# Model assessment methodology

Google Colab via GitHub

Accuracy · Precision · Recall · F1-Score · AUC · Interpretability · Computational efficiency

Data superset

Hajela et al. (2020)
- Crime
- Location

K-Means → Naïve Bayes
K-Means → Decision Tree

Cichosz (2020)
- Crime
- Location
- Points of interest

Random Forest
Decision Tree

Rummens et al. (2017)
- Crime
- Location
- Demographics

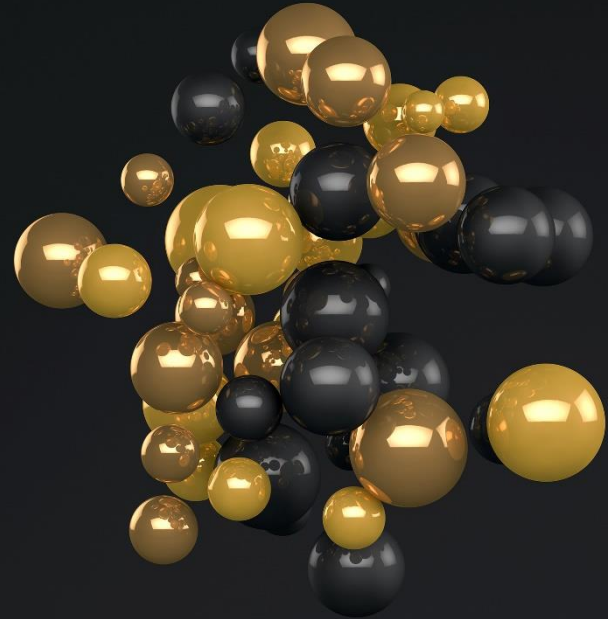Logistic Regression
Artificial Neutral Network

# Ethical considerations

- Data used should not contain personal information

- Outcome will provide a more ethical model for machine learning model assessment due to inclusion of interpretability and computational efficiency metrics
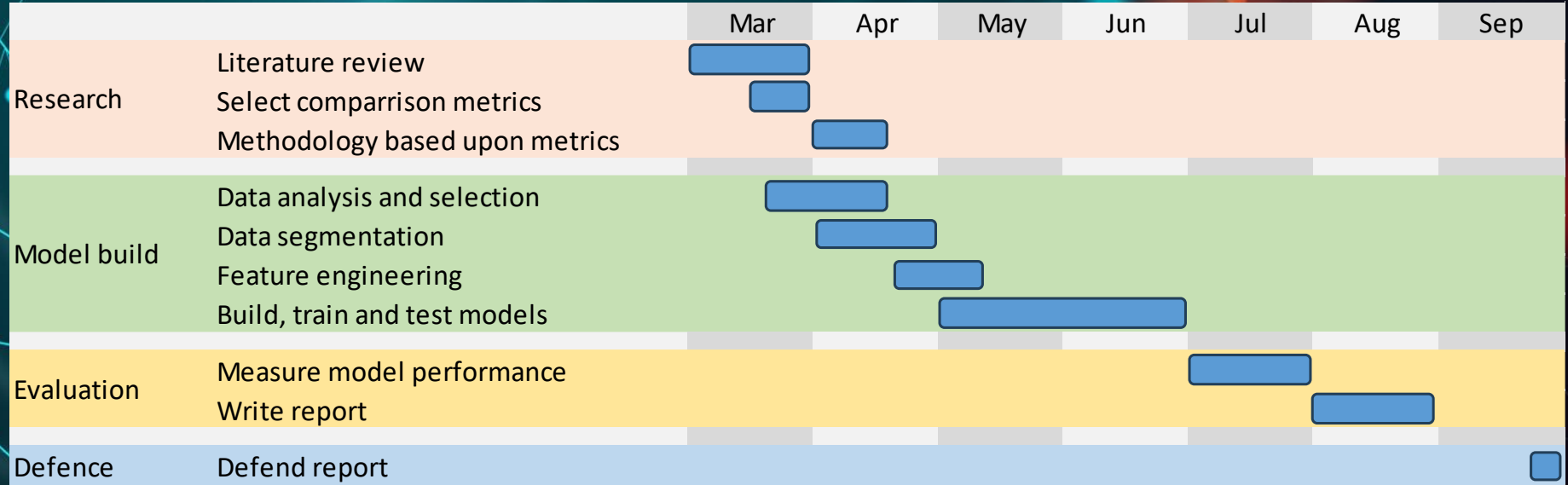
# Artefacts to be created

- Three datasets; one for each approach

- Seven trained and tested machine learning models, all available on GitHub

- One report comparing the results of the six solutions with a recommendation

# Schedule

| | | Mar | Apr | May | Jun | Jul | Aug | Sep |
|---|---|---|---|---|---|---|---|---|
| **Research** | Literature review | ▬ | | | | | | |
| | Select comparrison metrics | ▬ | | | | | | |
| | Methodology based upon metrics | | ▬ | | | | | |
| **Model build** | Data analysis and selection | | ▬ | | | | | |
| | Data segmentation | | ▬ | | | | | |
| | Feature engineering | | | ▬ | | | | |
| | Build, train and test models | | | ▬▬▬ | | | | |
| **Evaluation** | Measure model performance | | | | | ▬ | | |
| | Write report | | | | | | ▬ | |
| **Defence** | Defend report | | | | | | | ▬ |

# References

Ahishakiye, E., Taremwa, D., Omulo, E.O. & Niyonzima, I. (2017) Crime prediction using decision tree (J48) classification algorithm. *International Journal of Computer and Information Technology*, 6(3): 188-195.

Cichosz, P. (2020) Urban crime risk prediction using point of interest data. *ISPRS International Journal of Geo-Information*, 9(7): 459. DOI: https://doi.org/10.3390/ijgi9070459

Feaviour, L (2024) *Machine learning as a tool for police to reduce crime*. MSc Literature review, University of Essex

Hajela, G., Chawla, M. & Rasool, A. (2020) A clustering based hotspot identification approach for crime prediction. *Procedia Computer Science*, 167: 1462-1470. DOI: https://doi.org/10.1016/j.procs.2020.03.357

Iqbal, R., Murad, M.A.A., Mustapha, A., Panahy, P.H.S. & Khanahmadliravi, N. (2013) An experimental study of classification algorithms for crime prediction. *Indian Journal of Science and Technology*, 6(3): 4219-4225. DOI: https://dx.doi.org/10.17485/ijst/2013/v6i3.6

Jain, V., Sharma, Y., Bhatia, A. & Arora, V. (2017) Crime prediction using K-means algorithm. *GRD Journals-Global Research and Development Journal for Engineering*, 2(5): 206-209.

Rummens, A., Hardyns, W. & Pauwels, L. (2017) The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. *Applied geography*, 86: 255-261. DOI: https://doi.org/10.1016/j.apgeog.2017.06.011

Russell, S. & Norvig, P. (2021) *Artificial Intelligence: A Modern Approach*. 4th ed. Harlow: Pearson Education Limited.

Safat, W., Asghar, S. & Gillani, S.A. (2021) Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. *IEEE access*, 9: 70080-70094. DOI: https://doi.org/10.1109/ACCESS.2021.307811

Stalidis, P., Semertzidis, T. & Daras, P. (2021) Examining deep learning architectures for crime classification and prediction. *Forecasting*, 3(4): 741-762. DOI: https://doi.org/10.3390/forecast3040046

Zaidi, N.A.S., Mustapha, A., Mostafa, S.A. & Razali, M.N. (2020) 'A classification approach for crime prediction'. *Applied Computing to Support Industry: Innovation and Technology*. Ramadi, Iraq,15–16 September. Switzerland: Springer International Publishing. 68-78.